MONTI

Monografías de Traducción e Interpretación
Monografies de Traducció i d'Interpretació
Monographs in Translation and Interpreting
Monographies de Traduction et d'Interprétation
Monographien zur Translation

13
2021

CTS spring-cleaning: A critical reflection

Reflexión crítica en los estudios de traducción basados en corpus

Calzada Perez, Maria &
Sara Laviosa (eds.)

ISSN-e 1989-9335
ISSN 1889-4178

Universitat d'Alacant
Universidad de Alicante

Universitat Jaume I

Universitat de València

MONTI

13 (2021)

Universitat d'Alacant - Universitat Jaume I - Universitat de València

Número de artículos propuestos para *MonTI 13* (2021) / Number of contributions submitted to *MonTI* 13 (2021): Diecinueve / Nineteen.
Número de artículos aceptados en *MonTI 13* (2021) / Number of articles accepted in *MonTI 13* (2021): Nueve/ Nine (47,4 %)

La revista *MonTI* está indexada en / *MonTI* is indexed in:
BITRA, Carhus Plus+, Dialnet, ESCI (Web of Science); DICE, ERIH-PLUS, FECYT, ISOC, Latindex, Redalyc, Scopus & TSB.

Website: https://web.ua.es/es/monti

MonTI ha recibido ayudas económicas de la Facultad de Filosofía y Letras (Universitat d'Alacant), del Vicerectorat d'Investigació i Doctorat (Universitat Jaume I) y de la Universitat de València (Departament de Filologia Anglesa i Alemanya, Departament de Filologia Francesa i Italiana, Departament de Teoria dels Llenguatges i Ciències de la Comunicació).

Publicacions de la Universitat d'Alacant
03690 Sant Vicent del Raspeig
publicaciones@ua.es
https://publicaciones.ua.es
Telèfon: 965 903 480

ISSN-e 1989-9335
ISSN 1889-4178
Dipòsit legal: A-257-2009

Composició:
Marten Kwinkelenberg

MonTI está editada por las universidades de Alicante (Departamento de Traducción e Interpretación), Jaume I (Departament de Traducció i Comunicació) y València (Departaments de Filologia Anglesa i Alemanya, de Filologia Francesa i Italiana i de Teoria dels llenguatges i Ciències de la Comunicació).

MARÍA CALZADA & SARA LAVIOSA (EDS.)

MONTI 13 (2021)

REFLEXIÓN CRÍTICA EN LOS ESTUDIOS DE TRADUCCIÓN BASADOS EN CORPUS

CTS SPRING-CLEANING: A CRITICAL REFLECTION

UNIVERSITAT D'ALACANT
UNIVERSITAT JAUME I
UNIVERSITAT DE VALÈNCIA

ÍNDICE

Recibido / Received: 08/02/2021

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.01

Para citar este artículo / To cite this article:
Calzada Pérez, María & Sara Laviosa. (2021) "Twenty-five years on: Time to pause for a new agenda for CTIS." In: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection*. *MonTI* 13, pp. 7-32.

TWENTY-FIVE YEARS ON: TIME TO PAUSE FOR A NEW AGENDA FOR CTIS[1]

María Calzada Pérez
calzada@uji.es
Universitat Jaume I

Sara Laviosa
sara.laviosa@uniba.it
University of Bari Aldo Moro

Abstract

The introduction of corpora in descriptive and applied translation and interpreting studies goes back to the 1990s, when the corpus linguistic approach was making considerable progress in descriptive and applied language studies. Twenty-five years on, Corpus-Based Translation and Interpreting Studies (CTIS) is a well-established field of interdisciplinary research worldwide. Its growth goes hand in hand with technological advancements, which make it possible to design, create and share monolingual and multilingual spoken, written and multimodal corpora as resources for theoretical, descriptive and applied research in both translation and interpreting studies. We believe this is the right time to pause and reflect on the achievements and criticalities of this variegated area of scholarship and practice in order to look to the future with renewed confidence and awareness of the challenges that lie ahead.

1. This article was partly carried out within the research project *Representaciones originales, traducidas e interpretadas de la(s) crisis de refugiados: triangulación metodológica desde el análisis del discurso basado en corpus (RE-CRI)*, financially supported by Ministerio de Ciencia e Innovación (PID2019-108866RB-I00).

Keywords: Corpus-based Translation Studies; Corpus-based Interpreting Studies; Corpus Linguistics; Corpus-assisted Discourse Analysis; Contrastive Linguistics.

Resumen

La introducción de los corpus en los estudios descriptivos y aplicados de traducción e interpretación se remonta a la década de 1990. Han pasado ya (más de) 25 años y los estudios de traducción e interpretación basados en corpus (CTIS) son un campo de investigación interdisciplinar bien establecido en todo el mundo. Su crecimiento va de la mano de los avances tecnológicos, que permiten diseñar, crear y compartir corpus monolingües y multilingües (orales, escritos y multimodales) como recursos para la investigación teórica, descriptiva y aplicada en los estudios de traducción e interpretación. Creemos que es el momento oportuno para hacer una pausa y reflexionar sobre los logros y las carencias de este variado ámbito de la erudición y la práctica, con el fin de mirar al futuro con renovada confianza y conciencia de los retos que nos esperan.

Palabras clave: Estudios de traducción basados en corpus; Estudios de interpretación basados en corpus; Lingüística de corpus; Análisis del Discurso Asistido por Corpus; Lingüística contrastiva.

1. The origins of Corpus-based Translation and Interpreting Studies

Corpus-based Translation Studies (CTS) denotes an area of research that adopts and develops the methodologies of Corpus Linguistics (CL) to analyse translation practices for theoretical, descriptive and applied purposes. CL is an approach to language studies, which is based on the analysis of corpora, i.e. collections of authentic texts held in electronic form and assembled according to specific design criteria. In this article, we trace the development of CTS from its advent in the 1990s to the present day, highlighting its main achievements in various ambits of enquiry, with a view to providing the background to the collection of papers selected for this volume of *MonTI*.

After Mona Baker's (1993) seminal paper, where a research agenda for the corpus-based approach to translation studies was outlined, the first collection of papers devoted to this research area was published in 1998 in a special issue of *Meta* entitled *L'Approchee Basée sur le Corpus/The Corpus-Based Approach*, guest edited by Sara Laviosa. The papers commissioned for this issue were grouped under two main headings: "Theoretical Research"

and "Empirical and Pedagogical Studies." The first group outlines the scope, object of study, and methodology of the emergent corpus-based approach. The second group consists of empirical and pedagogical studies of the product and process of translation. The concluding paper by Maria Tymoczko draws on the insights provided by these essays and points out that investigating translation by means of corpora serves "to address not simply questions of language and linguistics, but also questions of culture, ideology, and literary criticism" (1998: 653).

More specifically, CTS is viewed as having a central role within the whole discipline of Translation Studies, because (1) it is committed to integrate linguistic and cultural studies and explore their interrelationship, (2) it shows awareness of the effect of ideology on the theory, practice and pedagogy of translation, and (3) it aims to adapt modern technologies to enhance theory, empirical research and practice for the benefit of translator training and the work of the practising professional translator. Hence, Tymoczko (1998: 658) encourages a view of CTS that offers "the opportunity to reengage the theoretical and pragmatic branches of Translation Studies, branches which over and over again tend to disassociate, developing slippage and even gulfs." Indeed, these early corpus-based studies already illustrate some of the main lines of enquiry that, as envisaged by Tymoczko, would be developed in the years that followed. We now critically review each of these lines of enquiry in turn, with a view to showing how this scholarly research has enhanced our understanding of translation and translating up until now, and will continue to do so in future.

Three research domains can be identified in the late 1990s, each of them concerning issues and topics that fall within the realm of one of the three branches of the discipline as a whole, namely theoretical, descriptive and applied translation studies. As regards theory, the paper authored by Mona Baker (1998) deals with the rationale and motivation for investigating the product and process of translation through corpora expounding on the research agenda she had launched five years earlier. Baker discusses the need to develop a coherent corpus-based methodology for identifying the distinctive features of translational language. She argues that the aim of this research endeavour, which builds upon the studies of scholars working in the descriptive and target-oriented approach to Translation Studies, is not merely

to unveil the nature of the 'third code' per se, but to understand the specific constraints, pressures and motivations that influence the act of translating intended as a mediated communicative event. These considerations seem to echo Tymoczko's appraisal of CTS as an approach whose "[m]odes of interrogation – as well as care in the encoding of metatextual information about translations and texts – allow researchers to move from text-based questions to context-based questions" (1998: 653).

In a similar vein, Miriam Shlesinger posits that just as translation is a communicative event shaped by its own goals, pressures and context of production, so too is interpreting. The term 'interpreting' is intended as "the production of oral output based on other-language input which may be either written (to be read) or unwritten (impromptu)" (1998: 486-487). Shlesinger proposes to extend Baker's notion of monolingual comparable corpora (consisting of a corpus of original texts in one language and a corpus of comparable texts translated into that same language) so as to comprise three collections of texts in the same language: interpreted speeches from a variety of source languages, original spoken texts produced in similar settings, and written translations of source oral texts delivered in analogous circumstances. This novel design would permit not only the study of interpreted texts as distinct instantiations of oral discourse, but also the identification of those regular patterns of language use that distinguish interpreting from written translation. Furthermore, Shlesinger proposes to adapt the traditional uniderectional parallel corpus design so as to include three sets of texts: source language texts, their interpreted versions, and their written translations. The particular advantage of this corpus type, she argues, is that it makes it possible to investigate language- and direction-specific features of the interpreted output along with their possible interaction with personal variable such as gender, extent of professional experience, or language background. On the basis of these considerations, Shlesinger envisages that the corpus-based study of interpreting will help scholars to continue to explore the common ground between interlingual written and spoken mediation as well as define what sets interpreting apart from translation (1998: 490-491).

With regard to parallel corpora, Kirsten Malmkjær (1998) explains the advantages of using these resources for contrastive and translation studies. For contrastive linguists, parallel corpora are valuable for investigating the

differences and similarities in language use. For translation scholars, they are valuable for identifying translational norms. She then discusses two main problems connected with the use of parallel corpora for answering questions arising from within Translation Studies in particular. The first problem is that KWIC concordance lines do not always offer sufficient linguistic context to investigate features of whole texts. There exists, therefore, a risk that some aspects of translational behaviour may be revealed, while others may be overlooked. The second difficulty is related to the way parallel corpora are designed so as to include only one translation for each source text. This may hide an important aspect of the translational phenomenon, namely the differences existing between the various translations of the same original work. To remedy these shortcomings, Malmkjær suggests complementing norm-oriented studies, which require large amounts of text, with smaller and carefully constructed corpora which consist of one source text and as many translations of it as possible, so that in-depth investigations of entire texts can be performed. There are two advantages in combining these two different types of corpora. First, the findings would be richer, and second, they would be more rigorous, given that the larger corpus could be checked against the individual cases examined in the smaller corpus. Malmkjær argues that this methodology caters for the needs of both the contrastive linguist and the translation scholar, bringing them closer to one another in a relationship of mutual cooperation and encouraging synergies with bordering fields that, at any rate, have always been part of translation studies practices.

While Baker and Shlesinger explain the rationale for exploring translation and interpreting through corpora, and Malmkjær proposes to refine corpus-based methodology to address questions germane to contrastive and translation studies, Sandra Halverson (1998) discusses the issue of representativeness in the design of general purpose translation corpora and provides a coherent theoretical framework within which data and methodology form a coherent whole so as to ensure the comparability of empirical findings. To this end, Halverson proposes a prototypical conceptualisation of the object category as opposed to a classical one. In this approach, the target population is regarded as a prototype category whose centre is occupied – but only for the cultures of industrialised western countries – by professional translations, whereas the peripheral positions are filled in by clusters of

different typologies, for instance those carried out by trainees or non-professional translators or those between one's own best language and another language. The relationship between the centre and the periphery within the prototype is not one of inclusion or exclusion of the elements belonging to the category, but of resemblance. Therefore, the boundaries of the different groups of translations are not impermeable. For the researcher this means that, in order to ensure representativeness, a sample corpus of the population of translated texts will have to be made up of an array of subcorpora enjoying different degrees of significance and all being regarded as legitimate objects of study. Given that prototypes are by definition culture-bound, corpus-based findings cannot be generalised beyond the specific target population that a given corpus represents.

Summing up, these early theoretical reflections focus on why, what and how we study translation and interpreting through corpora. We now move on to review the findings obtained from the empirical corpus studies of translation published in the special issue of *Meta* in 1998. Sara Laviosa's (1998b) investigation of English translated text is based on a subcorpus of the English Comparable Corpus (ECC). It comprises two collections of narrative prose in English: one is made up of translations from a variety of source languages (mainly romance languages), the other includes original English texts produced during a similar time span. The study reveals four patterns of lexical use in translated versus original texts: a relatively lower proportion of lexical words versus grammatical words, a relatively higher proportion of high-frequency versus low-frequency words, relatively greater repetition of the most frequent words, and less variety in the words most frequently used. Laviosa proposes to call these regularities in English translated text 'core patterns of lexical use' to indicate that, given their occurrence in both the newspaper (see Laviosa 1998a) and narrative prose subcorpora of ECC, they may prove typical of translational English in general.

Still within the quest for regular features of translational language, Linn Øverås investigates explicitation. She uses two subcorpora consisting of English and Norwegian translations of fiction, taken from the bi-directional English-Norwegian Parallel Corpus (ENPC). Øverås hypothesises a rise in the level of lexical and grammatical cohesion when translating from Norwegian into English and from English into Norwegian. The comparison

of the distribution of explicitating and implicitating shifts reveals a general tendency to explicitate in both translational English and translational Norwegian, notwithstanding a lower level of explicitation in Norwegian-English translations. Øverås finds two types of increased cohesion, i.e. addition and specification. Addition involves the insertion of grammatical or lexical items not present in the source text. Specification involves the expansion or substitution of grammatical and lexical items present in the source text. Øverås considers several factors that can account for explicitation, for example, the stylistic preferences in the source and the target languages, the obligatory shifts resulting from target language grammatical rules or from culture-specific translation norms as well as the constraints inherent in translation as mediated communicative event. She also suggests linking explicitation with neutralization (the tendency to use common, unmarked collocations or similes rather than metaphors) because they both have the effect of achieving greater readability of the target text. In so doing, she implicitly highlights the interrelationship between different posited regular patterns of the 'third code'.

Moving on from the study of patterns to the study of shifts, Jeremy Munday reports on the preliminary findings of the analysis of Edith Grossman's translation, *Seventeen Poisoned Englishmen*, of a short story by Gabriel García Márquez, *Diecisiete ingleses envenenados*. Munday uses a variety of basic corpus linguistic analytical methods – word frequency lists, descriptive statistics and concordances – to explore texts inductively. Word frequency lists are first obtained for both source and target texts and then compared for identifying useful areas of investigation. Munday uses intercalated text, i.e. a text obtained by manually keying in the translated text between the lines of the source text. He then runs concordances of this intercalated text and uses them to carry out a contextualised comparative analysis of all the instances of selected lexical items in order to examine the shifts that build up cumulatively over the whole text as a result of the choices taken by the translator. Such analysis is carried out to understand the decision-making process underpinning the product of translation and infer the translator's textual-linguistic norms. Munday's approach is therefore descriptive, product- and process-oriented and data-driven. He derives his hypotheses from observing differences that occur in the parallel frequency

lists and during the manual construction of the intercalated text. These initial hypotheses are then investigated with the aid of additional automatic methods of analysis such as aligned concordance lines. Munday's investigation of the first 800 words of his full-text parallel corpus reveals shifts in cohesion and word order that occur over the whole text and have the effect of moving the narrative viewpoint from the first to the third person and thereby distancing the reader from the thoughts, experiences and feelings of the main character in the story.

From the perspective of contrastive linguistics, Belinda Maia analyses the frequency and nature of the SVO sentence structure in English and Portuguese, particularly in those cases where the subject is realized by the first person pronoun *I* and *eu*, respectively, or by a name. The corpus she analyses is a small bidirectional parallel corpus comprising an English novel and its Portuguese translation and a Portuguese novel and its English translation. Maia considers this corpus design appropriate for comparing how the same situation is represented in the two languages in the parallel subcorpus, while the bilingual comparable subcorpus permits additional comparisons between the original languages and between the translational and the non-translational varieties of the same language. The discrepancies she observes in the frequencies of personal subjects (realized by either names or pronouns) suggest that, contrary to English language use, the seemingly subjectless V+O sentence structure is the norm in original Portuguese, and translational Portuguese is influenced by English norms. Moreover, while the use of *I* is syntactically necessary in English, the occurrence of the Portuguese equivalent *eu* seems to be related to pragmatic factors such thematisation, topicalisation and emphasis, while the verb acts as the normal theme of a high proportion of Portuguese sentences. On the basis of these results, Maia argues that the flexibility of word order and the wider variation of thematisation in Portuguese in relation to English allow for more subtlety in communication.

Like Maia, Jarle Ebeling regards parallel corpora as suitable sources of data for investigating the differences and similarities between languages, and adopts the notion of translation equivalence as a methodology for contrastive analysis. Ebeling uses the ENPC, a bidirectional parallel corpus of Norwegian and English texts, to examine presentative English

there-constructions and the Norwegian equivalent *det*-constructions in original and translated English as well as original and translated Norwegian respectively. The corpus of original English reveals that *be* is by far the most frequent verb occurring in these structures, while Norwegian allows a much wider set of verbs, some of which in the passive voice. Ebeling's analysis of the Norwegian translation equivalents of the English *there be*-constructions reveals the influence of the target language. He finds that translators widen the range of *det*-constructions by using a) other verbs over and above those of existence, b) *ha*-existentials (corresponding to *have*-existentials in English), and c) *det*-constructions with passives. In Ebeling's view, this wider choice renders the meaning expressed in the translation more specific, that is more informative, compared with the original. Conversely, the English translations of *det*-constructions with lexical verbs in the active voice are very frequently rendered with *there be*-constructions, and this leads to less specification or 'despecification'. These results partly confirm the predictions put forward on the evidence provided by the analysis of the original corpora and throw new light on the assumed relationship of equivalence between these two structures in English and Norwegian.

With regard to the applied branch of corpus studies, the research carried out by Federico Zanettin and Lynne Bowker are of particular interest for translator training. Zanettin demonstrates how small bilingual comparable corpora are useful to explore the stylistic features of a particular text genre by comparing words and phrases that have a strong formal resemblance such as proper names and cognates or are based on lexicographic translation equivalents. Zanettin provides examples of such searches carried out in class with an Italian-English comparable corpus of leading daily newspapers. The way in which President François Mitterand is talked about in the two languages, for example, presents interesting differences: *François Mitterand* or simply *Mitterand* is commonly used in Italian, while English prefers *President Mitterand* or *President François Mitterand* or *Mr Mitterand*. Also, equivalent verbs typically used to introduce direct and reported speech have different frequencies as well as syntactic and collocational profiles in the two languages. Even cognates such as *prezzi* and *prices* show different collocational and colligational patterns. These data-driven learning

investigations are valuable for refining contrastive knowledge of the source and target languages and enhancing translation skills.

Still within a pedagogic perspective, Bowker addresses two main problems usually faced by students training to become professional translators in specialized subject domains. One difficulty is shown by the occurrence of terminological errors resulting from poor subject-specific knowledge. The other is shown by the occurrence of errors due to a lack of specialized writing skills in the target language. Bowker's pilot study reports on a translation experiment undertaken with a group of fourth-year undergraduate L1 English students at Dublin City University who carried out two translations from L2 French of two semi-specialized passages on optical scanners. One translation was completed with the use of bilingual and monolingual dictionaries together with non-lexicographic reference materials (e.g. manuals and brochures). The other translation was carried out with a bilingual dictionary and a 1.4 million-word specialized monolingual corpus of English articles on optical scanners, which was compiled from *Computer Select* on CD-ROM. The software used to analyse the corpus was *WordSmith Tools*. The findings reveal that the corpus-aided translations were of higher quality in respect of subject field understanding (*sensibilité aux nuances* was accurately rendered as *whatever their sensitivity to colour*); correct term choice (*vitre/glass paten* or *scan bed*); and idiomatic expression (*photodiodes sensibles à la lumière/light-sensitive photodiodes* or *photosensitive diodes*). Bowker observes that, although there was no improvement with regard to grammar or register, the use of a specialized monolingual target corpus was not associated with poorer performance.

Revisiting the past, as we have done here, enables us to appreciate more than ever the value of this early research, which sowed the seed of the variegated lines of enquiry that, since the turn of the century, have contributed to the establishment, consolidation and growth of corpus studies of translation and interpreting in the pure and applied branches of the discipline, pushing the whole field of scholarship towards empiricism and interdisciplinarity.

2. CTIS: Twenty-five years on

More than twenty-five years have passed since Baker (1993) put forward a research agenda for using corpora in descriptive translation studies and the first collection of papers on corpus-based translation and interpreting studies was published (Laviosa 1998). For the past two decades, Corpus-based Translation and Interpreting Studies (CTIS) has grown to such an extent that, already at the beginning of the 2000s, Baker (2004, 169) states that we "have too much rather than too little to go on."

Throughout these recent decades, many corpora have been compiled and extensively reviewed by the literature (e.g., Hu 2016; Laviosa 2002; Olohan 2004). Federico Zanettin (2012: 10) captures this overwhelming proliferation visually, as in Figure 1:

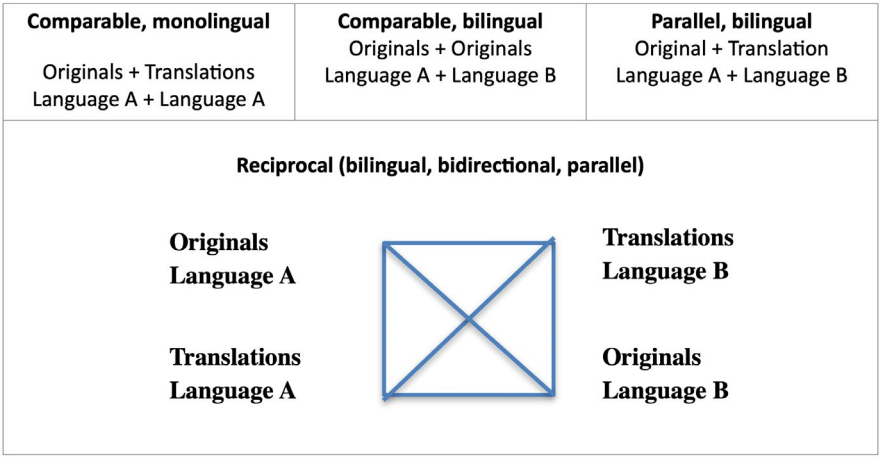

Figure 1: Zanettin's (2012) typology of translation-related corpora.

And Figure 1 does not exhaust corpus mapping. Zanettin (2012: 11) himself complements it with other cases of corpus typologies:

> A bilingual, reciprocal corpus may be graphically represented as a square cut across by diagonal lines, in which the different subcorpora stand at the corners. Multilingual, reciprocal parallel corpora may generate complex models described as a star and diamond configurations (Johansson 2003:

139–142). In a star model, there are multiple translations of the same texts in different languages. The diamond model includes source texts in more than two languages as well as their translations in all the other languages.

With all these corpora, as Bernardini & Kenny (2020) confirm in a concise but thorough review, much of what was already underway in 1998 has been pursued further, especially with regard to the most prominent of all CTIS research paths — the study of the distinctive features of translational language (e.g. Othman 2020; Váradi 2007; Xiao 2010).

It is true that, with the turn of the millennium, early corpus-based studies have been complemented by newly-conceived alternative research lines, but, admittedly, to a comparatively much lesser extent. This is the case, for example, of explorations into translator style (e.g. Saldanha 2011). It is also the case of tentative incursions into multimodality, examining under-represented areas such as interpreting (e.g. Hu & Tao 2014) or audiovisual translation (e.g. Baños et al. 2013). It is indeed the case of incipient approaches to translational data "produced under relatively new conditions" (Bernardini & Kenny 2020: 113), such as the work on (previously inexistent) modalities and genres like web localisation and social media translation (e.g. Jiménez-Crespo 2015). Nevertheless, as mentioned above, well into the 21st century, these alternative research routes are still in the periphery of CTIS.

The beginning of the 1990s, then, saw the official birth of CTIS thanks to a change in paradigm within TIS, made possible by the increasing strength of descriptive translation studies and the proliferation of ever more powerful electronic tools. In less than a decade, CTIS grew steadily (see Santamaría Urbieta & Alcalde Peñalver, in this volume), and by the end of the 20th century, there was enough material for Laviosa (1998) to review the field for the first time. It seems to us that, in 2021, the time is particularly ripe for another, critical (self-)reflexive pause. We are not alone in this consideration.

De Sutter & Lefer (2020), for example, put forward the same proposal in a particularly vehement manner when rallying for a new, updated research agenda for empirical translation studies, which they believe is "indispensable if we want findings in empirical translation studies to be accurate, reliable and generalisable so that we can start building solid, stable theories." (p. 7) Both authors (De Sutter & Lefer 2020: 2) believe that this agenda would

make it possible to advance in the study of largely unanswered questions relating to:

> [the] mechanisms that shape translation, how these mechanisms interact, and to what extent this interaction functions differently than in other types of monolingual and bilingual written language production.

In their view, this new agenda, which is to be built upon particularly solid empirical pillars (and which is to be characterised by statistical sophistication, interdisciplinarity and a multi-methodological framework), is essential to overcome four main potentially lethal risks for CTIS:

(1) CTIS research is currently still focusing (to an excessive extent) on linguistic differences between translated and non-translated texts in order to identify distinctive translational features. Nevertheless, similarities are also an essential part of the picture to understand the process and product of translation (and interpreting), and this part of the picture, for De Sutter & Lefer (2020: 4), is mostly overlooked. In fact, for them:

> with the benefit of hindsight, it is a questionable approach to assume first and foremost differences when translated texts in a given language that are produced by highly skilled, native-language professionals are compared with texts in the same language produced by presumably equally skilled language professionals (journalists, writers, spokesmen, etc.), with the only obvious difference being the circumstances under which the texts are produced (bilingual vs monolingual language activation). […] In other words, subtle quantitative differences are likely to be found across translated and non-translated texts, alongside a massive number of commonalities.

(2) There is much room for improving the CTIS theoretical framework, which both scholars (De Stutter & Lefer 2020: 4-5) perceive as "under-developed." The reason for this is that:

> Instead of empirical research in translation studies giving rise to the falsification, verification or adaptation of the hypotheses of universal features of translation, as initially intended by Baker, universal features have been used repeatedly and uncritically to 'explain' specific patterns observed in the corpus data. In the process, translation universals have gradually lost power in that they have only been used as fixed, passe-partout post hoc explanations: whatever linguistic phenomenon is being studied, there

> will always be some translation universal available which can be used to rationalise the descriptive patterns uncovered in the data.

(3) There is a lack of sophistication (i.e. a basic, often incorrect, use of statistics) in CTIS research designs, which end up nurturing reductionist (monofactorial) studies. The result of this is that many CTIS studies often rely on a single explanatory factor (notably the distinction between translated and non-translated texts) to characterise the process and product of translation (and interpreting). De Sutter & Lefer (2020: 5) quote Gries (2018: 295) to boldly declare that:

> monofactorial observational studies have virtually nothing to contribute to corpus linguistics [because] (i) no phenomenon is monofactorial and (ii) even if one had a new monofactorial hypothesis of a phenomenon, it would still require multifactorial testing to determine (a) whether it either adds anything to what we already know about the phenomenon (by statistically controlling for what we already know) or (b) whether it replaces (parts of) what we already know about the phenomenon.

(4) CTIS remain largely isolated and fixed on their object of study: translated/interpreted products and process

> without taking into account theoretical and methodological developments in other, related, fields such as corpus linguistics (including learner corpus research), variational linguistics, contrastive linguistics, sociolinguistics, psycholinguistics and cognitive linguistics, to name but a few (2020: 5).

In the authors' views, their proposed research agenda is not totally new in the field. It is already being upheld by works such as Delaere & De Sutter (2013), Hu et al. (2019) and Kruger (2019), to name but a few. Furthermore, it is disseminating quickly with the following declaration of intentions:

> understanding of the governing principles underlying translation, and the constraints under which it operates, can be achieved, in our view, by re-adopting and updating the essential aspect of Baker's research program, i.e., looking over the disciplinary fence and carefully selecting corpus-linguistic, ethnographic, sociological, and psycholinguistic methods that are appropriate for studying central aspects of translation, as well as interpreting research outcomes in an emerging, bottom-up translation theory that builds on theories in neighboring disciplines, such as contact linguistics, second language acquisition research and psycholinguistics (2020: 2).

Like De Sutter & Lefer, we believe that another revamping impulse to Baker (1993)'s original agenda seems more than fitting with the turn of yet another decade, where ravishingly new texts are being created, and new contexts are being populated. It seems to us this the only way we can continue to follow Tymoczko's (1998: 653) advice "to move from text-based questions to context-based questions."

Nevertheless, we are more cautious in our evaluation of risks. It is true that differences have been the crux of the matter in the field for some time now and that similarities are undoubtedly an under-researched part of the translational picture. It is, however, intriguing to see how, in certain quarters, the debate seems to be incipiently shifting from translation and interpreting (as the objects of study of independent disciplines) to a somewhat larger space populated by "similar" subgenres, those of constrained communication, within which (C)TIS might be losing the limelight that has been so difficult to occupy. While it is true that translation theory is enhanced by high quality empirical research, it is also true that theoretical advances cannot and will not just come from empiricist minds. Disregarding non-empiricist approaches to translation and interpreting (like those coming from so many valuable corners within (C)TIS) would be disowning our (C)TIS history. We would indeed do well to pursue more sophistication (in our hopefully increasingly refined statistics-based approaches). Nonetheless, it seems to be equally true that the complex phenomena of translation and interpreting cannot be fully grasped by these or those metrics only and that our past has already warned us against the false illusion of indisputable truth (Venuti 2000). Finally, when has (C)TIS ever been isolated? If there is one quality in our studies, this is that we have never ceased to "look over the disciplinary fence." We have been continuously inspired by literary criticism; informed by linguistic analysis; empowered by (post-modern) cultural, postcolonial, queer, gay studies; nurtured by psycholinguistic methods; complemented by sociological approaches, and so forth.

In fact, it is by looking yet once again over the disciplinary fence that we realise CTIS is not the only field where pausing and self-reflexivity is being called for. Neighbouring fields, such as Corpus-Assisted Discourse Analysis (CADS), are addressing very similar needs, as the brilliant monograph published by Taylor & Marchi (2018) testifies. This is not surprising

in the least. CTIS and CADS share similar goals when "us[ing] corpora to study how social realities are constructed, represented and transmitted linguistically" (2018: 1) (through monolingual or multilingual artefacts). They both share similar foundations (some of which can be found in corpus linguistics and discourse analysis). They have both grown exponentially over the past decade, with the emergence/reinforcement of conference series (e.g. the corpora and discourse series in CADS; the CILC conferences in CTIS) and the increased prominence of well-known publications. And when Taylor & Marchi (2018: 2) point at areas of concern for CADS (i.e. lack of standardisation, hindrance of institutional barriers, concerns about decontextualised data, epistemological issues...), they ring a loud bell within CTIS circles. It is then hardly surprising that, after the astounding success of recent years, both CTIS (with the leadership of De Sutter & Lefer, among others) and CADS (with Taylor & Marchi *inter alios*) are converging into the very same conclusion: "The time has come to pause and reflect on what it is we do." In Taylor & Marchi's (2018: 2) words:

> This is not intended as an assault of the exciting work which is emerging but a recognition of the maturity of the methodology, which is now robust enough to withstand, and indeed demand, close scrutiny.

According to Taylor & Marchi (2018), close scrutiny (i.e. a self-reflexive form of awareness) is intended to fight three inextricably connected potential types of *malaise*: partiality, dusty corners and blind spots. In other words, focusing on certain areas while disregarding others brings about an incompleteness (i.e. a partiality) in research that leaves a series of dusty corners or overlooked features (such as similarities or absence) and under-researched contents (minoritised topics, text-types and languages that are hidden by dominant voices). It also creates black holes of undetected or under-analysed components (blind spots) that can only be illuminated with innovative approaches, including those inspired by triangulation. If we think about it, this argument overlaps with what is advocated by De Sutter & Lefer. An (excessive?) concentration on the distinctive translational features leaves behind dusty corners with overlooked features (such as similarities) and under-researched contents (e.g. many of the mechanisms that shape

translation) and black holes in translational theory than can only be explored with higher levels of complex and sophisticated multi-methods.

In line with De Sutter & Lefer (2020) and Taylor & Marchi (2018) among others, this volume is born with the intention to encourage CTIS practitioners to pause and look around; to explore dusty corners and blind spots; to fight partiality, while injecting doses of innovation into our work. It is ultimately intended to boost critical thinking, (self-)awareness and (self)reflexivity, without renouncing to our past. Quite the contrary, while embracing what we conceive of as a glorious past. Always looking outwards (over the fence), always gazing forward (while peeping into the rear mirror). In this spirit, the volume is organised in such a way that it covers four main themes: *Translation Features as a Starting Point*; *Neglected and Overlooked Areas of Study; Researching Original and Translated Communication Produced under New Conditions* and *Self-Reflexivity*.

We set off the volume by tightening our unbreakable bondage with the past when taking up CTIS' most prolific research concern: translational features. In a descriptive-explanatory proposal ("Explicitation and Implication in Translation: Combining Comparable and Parallel Corpus Methodologies"), Miguel Ángel Jiménez-Crespo & Maribel Tercedor-Sánchez concentrate on one of the most widely researched of these features, explicitation. They do so through both comparable and parallel corpora. Along these lines, not only do they identify cases of explicitation (and implicitation) regarding the under-researched issue of Latin-Greek terms (LG) in medical texts, but they also delve into the potential causes behind this explicitation (i.e. cross-linguistic interference, risk-aversion…). This approach is in line with other recent empirical works on the topic (see, for example, Delaere & De Sutter 2013; Kruger 2019). In a nutshell, Jiménez-Crespo & Tercedor-Sánchez present us with the unique and complex case of LGs, where inter and intra-linguist translation processes converge, and where explicitation is intertwined with determinologisation.

The next four contributions revolve around the theme of what we could call *Neglected and overlooked areas of study*. Now we enter realms that are not frequently visited by corpus-based translation and interpreting scholars, such as subtitling, travel journalism, simultaneous interpreting and operatic audio description. Stepping on the (partially) untrodden is gradually

taking on in CTIS. When entering these realms, all four articles look over the disciplinary fence (some do so on more than one occasion) in order to find adequate (multi)-methods with a view to analysing their neglected and overlooked objects of study.

Blanca Arias' "Using Corpus Pattern Analysis for the Study of Audiovisual Translation. A Case Study to Illustrate Advantages and Limitations", for example, turns to the Theory of Norms and Exploitations (from Television Studies) and the application of Corpus Pattern Analysis (CPA) (from lexical studies) in order to perform an analysis into creative (anomalous or non-canonical) collocations. The corpus chosen for her exploration consists of original English and translated Spanish subtitles from the first two episodes of the first season of television series *Castle* (2009), *Dexter* (2006) and *The Mentalist* (2008). Studies of creativity (or exploitations in Arias' terminology) such as this are certainly under-represented. They are particularly welcome when they conclude with the critical assessment of the advantages and disadvantages of importing (CPA) methodology.

Collocations are also the topic of the article by David Finbar Brett, Barbara Loranc, Antonio Pinna ("A Corpus-Driven Analysis of Adjective+Noun Collocations in Travel Journalism in English, Italian and Polish"), in which travel journalism is explored through the corpus-driven lens. With the aid of three comparable corpora of travel reportage in English, Italian and Polish, the authors explore a variety of issues: (a) similarities and differences in the frequencies of adjective/noun collocations across languages; (b) similarities and in the denotative effects of some of the most frequent adjective/noun collocations; (c) syntactic variability of the collocates; (d) connectivity of some of the most productive collocates; and (e) collocational behaviour within selected themes. Among the paper's innovations, the cross-linguistic exploration of collocations in travel writing (certainly an under-researched topic of analysis) immediately springs to mind. Nevertheless, there are other sources of methodological innovation. Firstly, imported tools from the Social Network Analysis (such as Gephi) are used to reinforce the notion of collocational connectivity (which is built upon the concept of relational networks). Secondly, automatised protocols of compilation and annotation (with tailor-made Perl scripts created by the authors themselves) are a necessary step towards those higher levels of (programming) sophistication requested

by De Sutter & Lefer (2020) for their new agenda. We believe greater programming autonomy by CTIS researchers is an increasingly unavoidable requirement in the field.

Collocation and methodological sophistication are yet again the main ingredients of Marta Kajzer-Wietrzny & Łukasz Grabowski's "Formulaicity in Constrained Communication: An Intermodal Approach." Here collocation revolves around the notion of the bigram. For its part, sophistication is of a different type to that of the previous contribution and results from proficient knowledge of statistics. Mastering increasingly complex statistics (like being autonomous in programming, as shown in the previous article) is also becoming an unavoidable requirement for present and future CTIS research. In short, this article fits a Poisson regression model with fixed and random effects to dissect the use of bigrams in subcorpora from the intermodal EPTIC corpus. The result is a comparative study of three types of parliamentary speeches in English: translated speeches from Polish originals; interpreted speeches from Polish originals and original speeches produced by Polish natives for whom English is a second language. The neglected area of simultaneous interpreting and the increasingly widespread interest in constrained communication are particularly suitable for the present volume. The multidimensional nature of the study is also worth noting. In De Sutter & Lefer's (2020) manner, this article goes beyond the monofactorial explanation of the communicative phenomena under analysis to assess causation along five dimensions: (1) (mono-bilingual) language activation; (2) (spoken, written-multimodal) modality and register; (3) (unmediated/mediated) text production; (4) (native, non-native) proficiency; and (5) (high/low) task expertise. Finally, the paper resorts to a set of under-used tools within CTIS (such as Formulib, R or ad-hoc scripts in Python) to complement our basic toolkit.

"The Hierarchization of the Operatic Signs through the Lens of Audio Description: A Corpus Study" by Irene Hermosa-Ramírez also deals with a variety of very dusty corners that certainly benefit from the exploration with corpus-based tools. CTIS has not often been to the opera, nor have we indulged in examining operatic audio description. Spotting similarities and differences between productions from two famous opera houses (the Teatro Real, in Madrid, and the Liceu, in Barcelona) has not been a common aim

for the field. Complementary semiotic analyses do not usually come hand in hand with well-established corpus-based instruments. All this dusting is precisely what Hermosa-Ramírez does in her article. She compiles two corpora with audio descriptions of three famous operatic productions (*Aida*, *The Magic Flute* and *Carmen*) as represented at the Teatro Real and the Liceu and, with the aid of Sketch Engine, she compares mean sentence length, open class word frequencies, POS distributions and TTRs. Having done this, she complements her results with a semiotic analysis based on Rędzioch-Korkuz (2016) and TRACCE narratology tagset. Furthermore, she shows that triangulation is not the only possibility when using mixed-method designs. Methodological complementation is a powerful way to fight partiality.

The theme of *Researching Original and Translated Communication Produced under New Conditions* strongly links the next two articles, dealing with topics that were mostly unheard of in the 1990s.

In "Los estudios de corpus y la localización: una propuesta de análisis para material interactivo / Corpus studies and localization: a research proposal for interactive material", by Laura Mejías Climent, English original and Spanish translated dubbing synchronies of three video games (*Batman: Arkham Knight, Assassin's Creed Syndicate* and *Rise of the Tomb Raider*) are the main focus of the study. This topic in itself is already pretty innovative. Triangulation of data from qualitative methods, quantitative results and professional knowledge drawn from semi-structured interviews adds complexity to the standpoints from which the study is performed. The theoretical inspiration from localisation, audiovisual translation and corpus-based studies generates synergies that result in some sophisticated components. For example, multimodality here incorporates not just audiovisual stimuli but also tactile inputs. Moreover, the unit of analysis distances itself from the most traditional understanding of original and translated text and concentrates on "game situations." This generates the most innovative aspect of this article's already innovative proposal. Original and translated texts do not pre-date this research but are both a result and a means of it. In order for the researcher to research, she has to play the games and compare her moves. Setting out to study inexistent text was not a common goal during the 1990s. However, if we come to think about it, it is particularly characteristic of our

current "liquid" times (Bauman 2000), where the materiality of virtual/real, existent/inexistent texts is difficult to grasp.

Leticia Moreno-Pérez & Belén López-Arroyo, for their part, have an equally updated focus (i.e., a writing generator and its use for the translating profession) in "A Typical Corpus-Based Tool to the Rescue: How a Writing Generator Can Help Translators Adapt to the Demand of the Market". As is well known, nowadays a writing generator is a tool that helps non-native speakers to generate text (within specialised genres) in a foreign language. Developing a writing generator is an undoubtedly sophisticated task that enriches our current CTIS toolkit. It relies on file managers, tagger builders and taggers that produce information on prototypical rhetorical structures (moves and steps), lexicon-grammatical patterns and terminological and phraseological glossaries. Its application to translation within the field of oenology shows CTIS is not only able to look over the disciplinary fence but also (and mainly) to listen to the demands of the market, narrowing the breach between academia and the profession.

Last but certainly not least, the volume concludes with *Self-Reflexivity*, the ultimate goal of this venture. Self-reflexivity can be of two types. The first type is that of individual self-reflexivity, in which researchers analyse their prior/contemporary production to reinforce their theoretical and methodological standing. The second type of self-reflexivity is collective and disciplinary-bound; it looks over the field fence while focusing on the domain kept within that fence.

"Autocrítica de publicaciones previas basadas en corpus: análisis DAFO / Self-Criticism of Previous Corpus-Based Publications: SWOT Analysis", by Alexandra Santamaría Urbieta & Elena Alcalde Peñalver, is an example of the first type of self-reflexivity. The authors review four of their own joint publications in which corpus-based methods are applied. For this purpose, they resort to the SWOT (Strengths, Weaknesses, Opportunities and Threats) methodology, which is particularly prolific in businesses and marketing studies and which provides a useful scaffolding for self-criticism and self-awareness. We believe this is a good way for CTIS partitioners to contribute to the revamping of Baker's (1993) agenda.

And it is precisely Mona Baker who is a powerful inspiring voice in the concluding contribution to the volume, "From Text to Data: Mediality in

Corpus-based Translation Studies", by Jan Buts & Henry Jones. This final article is produced within the Genealogies of Knowledge (GoK) project framework, where Baker (a Professor Emeritus at present) has a prominent role. Indeed, Buts' & Jones' proposal is an example of the second type of self-reflexivity, whereby scholars dare to ponder on fundamental pillars of the discipline precisely as the only manner to make it stronger. One of such pillars within CTIS is the electronic media "in and through which translations are stored, transmitted and – by extension – studied (Armstrong 2020; Pérez-González 2014)" (see contribution in this volume). These media are not mere containers that serve to preserve and convey meaning and knowledge in an untainted manner. On the contrary, they are transformative environments that deeply affect our relation with (and understanding of) texts. On the whole, CTIS seems to have pushed aside issues of mediality, which have been confined to some of our dustiest corners. Yes, as Buts & Jones argue, the limitations and restrictions of technologies have been discussed before. However, as the authors advocate: "[w]hatever the cause, the convertibility of the sign and its attachment to the binary standard are yet to be consistently questioned." Being aware that the electronic media upon which CTIS depends and the software tools with which CTIS performs its analyses lead to the application of certain (pattern-seeking) methodologies rather than others (more focused on structures and narratives) is the first step to combatting partiality. Designing new visualisation tools and critically applying them to the examination of political and scientific concepts — as Buts & Jones do as part within the GoK project — seems to us a powerful initiative to contribute to CTIS new agenda.

In the dawn of the second decade of the 21st century, when in many parts of the world citizens in lockdowns hold their breath without a clear idea of what to do next, a need to stop and think critically is indeed more pressing than ever.

References

BAKER, Mona. (1993) "Corpus Linguistics and Translation Studies — Implications and Applications." In: Baker, Mona; Gill Francis & Elena Tognini-Bonelli (eds.) 1993. *Text and Technology*. Amsterdam: John Benjamins, pp. 233-50.

BAKER, Mona. (1998) "Réexplorer la langue de la traduction: une approche par corpus." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 480-485.

BAKER, Mona. (2004) "A Corpus-Based View of Similarity and Difference in Translation." *International Journal of Corpus Linguistics* 9:2, pp.167-93.

BAÑOS, Rocío; Silvia BRUTI & Serenella ZANOTTI. (2013) "Corpus Linguistics and Audiovisual Translation: In Search of an Integrated Approach." *Perspectives* 21:4, pp. 483-90.

BAUMAN, Zygmunt. (2000) *Liquid Modernity*. Cambridge: Polity Press.

BERNARDINI, Silvia & Dorothy KENNY. (2020) "Corpora." In: Baker, Mona & Gabriela Saldanha (eds.) 2020. *Routledge Encyclopedia of Translation Studies*. Third edition. London & New York: Routledge, pp. 112-15.

BOWKER, Lynne. (1998) "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 631-651.

DE SUTTER, Gert & Marie-Aude LEFER. (2020) "On the Need for a New Research Agenda for Corpus-Based Translation Studies: A Multi-Methodological, Multifactorial and Interdisciplinary Approach." *Perspectives* 28:1, pp. 1-23.

DELAERE, Isabelle & Gert DE SUTTER. (2013) "Applying a Multidimensional, Register-Sensitive Approach to Visualize Normalization in Translated and Non-Translated Dutch." *Belgian Journal of Linguistics* 27, pp. 43-60.

EBELING, Jarle. (1998) "Contrastive Linguistics, Translation, and Parallel Corpora." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 602-615.

HALVERSON, Sandra. (1998) "Translation Studies and Representative Corpora: Establishing Links between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 494-514.

HU, Kaibao. (2016) *Introducing Corpus-Based Translation Studies. New Frontiers in Translation Studies*. Berlin & Heidelberg: Springer Berlin Heidelberg.

Hu, Kaibao & Qing Tao. (2014) "The Chinese-English Conference Interpreting Corpus: Uses and Limitations." *Meta* 58:3, pp. 626-42.

Hu, Xianyao; Richard Xiao & Andrew Hardie. (2019) "How Do English Translations Differ from Non-Translated English Writings? A Multi-Feature Statistical Model for Linguistic Variation Analysis." *Corpus Linguistics and Linguistic Theory* 15:2, pp. 347-82.

Jiménez-Crespo, Miguel A. (2015) "Translation Quality, Use and Dissemination in an Internet Era: Using Single-Translation and Multi-Translation Parallel Corpora to Research Translation Quality on the Web." *The Journal of Specialised Translation* 23, pp. 39-63.

Kruger, Haidee. (2019) "*That* Again: A Multivariate Analysis of the Factors Conditioning Syntactic Explicitness in Translated English." *Across Languages and Cultures* 20:1, pp. 1-33.

Laviosa, Sara. (1998a) "The English Comparable Corpus: A Resource and a Methodology." In: Bowker, Lynne; Michael Cronin; Dorothy Kenny & Jennifer Pearson (eds.) 1998. *Unity in Diversity? Current Trends in Translation Studies*. Manchester: St. Jerome, pp. 101-112.

Laviosa Sara. (1998b) "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 557-570.

Laviosa, Sara (ed.) (1998) *Special Issue: L'approche Basée Sur Le corpus/The Corpus Based Approach. Meta. Journal Des Traducteurs/Meta. Translators' Journal* 43:4.

Laviosa, Sara. (2002) *Corpus-Based Translation Studies: Theory, Findings, Applications*. Amsterdam: Rodopi.

Maia, Belinda. (1998) "Word Order and the First Person Singular in Portuguese and English." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 589-601.

Malmkjær, Kirsten. (1998) "Love thy Neighbour: Will Parallel Corpora Endear Linguistics to Translators?" In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 534-541.

Munday, Jeremy. (1998) "A Computer-Assisted Approach to the Analysis of Translation Shifts." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 542-556.

Olohan, Maeve. (2004) *Introducing Corpora in Translation Studies*. London & New York: Routledge.

OTHMAN, Waleed. (2020) "An SFL-based model for investigating explicitation-related phenomena in translation." In: Calzada Pérez, María & Jeremy Munday (eds.) *Meta: Journal des traducteurs* 65:1, pp. 193-210.

ØVERÅS, Linn. (1998) "In Search of the Third Code: An Investigation of Norms in Literary Translation." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 571-588.

RĘDZIOCH-KORKUZ, Anna. (2016) *Opera Surtitling as a Special Case of Audiovisual Translation*. Bern: Peter Lang.

SALDANHA, Gabriela. (2011) "Translator Style: Methodological Considerations." *The Translator* 17:1, pp. 25-50.

SHLESINGER, Miriam. (1998) "Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 486-493.

TAYLOR, Charlotte & Anna MARCHI (eds.) (2018) *Corpus Approaches to Discourse: A Critical Review*. Milton Park, Abingdon, Oxon & New York: Routledge.

TYMOCZKO, Maria. (1998) "Computerized Corpora and the Future of Translation Studies." *Meta* 43:4, pp. 652-60.

VÁRADI, Tamas. (2007) "NP Modification Structures in Parallel Corpora." In: Rogers, Margaret & Gunilla Anderman (eds.) 2007. *Incorporating Corpora. The Linguist and the Translator.* Clevedon: Multilingual Matters.

VENUTI, Lawrence. (2000) "¿Será Útil La Teoría de La Traducción Para Los Traductores?" *Vasos Comunicantes* 16, pp. 26-34.

XIAO, Richard. (2010) "How Different Is Translated Chinese from Native Chinese?: A Corpus-Based Study of Translation Universals." *International Journal of Corpus Linguistics* 15:1, pp. 5-35.

ZANETTIN, Federico. (1998) "Bilingual Comparable Corpora and the Training of Translators." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 616-630.

ZANETTIN, Federico. (2012) *Translation-Driven Corpora Corpus Resources for Descriptive and Applied Translation Studies*. Translation Practices Explained. Manchester & Kinderhook: St. Jerome Pub.

BIONOTE / NOTA BIOGRÁFICA

María Calzada Pérez is Full Professor of Translation Studies at the Universitat Jaume I. (Spain). Her research mainly focuses on corpus-based translation studies, institutional translation (especially at the European Parliament), ideology, and translator-teaching. She is Coordinator of the ECPC (European Comparable and Parallel Corpora of Parliamentary Speeches) research group. She has produced research, such as (i) *Transitivity in Translating: The Interdependence of Texture and Context* (Peter Lang, 2007); (ii) "Five Turns of the Screw. A CADS analysis of the European Parliament" (*Journal of Language and Politics* 16:3, 2017) (iii). "Corpus-based methods for Comparative Translation and Interpreting Studies" (*Translation and Interpreting Studies* 12:2, 2017); and (iv) "What is kept and what is lost without translation? A corpus-assisted discourse study of the European Parliament's original and translated English" (*Perspectives* 26:2, 2017). She is also editor of volumes like: *Apropos of Ideology* (St. Jerome, 2003).

Sara Laviosa is Associate Professor in English Language and Translation Studies at the Department of Lettere Lingue Arti Italianistica e Culture Comparate (LeLiA), Università degli Studi Aldo Moro, Italy. Her research interests are in corpus-based translation studies and pedagogic translation. She is the author of *Corpus-based Translation Studies* (Rodopi/Brill, 2002), *Translation and Language Education* (Routledge, 2014) and *Linking Wor(l)ds* (Liguori, 2020). She is the co-author (with Adriana Pagano, Hannu Kemppanen & Meng Ji) of *Textual and Contextual Analysis in Empirical Translation Studies* (Springer, 2017). She is the guest editor of *L'Approche Basée sur le Corpus/The Corpus-based Approach* (Special Issue of *Meta* 43:4, 1998), *Translation in the Language Classroom* (Special Issue of *The Interpreter and Translator Trainer* 8:1, 2014). She is the co-editor (with Maria González-Davies) of *The Routledge Handbook of Translation and Education* (Routledge, 2020) and (with Meng Ji) of *The Oxford Handbook of Translation and Social Practices* (OUP, 2020). She is the Founder and Editor of the journal *Translation and Translanguaging in Multilingual Contexts* (John Benjamins).

Recibido / Received: 08/02/2021

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.01

Para citar este artículo / To cite this article:
Calzada Pérez, María & Sara Laviosa. (2021) "Un cuarto de siglo después: Tiempo para reflexionar sobre una nueva agenda de los ETBS." In: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection*. *MonTI* 13, pp. 33-61.

UN CUARTO DE SIGLO DESPUÉS: TIEMPO PARA REFLEXIONAR SOBRE UNA NUEVA AGENDA DE LOS ETBS[1]

María Calzada Pérez
calzada@uji.es
Universitat Jaume I

Sara Laviosa
sara.laviosa@uniba.it
University of Bari Aldo Moro

Abstract

The introduction of corpora in descriptive and applied translation and interpreting studies goes back to the 1990s, when the corpus linguistic approach was making considerable progress in descriptive and applied language studies. Twenty-five years on, Corpus-Based Translation and Interpreting Studies (CTIS) is a well-established field of interdisciplinary research worldwide. Its growth goes hand in hand with technological advancements, which make it possible to design, create and share monolingual and multilingual spoken, written and multimodal corpora as resources for theoretical, descriptive and applied research in both translation and interpreting studies. We believe this is the right time to pause and reflect on the achievements and criticalities of this variegated area of scholarship and practice in order to look to the future with renewed confidence and awareness of the challenges that lie ahead.

1. Este artículo se ha llevado a cabo en el marco del proyecto de investigación *Representaciones originales, traducidas e interpretadas de la(s) crisis de refugiados: triangulación metodológica desde el análisis del discurso basado en corpus (RE-CRI)*, financiado por el Ministerio de Ciencia e Innovación (PID2019-108866RB-I00).

Keywords: Corpus-based Translation Studies; Corpus-based Interpreting Studies; Corpus Linguistics; Corpus-assisted Discourse Analysis.

Resumen

La introducción de corpus en los estudios descriptivos y aplicados de la traducción e interpretación se remonta a la década de 1990. Tras 25 años, los estudios de traducción e interpretación basados en corpus (ETBC) son un campo de investigación interdisciplinar bien establecido en todo el mundo cuyo crecimiento va de la mano de avances tecnológicos que permiten diseñar, crear y compartir corpus monolingües y multilingües (orales, escritos y multimodales) como recursos para la investigación teórica, descriptiva y aplicada en los estudios de traducción e interpretación. Creemos que es el momento oportuno de hacer una pausa y reflexionar sobre los logros y las carencias de este variado ámbito de la erudición y la práctica con el fin de mirar al futuro con renovada confianza y conciencia de los retos que nos esperan.

Palabras clave: Estudios de traducción basados en corpus; Estudios de interpretación basados en corpus; Lingüística de corpus; Análisis del Discurso Asistido por Corpus; Lingüística contrastiva.

1. Los orígenes de los estudios de traducción e interpretación basados en corpus

Los estudios de traducción e interpretación basados en corpus (ETBC) designan un área de investigación que adopta y desarrolla metodologías de la lingüística de corpus (LC) para analizar traducciones con finalidades teóricas, descriptivas y aplicadas. A su vez, la LC es un enfoque de los estudios lingüísticos basado en el análisis de corpus (colecciones de textos reales en formato electrónico compilados en función de criterios específicos). En este artículo, ofrecemos un seguimiento del desarrollo de los ETBC desde su aparición en la década de 1990 hasta el presente y destacamos los logros principales en distintos ámbitos de investigación con el objetivo de ofrecer un contexto al conjunto de artículos seleccionados en este volumen de *MonTI*.

Tras el trabajo seminal de Mona Baker (1993), en el que se perfila una hoja de ruta para un enfoque de los estudios de traducción e interpretación basados en corpus, en 1998 se publica el primer volumen dedicado a la investigación en esta área en un número especial de *Meta* titulado *L'Approchee*

Basée sur le Corpus/The Corpus-Based Approach, editado por Sara Laviosa. Los artículos de este número se agrupaban en torno a dos temas fundamentales: "Investigación teórica" y "Estudios empíricos y didácticos". En el primer bloque se exponía el objetivo, el objeto de estudio y la metodología del enfoque emergente basado en corpus. El segundo grupo consistía en estudios empíricos y didácticos de la traducción, entendida como proceso y como producto. El artículo final del número, a cargo de Maria Tymoczko, partía del conocimiento generado por estos trabajos y destacaba que la investigación en traducción mediante corpus tenía, como finalidad, "to address not simply questions of language and linguistics, but also questions of culture, ideology, and literary criticism" (1998: 653).

Los ETBC se consideran centrales en el conjunto de la disciplina de los estudios de traducción, porque (1) nacen con la voluntad de integrar los estudios lingüísticos y culturales y explorar la relación entre ambos, (2) revelan una actitud consciente sobre el efecto de la ideología en la teoría, la práctica y la didáctica de la traducción y (3) tienen como objetivo adaptar las nuevas tecnologías para avanzar en la teoría, la investigación empírica y la práctica en beneficio de la formación de traductores y la práctica profesional. Por este motivo, Tymoczko (1998: 658) potencia una visión de los ETBC que ofrece "the opportunity to reengage the theoretical and pragmatic branches of Translation Studies, branches which over and over again tend to disassociate, developing slippage and even gulfs". De hecho, estos primeros estudios basados en corpus ya ilustraban algunas de las principales líneas de investigación que, como apuntaba Tymoczko, se desarrollarían en los años siguientes, de las cuales este trabajo propone una revisión crítica que profundice en cómo la investigación académica ha potenciado la comprensión de los estudios de traducción y la labor traductora hasta la actualidad y cómo lo hará en el futuro.

A finales de la década de 1990 identificamos tres ámbitos de investigación. Cada uno de ellos se refiere a aspectos y temáticas que se encuentran bajo el paraguas de una de las tres ramas de la disciplina entendida como un todo: los estudios teóricos, los estudios descriptivos y los estudios aplicados. En lo que concierne a la teoría, el artículo de Mona Baker de 1998 aborda los fundamentos y los objetivos de la investigación en traducción como proceso y como producto mediante corpus que profundicen en aquella primera

propuesta de investigación que ella misma había puesto sobre la mesa cinco años antes. Baker trata la necesidad de desarrollar una metodología basada en corpus que sea coherente y que identifique los distintos aspectos del lenguaje de la traducción. Sostiene, además, que el objetivo de este esfuerzo investigador, que se basa en los estudios de los académicos que trabajan en enfoques descriptivos y orientados al producto, no es únicamente revelar la naturaleza del "tercer código" *per se*, sino comprender las restricciones específicas, las presiones y las motivaciones que influyen en la tarea de traducir entendida como actividad comunicativa. Estas consideraciones constituyen una suerte de eco del análisis de Tymoczko a propósito de los ETBC como un enfoque cuyos "[m]odes of interrogation – as well as care in the encoding of metatextual information about translations and texts – allow researchers to move from text-based questions to context-based questions" (1998: 653).

De forma similar, Miriam Shlesinger postula que, así como la traducción es una acción comunicativa determinada por sus propios objetivos, presiones y contexto de producción, lo mismo sucede con la interpretación, entendida como "the production of oral output based on other-language input which may be either written (to be read) or unwritten (impromptu)" (1998: 486-487). Shlesinger propone, pues, extender la noción de corpus monolingüe comparable de Baker (que consiste en un corpus de textos originales en una lengua y un corpus de textos comparables traducidos en esa misma lengua) para que incluya grupos de textos diferentes en la misma lengua: textos orales interpretados de diversas lenguas, textos orales originales producidos con características similares y traducciones escritas de textos orales realizados en circunstancias análogas. Este diseño innovador permitiría no solo el estudio de textos interpretados como manifestaciones distintas del discurso oral, sino también la identificación de patrones lingüísticos regulares de uso que diferencian la interpretación de la traducción escrita. Además, Shlesinger propone adaptar el diseño tradicional de corpus paralelo unidireccional para que incluya tres tipos de textos: textos en lengua original, sus versiones interpretadas y sus versiones traducidas escritas. Para ella, la ventaja concreta de este tipo de corpus radica en que permiten investigar aspectos específicos de la lengua y de la combinación y dirección lingüística de los textos interpretados, sin olvidar otras posibles variables como el género, la experiencia profesional o el *background* lingüístico del intérprete. A partir

de estas consideraciones, Shlesinger prevé que los estudios de interpretación basados en corpus ayudarán a seguir analizando los aspectos comunes que comparten la mediación escrita y la oral, así como a definir aquellos aspectos que permiten diferenciar la interpretación de la traducción (1998: 490-491).

En lo que respecta a corpus paralelos, Kirsten Malmkjær (1998) explica las ventajas de usar estas fuentes tanto en estudios contrastivos como en estudios de traducción propiamente dichos. En cuanto a la lingüística contrastiva, los corpus paralelos son muy útiles para investigar las diferencias y similitudes de la lengua en uso. Para los investigadores en estudios de traducción, son igualmente muy valiosos para identificar normas de traducción pese a dos problemas básicos que la propia Malmkjær se encarga de presentar. El primer problema son las líneas de concordancias KWIC, que no siempre ofrecen un contexto lingüístico suficiente para investigar características de los textos. Además, se corre el riesgo de que algunos aspectos del comportamiento de las personas que traducen queden de manifiesto, mientras que otros puedan pasar desapercibidos. La segunda dificultad está relacionada con el modo en que los corpus paralelos están diseñados para incluir únicamente una traducción para cada texto original, cosa que podría ocultar un aspecto importante del fenómeno translatorio, esto es, las diferencias que existen entre las distintas traducciones de un mismo original. Para paliar estas deficiencias, Malmkjær sugiere complementar los estudios orientados a identificar normas, que requieren de una gran cantidad de textos, con un corpus más reducido, pero con una mayor elaboración que consista en un texto en lengua original y tantas traducciones como sea posible, de manera que puedan llevarse a cabo investigaciones de mayor calado. La combinación de estos dos tipos de corpus representa dos ventajas. La primera es que los resultados pueden ser más completos, y la segunda es que pueden tener mayor rigor, ya que un corpus más extenso permite una comparación mejor que un número inferior de casos analizados en un corpus menor. Malmkjær considera que esta metodología atiende tanto a las necesidades de la lingüística contrastiva como a las de los estudios de traducción e interpretación, uniéndolos en una relación de cooperación mutua y potenciando sinergias con otras disciplinas colindantes que, de algún modo, siempre han formado parte de la dimensión aplicada de los estudios de la traducción.

Al tiempo que Baker and Shlesinger exponen los fundamentos para investigar la traducción y la interpretación mediante corpus y Malmkjær propone perfeccionar la metodología basada en corpus para plantear preguntas relevantes en el marco de los estudios contrastivos y los estudios de traducción, Sandra Halverson (1998) analiza el tema de la representatividad en el diseño de un corpus general de traducción y proporciona un marco teórico coherente en el que los datos y la metodología conforman un todo que garantiza la comparación de los resultados empíricos. Con este objetivo en mente, Halverson propone una conceptualización prototípica de la categoría objeto opuesta a la clásica. Bajo este enfoque, la población objeto de estudio deviene una categoría prototípica cuyo centro lo ocupan traducciones profesionales (en culturas de países occidentales industrializados, esto es), mientras que las posiciones periféricas estarían ocupadas por grupos de categorías diferentes como, por ejemplo, aquellas traducciones que puedan proporcionar estudiantes de traducción, traductores no profesionales o traducciones hacia una lengua que no es la lengua materna de quien traduce. La relación entre el centro y la periferia dentro de ese prototipo no sería la de la inclusión o exclusión de los elementos que pertenecen a la categoría, sino la de similitud. Además, los límites de los distintos grupos de traducciones no serían impermeables, lo que, para quien investiga, implicaría que, con el objetivo de asegurar la representatividad del corpus, este debería estar formado por muestras de poblaciones diferentes mediante subcorpus con diferentes grados de significancia. Teniendo en cuenta que los prototipos están, por definición, vinculados a la cultura, las conclusiones de este tipo de estudios no podrían generalizarse más allá de la población específica que representa un corpus determinado.

En resumen, estas reflexiones teóricas iniciales se centran en el porqué, el qué y el cómo estudiamos la traducción y la interpretación mediante corpus. Nos centramos ahora en la revisión de los resultados obtenidos por los estudios publicados en el volumen especial de *Meta* de 1998. La investigación de Sara Laviosa (1998b) sobre el texto inglés traducido se basa en un subcorpus del English Comparable Corpus (ECC). Dicho corpus comprende dos colecciones de prosa narrativa en inglés: uno hecho a partir de traducciones de distintas lenguas (fundamentalmente lenguas románicas), y uno de textos originales en inglés del mismo período. El estudio revela

cuatro patrones de selección léxica en los textos traducidos en contraste con los textos originales: una proporción menor (relativa) de palabras léxicas frente a palabras gramaticales, cierta proporción menor de palabras de alta frecuencia frente a palabras de baja frecuencia, una mayor repetición de palabras más frecuentes y una menor variedad en las palabras usadas con más frecuencia. Laviosa (1998a) propone denominar "*core patterns of lexical use*" a estas regularidades de los textos en inglés traducidos para indicar que, dada su presencia tanto en el subcorpus de prosa periodística como en el de prosa narrativa, pueden resultar típicos del inglés traducido en general.

Sin dejar de lado las características regulares del lenguaje traducido, Linn Øverås (1998) se centra en la explicitación. Para ello, utiliza dos subcorpus de textos de ficción en noruego y en inglés del English-Norwegian Parallel Corpus (ENPC) y parte de la hipótesis de que aumentaría el nivel de cohesión léxica y gramatical cuando se traduce del noruego al inglés y viceversa. La comparación de la distribución de los cambios de explicitación e implicitación en los textos revela una tendencia general a explicitar tanto en el inglés traducido como en el noruego traducido, a pesar de haber encontrado un nivel más bajo de explicitación en las traducciones del noruego al inglés. Øverås encuentra dos tipologías en el aumento de la cohesión: la adición y la especificación. La adición conlleva una inserción de ítems gramaticales y léxicos que no están presentes en el texto de partida. La especificación comporta la expansión o substitución de los ítems gramaticales y léxicos que no están presentes en el texto de partida. Øverås tiene en cuenta diversos factores que pueden dar cuenta de la explicitación, entre los que se incluyen las preferencias estilísticas de la lengua de partida y de llegada, los cambios obligatorios resultado de las reglas de la lengua de llegada o las normas culturales y restricciones inherentes de la traducción entendida como acto de comunicación. Asimismo, sugiere una relación entre la explicitación y la neutralización (la tendencia a utilizar colocaciones comunes no marcadas o símiles en lugar de metáforas) porque ambas tienen el efecto de conseguir una mayor legibilidad del texto de llegada. Al hacerlo, de forma implícita subraya la relación intrínseca entre diferentes patrones regulares postulados del "tercer código".

Si pasamos del estudio de patrones al estudio de cambios (*shifts*), Jeremy Munday (1998) nos presenta los resultados iniciales obtenidos tras análisis

de la traducción *Seventeen Poisoned Englishmen* que Edith Grossman hace del cuento de García Márquez *Diecisiete ingleses envenenados*. Munday recurre a una variedad de métodos básicos de análisis lingüístico de corpus (listas de frecuencia de palabras, estadística descriptiva y concordancias) para examinar los textos de forma inductiva. Las listas de frecuencia de palabras se obtienen en primera instancia de los textos en lengua de partida y lengua de llegada y luego se comparan para identificar áreas de investigación interesantes. Munday utiliza textos intercalados, esto es, introduce de forma manual los textos traducidos entre las líneas del texto original. Posteriormente, establece concordancias de este texto intercalado y las usa para llevar a cabo un análisis comparativo contextualizado de todos los ítems de selección léxica con el objetivo de examinar los cambios que se han ido acumulando sobre el texto en su conjunto como resultado de las elecciones llevadas a cabo por la traductora. Dicho análisis se efectúa para comprender el proceso de toma de decisiones que apuntala el producto final y para inferir las normas lingüístico-textuales de la traducción. El enfoque de Munday es, por tanto, descriptivo, orientado al proceso y al producto y basado en datos, estableciendo hipótesis a partir de la observación de las diferencias que se dan en las listas paralelas durante la construcción manual del texto intercalado. A partir de estas hipótesis iniciales de frecuencia lleva a cabo su investigación con la ayuda de métodos de análisis automáticos adicionales como el alineamiento de concordancias. El análisis de las primeras 800 palabras de su corpus paralelo revela cambios en la cohesión y el orden de palabras que tienen lugar en todo el texto y cuyo efecto es el de desplazar el punto de vista narrativo de la primera a la tercera persona, con lo cual se crea una distancia entre el lector y los pensamientos, experiencias y sentimientos del protagonista principal del cuento.

Desde la lingüística contrastiva, Belinda Maia (1998) analiza la frecuencia y la naturaleza de la estructura oracional SVO en inglés y en portugués, especialmente aquellos casos en los que el sujeto aparece como el pronombre de primera persona *I* y *eu*, respectivamente, o un sustantivo. El corpus que analiza es un corpus paralelo bidireccional reducido que comprende una novela en inglés con su traducción al portugués y una novela portuguesa con su correspondiente traducción al inglés. Maia considera que el diseño de este corpus es apropiado para comparar cómo una misma situación se

representa en las dos lenguas en el subcorpus paralelo, mientras que el subcorpus comparable bilingüe permite comparaciones adicionales entre las lenguas originales y entre las variedades traducidas y no traducidas dentro de una misma lengua. Las discrepancias que observa en las frecuencias del sujeto persona (ya sea como nombre o pronombre) sugieren que, al contrario que en el uso habitual en inglés, la ausencia de sujeto en la estructura oracional V+O es la norma en el original portugués, mientras que el portugués traducido denota la influencia de las normas del inglés. Es más, mientras que el uso de *I* es sintácticamente necesario en inglés, la ocurrencia *eu* equivalente del portugués parece estar relacionada con factores pragmáticos como la tematización, la topicalización y el énfasis, en tanto que el verbo actúa como tema normal en un alto índice de frases en portugués. A partir de estos resultados, Maia concluye que la flexibilidad del orden de palabras y una variabilidad más amplia de tematización en portugués en relación con el inglés permiten una comunicación más sutil.

Al igual que Maia, Jarle Ebeling (1998) considera los corpus paralelos como una fuente adecuada de datos para la investigación de las diferencias y similitudes entre lenguas, y adopta la noción de equivalencia traductora como metodología para el análisis contrastivo. Ebeling utiliza el ENPC, un corpus paralelo bidireccional de textos en noruego y en inglés, para examinar las construcciones presentacionales con *there* y las construcciones con *det* equivalentes en noruego, tanto en inglés original y traducido como en noruego original y traducido, respectivamente. El corpus en inglés original revela que *be* es el verbo más frecuente, mientras que el noruego permite la aparición de una gran variedad de verbos, algunos en voz pasiva. El análisis de Ebeling de las equivalencias traductoras de las construcciones *there be* del inglés revela la influencia de la lengua de llegada. Constata que los traductores amplían el rango de las construcciones con *det* mediante el uso de a) otros verbos más allá de los existenciales, b) existenciales con *ha* (que se corresponden a las existenciales con *have* en inglés, y c) construcciones con *det* mediante frases pasivas. Desde el punto de vista de Ebeling, esta variabilidad otorga al significado expresado en la traducción una mayor especificidad, haciéndolo más informativo que si lo comparamos con el original. Sin embargo, las traducciones al inglés de las construcciones con *det* con verbos léxicos en la voz activa se transforman en construcciones con

there be, lo que lleva a una menor especificación. Estos resultados confirman, en parte, las predicciones planteadas a partir de la evidencia ofrecida por el análisis de los corpus originales y arroja luz en la relación de equivalencia asumida entre las dos estructuras en inglés y en noruego, respectivamente.

En lo concerniente a la rama aplicada de los estudios de corpus, las investigaciones llevadas a cabo por Federico Zanettin (1998) y Lynne Bowker (1998) son de interés particular para la formación de traductores. Zanettin demuestra cómo los corpus comparables bilingües reducidos son útiles para analizar aspectos estilísticos de un género textual determinado gracias a la comparación de palabras y frases que tienen, tanto una similitud formal muy elevada (como los nombres propios y los cognados) como aquellas que se basan en equivalentes de traducción lexicográficos. Zanettin ofrece ejemplos de búsquedas llevadas a cabo en el aula con un corpus comparable en italiano e inglés formado por textos periodísticos de rotativos destacados. Por ejemplo, el modo en que se habla del presidente François Mitterand en las dos lenguas presenta diferencias interesantes: *François Mitterand* o simplemente *Mitterand* es el uso más frecuente en italiano, mientras que el inglés prefiere *President Mitterand*, *President François Mitterand* o *Mr Mitterand*. Además, algunos verbos equivalentes que se usan normalmente para introducir el discurso directo o el discurso indirecto presentan frecuencias distintas en las dos lenguas en los perfiles sintácticos y colocacionales. Incluso cognados como *prezzi* y *prices* muestran diferentes patrones colocacionales y coligacionales. Estas incursiones didácticas basadas en datos son valiosas en tanto en cuanto permiten mejorar el conocimiento contrastivo de las lenguas de partida y de llegada y potencian las destrezas traductoras.

Abundando en esta perspectiva didáctica, Bowker aborda dos problemas importantes que afectan a la formación que se ofrece a los estudiantes para que se conviertan en traductores profesionales de ámbitos especializados. Una dificultad es la ocurrencia de errores terminológicos como resultado de un conocimiento deficiente de los lenguajes de especialidad. La otra es la ocurrencia de errores debidos a una falta de destreza en la redacción de textos especializados. El estudio piloto de Bowker presenta un experimento de traducción llevado a cabo con un grupo de estudiantes de cuarto curso de L1 inglés en la Dublin City University. Los participantes realizaron dos traducciones de textos originales en francés (L2) de dos fragmentos de textos

semiespecializados sobre escáneres ópticos. Una traducción se hizo con la ayuda de diccionarios bilingües y monolingües, así como con materiales de referencia no lexicográficos (manuales y catálogos). La otra traducción se hizo con un diccionario bilingüe y un corpus monolingüe especializado de artículos en inglés sobre escáneres ópticos de 1.4 millones de palabras, compilado a partir de *Computer Select* en CD-ROM. El programa informático utilizado para analizar el corpus fue WordSmith Tools. Los resultados revelan que las traducciones hechas con la ayuda del corpus fueron de calidad más elevada en cuanto al campo temático (*sensibilité aux nuances* se tradujo perfectamente como *whatever their sensitivity to colour*); selección terminológica (*vitre/glass paten* o *scan bed*); y expresiones idiomáticas (*photodiodes sensibles à la lumière/light-sensitive photodiodes* o *photosensitive diodes*). Bowker observa que, aunque no se constató una calidad superior en cuanto a la gramática o el registro, el uso de un corpus monolingüe especializado en lengua de llegada no se asociaba a una reescritura más pobre.

Hacer una revisión del pasado como la que hemos presentado nos permite apreciar aún más si cabe el valor de los inicios de este tipo de trabajos, que constituyeron el germen de una gran diversidad de líneas de investigación. Dichos trabajos, con la llegada del nuevo siglo, han contribuido a la presencia, consolidación y crecimiento de los estudios de traducción e interpretación basados en corpus en la rama pura y la aplicada de la disciplina, al tiempo que ha impulsado a la comunidad investigadora hacia el empirismo y la interdisciplinariedad.

2. ETBC: un cuarto de siglo después

Han pasado más de 25 años desde que Baker (1993) propusiera una hoja de ruta para la investigación mediante corpus en los estudios descriptivos de la traducción y se publicara el primer compendio de artículos sobre estudios de traducción e investigación basados en corpus (Laviosa 1998). A lo largo de las dos primeras décadas del siglo XXI, los Estudios de Traducción e Interpretación basados en Corpus (ETBC) han experimentado un crecimiento tal que, ya desde el inicio de los 2000, Baker (2004, 169) afirmó que se había hecho mucho más de lo que quedaba por hacer.

Durante las últimas décadas se han compilado y revisado ampliamente multitud de corpus (Hu 2016; Laviosa 2002; Olohan 2004, entre otros). Federico Zanettin (2012: 10) ofrece una imagen de esta proliferación desmesurada que presentamos en la Figura 1:

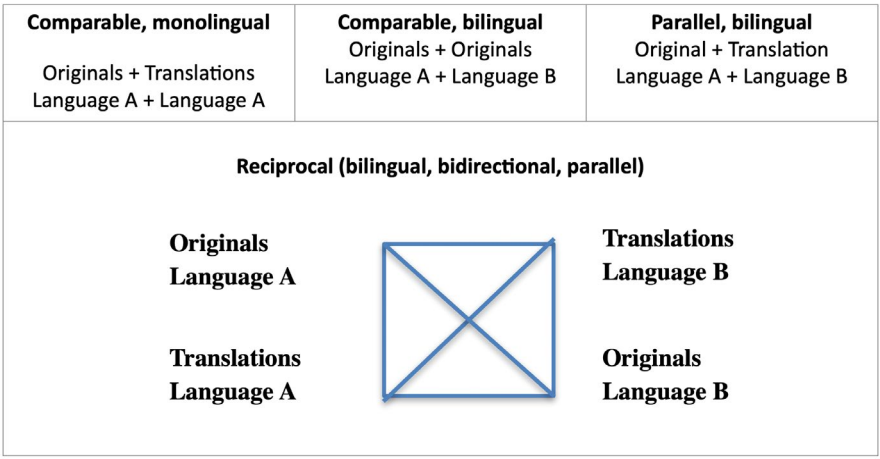

Figura 1: Tipología de corpus de traducción (Zanettin 2012)

Aun así, la Figura 1 no representa una cartografía exhaustiva del uso de los corpus. El propio Zanettin (2012: 11) la complementa con otros tipos:

> A bilingual, reciprocal corpus may be graphically represented as a square cut across by diagonal lines, in which the different subcorpora stand at the corners. Multilingual, reciprocal parallel corpora may generate complex models described as a star and diamond configurations (Johansson 2003: 139–142). In a star model, there are multiple translations of the same texts in different languages. The diamond model includes source texts in more than two languages as well as their translations in all the other languages.

Con todos estos corpus, tal y como Bernardini & Kenny confirman en una breve, pero completa revisión (2020), muchos de los estudios en curso en 1998 tuvieron una continuidad, sobre todo aquellos relacionados con la línea de investigación más destacada de los ETBC, la del estudio de las características distintivas de la lengua translacional (Othman 2020; Váradi 2007; Xiao 2010, entre otros).

Es cierto que, con la llegada del nuevo milenio, aquellos estudios iniciales basados en corpus se han visto complementados por nuevas líneas de investigación alternativas de un alcance, sin ningún género de dudas, sustancialmente menor. Es el caso, por ejemplo, de los estudios relacionados con el estilo del traductor (Saldanha 2011), los diferentes estudios exploratorios sobre multimodalidad que examinan áreas infrarrepresentadas como la interpretación (Hu & Tao 2014) o la traducción audiovisual (Baños *et al*. 2013) o las primeras aproximaciones a traducciones "produced under relatively new conditions" (Bernardini & Kenny, 2020: 113), entre las que podemos citar trabajos sobre modalidades y géneros inexistentes hasta la fecha como la localización de páginas web o la traducción en redes sociales (Jiménez-Crespo 2015). Aun así, como hemos comentado anteriormente, y bien entrados en el siglo XXI, estos senderos alternativos de investigación todavía se hallan en la periferia de los ETBC.

Así pues, a principios de la década de 1990 se produjo el nacimiento oficial de los ETBC gracias a un cambio de paradigma dentro de los estudios de traducción e interpretación que fue posible gracias a la creciente fuerza de los estudios descriptivos de la traducción y la proliferación de herramientas electrónicas cada vez más potentes. En menos de una década, los ETBC crecieron de forma constante (véase Santamaría Urbieta y Alcalde Peñalver, en este volumen) hasta que, a finales del siglo XX, existe ya suficiente material como para que Laviosa (1998) revise el campo por primera vez. Nos parece que, en 2021, el momento es particularmente propicio para otra pausa crítica (auto)reflexiva; una opinión, además, compartida por muchos.

De Sutter y Lefer (2020: 7), por ejemplo, plantean esta misma cuestión de forma especialmente vehemente al abogar por una hoja de ruta actualizada en los estudios empíricos de traducción, algo que consideran "indispensable si queremos que los hallazgos en los estudios empíricos de la traducción sean precisos, fiables y generalizables para poder empezar a construir teorías sólidas y estables". Esa nueva hoja de ruta, afirman (De Sutter y Lefer 2020: 2), permitirá avanzar en el estudio de cuestiones para las que todavía no tenemos respuesta, como puedan ser: "[the] mechanisms that shape translation, how these mechanisms interact, and to what extent this interaction functions differently than in other types of monolingual and bilingual written language production".

En su opinión, esta nueva hoja de ruta, que debe construirse sobre pilares empíricos sólidos (y debe caracterizarse por el refinamiento estadístico, la interdisciplinariedad y un marco multimetodológico), es esencial para superar los cuatro riesgos principales que pueden ser potencialmente letales para los ETBC:

(1) En la actualidad, la investigación en torno a los ETBC sigue centrándose (en exceso) en las diferencias lingüísticas entre los textos traducidos y los no traducidos con el fin de identificar los rasgos distintivos de la traducción. Sin embargo, las similitudes son también esenciales para entender el proceso y el producto de la traducción (y la interpretación), algo que, para De Sutter y Lefer (2020: 4), se suele pasar por alto en la mayoría de casos. De hecho, afirman que

> with the benefit of hindsight, it is a questionable approach to assume first and foremost differences when translated texts in a given language that are produced by highly skilled, native-language professionals are compared with texts in the same language produced by presumably equally skilled language professionals (journalists, writers, spokesmen, etc.), with the only obvious difference being the circumstances under which the texts are produced (bilingual vs monolingual language activation). […] In other words, subtle quantitative differences are likely to be found across translated and non-translated texts, alongside a massive number of commonalities.

(2) Hay todavía mucho margen de mejora respecto al marco teórico de los ETBC, el cual los autores perciben como "poco desarrollado" (De Stutter y Lefer 2020: 4-5):

> Instead of empirical research in translation studies giving rise to the falsification, verification or adaptation of the hypotheses of universal features of translation, as initially intended by Baker, universal features have been used repeatedly and uncritically to 'explain' specific patterns observed in the corpus data. In the process, translation universals have gradually lost power in that they have only been used as fixed, passe-partout post hoc explanations: whatever linguistic phenomenon is being studied, there will always be some translation universal available which can be used to rationalise the descriptive patterns uncovered in the data.

(3) Existe cierta falta de refinamiento (es decir, un uso básico de la estadística, a menudo, incorrecto) en los diseños de investigación de los ETBC, que

acaban por ser estudios de carácter reduccionista, monofactorial. El resultado es que muchos estudios se basan a menudo en un único factor explicativo (especialmente la distinción entre textos traducidos y no traducidos) para caracterizar el proceso y el producto de la traducción (y la interpretación). De Sutter y Lefer (2020: 5) citan a Gries (2018: 295) para afirmar, sin tapujos, que:

> monofactorial observational studies have virtually nothing to contribute to corpus linguistics [because] (i) no phenomenon is monofactorial and (ii) even if one had a new monofactorial hypothesis of a phenomenon, it would still require multifactorial testing to determine either (a) whether it either adds anything to what we already know about the phenomenon (by statistically controlling for what we already know) or (b) whether it replaces (parts of) what we already know about the phenomenon.

(4) Los ETBC permanecen en gran medida aislados y fijos en su objeto de estudio (los productos y procesos asociados con la traducción y la interpretación) "without taking into account theoretical and methodological developments in other, related, fields such as corpus linguistics (including learner corpus research), variational linguistics, contrastive linguistics, sociolinguistics, psycholinguistics and cognitive linguistics, to name but a few". (2020: 5).

Aun así, la hoja de ruta que proponen los autores para la investigación en el marco de los ETBC, tal y como ellos mismos afirman, no es totalmente novedosa. De hecho, trabajos previos como los de Delaere y De Sutter (2013), Hu et al. (2019) o Kruger (2019), por citar solo algunos, ya defienden postulados similares, a partir de ciertas premisas compartidas que están difundiéndose rápidamente en la academia:

> understanding of the governing principles underlying translation, and the constraints under which it operates, can be achieved, in our view, by re-adopting and updating the essential aspect of Baker's research program, i.e., looking over the disciplinary fence and carefully selecting corpus-linguistic, ethnographic, sociological, and psycholinguistic methods that are appropriate for studying central aspects of translation, as well as interpreting research outcomes in an emerging, bottom-up translation theory that builds on theories in neighboring disciplines, such as contact linguistics, second language acquisition research and psycholinguistics. (2020: 2)

Al igual que De Sutter y Lefer, creemos que un impulso renovador en la hoja de ruta original de Baker (1993) parece más que adecuado con el cambio de década, una década caracterizada por textos y contextos nunca antes vistos. En nuestra opinión, ese impulso es la única manera de seguir la propuesta de Tymoczko (1998: 653) de "pasar de la investigación basada en el texto a la investigación basada en el contexto".

No obstante, nuestra evaluación de los riesgos es algo más prudente que la de De Sutter y Lefer. Es cierto que las diferencias han sido el *quid* de la cuestión en este campo desde hace tiempo y que las similitudes son, sin duda, una parte poco investigada del panorama de la traducción. Sin embargo, resulta intrigante ver cómo, en ciertos sectores, el debate parece estar desplazándose incipientemente desde la traducción y la interpretación (como objetos de estudio de disciplinas independientes) hacia un espacio algo más amplio poblado por subgéneros "similares", los de la comunicación mediada, dentro del cual los estudios de traducción e interpretación, basados en corpus o no, podrían estar perdiendo el protagonismo que tanto les ha costado tener. Si bien es cierto que la teoría de la traducción se ve reforzada por la investigación empírica de alta calidad, también lo es que los avances teóricos no pueden provenir ni provendrán exclusivamente de los enfoques empíricos. Despreciar los enfoques no empíricos de la traducción y la interpretación (de los que tenemos tantísimos ejemplos en nuestra disciplina) sería renegar de nuestra propia historia. No se trata de negar que bien nos valdría buscar un mayor refinamiento en nuestros enfoques estadísticos (de hecho, así lo esperamos), sino de dar cuenta de que la complejidad de la traducción y la interpretación no puede abarcarse únicamente a partir de métricas. De hecho, el pasado ya nos ha advertido contra la falsa ilusión de la verdad indiscutible (Venuti, 2000). Además, ¿en qué momento los estudios de traducción e interpretación han dejado de tender puentes con otras disciplinas? Si algo nos caracteriza es que nunca hemos dejado de "mirar más allá de nuestra frontera como campo de especialidad". Nos hemos nutrido con la crítica literaria. Nos hemos informado con el análisis lingüístico. Nos hemos hecho fuertes con los estudios culturales de corte posmoderno, poscoloniales o *queer*. Nos hemos impregnado con métodos psicolingüísticos y hemos complementado nuestros estudios con enfoques sociológicos.

De hecho, cuando miramos más allá de esa frontera, nos damos cuenta de que los ETBC no son el único campo que requiere de una pausa para la reflexión. Disciplinas colindantes, como el Análisis del discurso basado en corpus (ADBC), se están enfrentando a cuestiones muy similares. Así lo atestigua la brillante monografía publicada por Taylor y Marchi (2018), cuyas conclusiones no nos sorprenden en absoluto. Tanto los ETBC como el ADBC comparten objetivos similares a la hora de "usar corpus para estudiar cómo se construyen, representan y transmiten lingüísticamente las realidades sociales" (2018: 1), sea a través de artefactos monolingües o multilingües. Ambas disciplinas comparten orígenes similares, algunos de los cuales pueden encontrarse en la lingüística de corpus y en el análisis del discurso. Ambas han crecido de manera exponencial en la última década, tanto con la aparición y consolidación de conferencias (Corpora and Discourse, en el caso del ADBC, o CILC en el caso de los ETBC) como con el aumento de publicaciones de gran relevancia. Y cuando Taylor y Marchi (2018: 2) señalan las áreas que más preocupan al ADBC (falta de estandarización, obstaculización mediante barreras institucionales, preocupación sobre datos descontextualizados, cuestiones epistemológicas, etc.), los ETBC no pueden sino verse reflejados. Así pues, no es de extrañar que, tras su éxito en los últimos años, tanto los ETBC (con contribuciones importantísimas como las de De Sutter y Lefer, entre otros) como el ADBC (con Taylor y Marchi a la cabeza, por nombrar a algunos autores) lleguen a la misma conclusión: "Ha llegado el momento de hacer una pausa y reflexionar sobre lo que hacemos". En palabras de Taylor y Marchi (2018: 2): "This is not intended as an assault of the exciting work which is emerging but a recognition of the maturity of the methodology, which is now robust enough to withstand, and indeed demand, close scrutiny".

Según Taylor y Marchi (2018), este escrutinio minucioso (que no es sino una forma autorreflexiva de conciencia) pretende combatir tres problemas inextricablemente conectados en la investigación: la parcialidad, los rincones poco transitados y los puntos ciegos. En otras palabras, el hecho de centrarse en determinados ámbitos y dejar de lado otros provoca que la investigación sea siempre incompleta (es decir, parcial), lo que deja, por omisión, ciertos rincones poco transitados, ciertos rasgos que suelen pasarse por alto (las similitudes, por ejemplo, o las ausencias) y ciertos ámbitos sin investigar

(temas minoritarios, tipos de texto y lenguas a la sombra de voces dominantes, etc.). A la vez, se crean todo tipo de agujeros negros formados por otros temas y ámbitos que pasan desapercibidos en la academia o se dejan de lado en la investigación (puntos ciegos) a los que solo puede llegarles la luz mediante enfoques innovadores, entre los que se incluyen aquellos inspirados en la triangulación de resultados. Si lo pensamos bien, este argumento se solapa con lo defendido por De Sutter y Lefer, ya que un acercamiento (¿excesivo?) a los rasgos distintivos de la traducción dejaría rincones poco transitados con elementos que pasarían desapercibidos (como las similitudes) y contenidos poco investigados (por ejemplo, muchos de los mecanismos que dan forma a la traducción), junto a verdaderos agujeros negros en la teoría de la traducción que solo podrían explorarse con investigaciones complejas que apliquen métodos y técnicas diferentes.

En línea con De Sutter y Lefer (2020) y Taylor y Marchi (2018), entre otros, este volumen nace con la intención de animar a los profesionales de los ETBC a detenerse y mirar a su alrededor; a explorar rincones poco transitados y puntos ciegos; a luchar contra la parcialidad e inyectar dosis de innovación en nuestro trabajo. En definitiva, el volumen pretende impulsar el pensamiento crítico, la (auto)conciencia y la (auto)reflexión sin renunciar a nuestro pasado. De hecho, más bien honrando un pasado que se nos antoja esplendoroso. Siempre mirando más allá de nuestras fronteras, siempre mirando hacia delante (sin dejar de mirar por el retrovisor, claro está). Con este espíritu, el volumen que presentamos está organizado a partir de cuatro temas principales: las características de la traducción como punto de partida; las áreas de estudio olvidadas y pasadas por alto; la investigación de la comunicación (original y traducida) en contextos contemporáneos y la reflexión.

Arrancamos el volumen estrechando nuestro inquebrantable vínculo con el pasado al retomar la preocupación investigadora más prolífica de los ETBC: los rasgos traslativos. En una propuesta descriptivo-explicativa ("Explicitation and Implication in Translation: Combining Comparable and Parallel Corpus Methodologies"), Miguel Ángel Jiménez-Crespo y Maribel Tercedor-Sánchez se centran, a través de corpus comparables y paralelos, en uno de los rasgos más investigados en nuestra disciplina: la explicitación. Y lo hacen no solo identificando casos de explicitación (e implicitación)

en relación con algo poco investigado en nuestra disciplina como puedan ser los términos del latín y el griego (LG) que aparecen en textos médicos, sino también profundizando en las posibles causas que subyacen a dicha explicitación (interferencias lingüísticas cruzadas, aversión al riesgo, etc.). El enfoque de Jiménez-Crespo y Tercedor-Sánchez está en línea con otros trabajos empíricos recientes sobre el tema (véase, por ejemplo, Delaere y De Sutter 2013; Kruger 2019), en lo que se nos antoja un caso único y complejo en el que convergen procesos de traducción inter e intralingüísticos donde la explicitación se entrelaza con la determinologización.

Las cuatro contribuciones siguientes giran en torno a lo que podríamos llamar *áreas de estudio olvidadas y pasadas por alto*. Nos adentramos en ámbitos poco transitados por los ETBC como el subtitulado, el periodismo de viajes, la interpretación simultánea y la audiodescripción operística. Al fin y al cabo, transitar por espacios desconocidos hasta la fecha es algo que los ETBC están comenzando a hacer con más fuerza que nunca. Al adentrarse en estos ámbitos, los cuatro artículos miran más allá de las fronteras de nuestra disciplina (algunos lo hacen en más de una ocasión) para encontrar enfoques (multi)metodológicos con lo que analizar ámbitos de nuestra disciplina tradicionalmente pasados por alto.

El artículo de Blanca Arias "Using Corpus Pattern Analysis for the Study of Audiovisual Translation. A Case Study to Illustrate Advantages and Limitations", por ejemplo, recurre a la Teoría de las normas y las explotaciones (originaria de los estudios televisivos) y a la aplicación del Análisis de patrones de corpus (APC) (originario de los estudios léxicos) para analizar colocaciones creativas (anómalas o no canónicas). El corpus escogido consiste en los subtítulos originales (en inglés) y traducidos (al español) de los dos primeros episodios de la primera temporada de las series de televisión *Castle* (2009), *Dexter* (2006) y *El mentalista* (2008). Los estudios sobre la creatividad (o explotaciones, en palabras de Arias) están ciertamente infrarrepresentados en nuestra disciplina. Por eso son de especial importancia, sobre todo cuando concluyen con la evaluación crítica de las ventajas e inconvenientes de importar enfoques metodológicos como el APC.

Las colocaciones son también el tema central del artículo de David Finbar Brett, Barbara Loranc y Antonio Pinna ("A Corpus-Driven Analysis of Adjective+Noun Collocations in Travel Journalism in English, Italian

and Polish"), en el que se explora el periodismo de viajes. Con la ayuda de tres corpus comparables de reportajes de viajes en inglés, italiano y polaco, los autores exploran diversas cuestiones: (a) similitudes y diferencias en las frecuencias de las colocaciones adjetivo/sustantivo en las distintas lenguas; (b) similitudes y efectos denotativos de algunas de las colocaciones adjetivo/sustantivo más frecuentes; (c) variabilidad sintáctica de las colocaciones; (d) conectividad de algunas de las colocaciones más productivas; y (e) comportamiento de las colocaciones en determinados temas. Entre los resultados innovadores del artículo destaca sobre todo la exploración translingüística de las colocaciones en la escritura de viajes (ciertamente, un tema de análisis poco investigado). Además, el artículo presenta otras fuentes de innovación metodológica. En primer lugar, se utilizan herramientas importadas del Análisis de redes sociales (como la plataforma Gephi) para reforzar la noción de conectividad colocacional (que se basa en el concepto de redes relacionales). En segundo lugar, destacan los protocolos automatizados de compilación y anotación (con scripts Perl creados por los propios autores), que son un paso necesario hacia esos niveles de refinamiento (en programación) que piden De Sutter y Lefer (2020). Sin duda, creemos que un requisito cada vez más ineludible en los ETBC será el dominio de técnicas y lenguajes de programación.

Las colocaciones y el refinamiento metodológico son, también, los principales ingredientes del artículo de Marta Kajzer-Wietrzny y Łukasz Grabowski: "Formulaicity in Constrained Communication: An Intermodal Approach", aunque en este caso las colocaciones giren en torno a la noción de bigrama y el refinamiento metodológico responda al dominio estadístico de sus autores. Conocer y dominar nociones complejas de estadística (al igual que el lenguaje de la programación, como se ha mostrado en el artículo anterior) se ha de convertir también en una exigencia ineludible en la investigación presente y futura de los ETBC. El artículo se basa en el modelo de regresión de Poisson con efectos fijos y aleatorios para diseccionar el uso de bigramas en diferentes subcorpus del corpus intermodal EPTIC. El resultado es un estudio comparativo de tres tipos de discursos parlamentarios en inglés: discursos traducidos a partir de originales polacos; discursos interpretados a partir de originales polacos y discursos originales producidos por nativos polacos para los que el inglés es su segunda lengua. Un ámbito

poco investigado por los ETBC, como es la interpretación simultánea, y el interés cada vez más extendido por la comunicación mediada son especialmente adecuados para el presente volumen. También cabe destacar el carácter multidimensional del estudio. Siguiendo a De Sutter y Lefer (2020), el artículo va más allá de la explicación monofactorial de los fenómenos comunicativos analizados y evalúa la causalidad en cinco dimensiones: (1) activación de la lengua (mono-bilingüe); (2) modalidad y registro (hablado, escrito-multimodal); (3) producción textual (no mediada/mediada); (4) competencia (nativa, no nativa); y (5) pericia en la tarea (alta/baja). Por último, el artículo utiliza un conjunto de herramientas infrautilizadas dentro de los ETBC (Formulib, scripts R o scripts ad-hoc en Python) para complementar el conjunto de herramientas de la investigación.

"The Hierarchization of the Operatic Signs through the Lens of Audio Description: A Corpus Study" de Irene Hermosa-Ramírez, también nos propone abordar una serie de rincones poco transitados que sin duda se benefician de las herramientas basadas en corpus. Los ETBC no han ido a menudo a la ópera, ni se han dedicado tradicionalmente a examinar la audiodescripción operística. Detectar las similitudes y diferencias en las producciones de dos grandes teatros (el Teatro Real de Madrid y el Liceu de Barcelona) no ha sido hasta ahora objeto de estudio principal en nuestro ámbito de estudio. De hecho, no se suelen aplicar análisis semióticos complementarios a los instrumentos canónicos de la investigación basada en corpus, que es precisamente lo que Hermosa-Ramírez se propone en su artículo. Para ello, la autora compila dos corpus con las audiodescripciones de tres famosas producciones operísticas (*Aida*, *La flauta mágica* y *Carmen*) representadas en el Teatro Real y en el Liceu y, con la ayuda de Sketch Engine, compara la longitud media de las frases, las frecuencias de palabras de clase abierta, las distribuciones de POS y los TTR. Posteriormente, la autora complementa sus resultados con un análisis semiótico basado en Rędzioch-Korkuz (2016) y el conjunto de etiquetas de narratología TRACCE, demostrando que la triangulación no es la única posibilidad cuando se utilizan métodos mixtos de análisis. La complementación metodológica, sin ir más lejos, es una forma poderosísima también de luchar contra la parcialidad.

La investigación de la comunicación (original y traducida) en contextos contemporáneos une a los dos artículos siguientes, que tratan temas que, en su mayoría, se dejaron al margen en los años noventa.

En "Los estudios de corpus y la localización: una propuesta de análisis para material interactivo", de Laura Mejías Climent, las sincronías en el doblaje original en inglés y traducido al español de tres videojuegos (*Batman: Arkham Knight*, *Assassin's Creed Syndicate* y *Rise of the Tomb Raider*) son el foco principal de estudio. El tema en sí mismo ya es bastante innovador, y la triangulación de datos procedentes de métodos cualitativos, resultados cuantitativos y conocimientos profesionales extraídos de entrevistas semiestructuradas añade complejidad al punto de vista desde el que se realiza el estudio. La inspiración teórica que busca el artículo en el ámbito de la localización, de la traducción audiovisual y los estudios basados en corpus genera sinergias que dan lugar a resultados más que significativos. Ejemplo de ello es la multimodalidad en este contexto, que incorpora no ya solo estímulos audiovisuales, sino también *input* táctil. Además, la unidad de análisis del estudio se aleja de la noción más tradicional de texto original y traducido y se traslada a las distintas "situaciones de juego", sin duda uno de los aspectos más novedosos de la ya de por sí innovadora propuesta de este artículo. Los textos originales y traducidos no son anteriores a la investigación, sino que son a la vez el resultado y el medio de la misma. Así, para que la investigadora pueda investigar, primero ha de jugar a los videojuegos y comparar sus resultados. Proponerse estudiar un texto *a priori* inexistente tampoco era un objetivo común durante los años noventa. Sin embargo, si lo pensamos bien, es bastante característico de los tiempos "líquidos" que nos ha tocados vivir (Bauman, 2000), en los que la materialidad de los textos virtuales/reales, existentes/inexistentes es difícil de aprehender.

Leticia Moreno-Pérez y Belén López-Arroyo, por su parte, adoptan un enfoque igualmente novedoso (un generador de escritura y su uso profesional) en "A Typical Corpus-Based Tool to the Rescue: How a Writing Generator Can Help Translators Adapt to the Demand of the Market". Como sabemos, un generador de escritura es una herramienta que ayuda a los hablantes no nativos a crear textos (de géneros de especialidad) en una lengua extranjera. El desarrollo de un generador de escritura es una tarea indudablemente refinada que enriquece el conjunto de herramientas de los

ETBC, apoyándose en gestores de archivos, constructores de etiquetas y etiquetadores que producen información sobre estructuras retóricas prototípicas (movimientos y pasos), patrones léxico-gramaticales y glosarios terminológicos y fraseológicos. Su aplicación a la traducción en el ámbito de la enología demuestra que los ETBC no solo son capaces de mirar más allá de su disciplina, sino también (y principalmente) de escuchar las demandas del mercado, reduciendo la brecha entre el mundo académico y la profesión.

Por último, pero no por ello menos importante, el volumen concluye con una reflexión crítica, objetivo último de la empresa que nos ocupa. La reflexión puede ser de dos tipos: individual, donde los investigadores analizan su producción anterior/actual para reforzar su posición teórica y metodológica, o colectiva y disciplinar, mirando más allá de lo que somos sin descuidar nuestra esencia.

"Autocrítica de publicaciones previas basadas en corpus: análisis DAFO", de Alexandra Santamaría Urbieta y Elena Alcalde Peñalver, es un ejemplo del primer tipo de reflexión. En él, las autoras revisan cuatro de sus propias publicaciones conjuntas en las que se aplican métodos basados en corpus. Para ello recurren a la metodología DAFO (Debilidades, Amenazas, Fortalezas y Oportunidades), especialmente prolífica en los estudios empresariales y de marketing, para proporcionar un andamiaje útil para la reflexión crítica y la autoconciencia. Creemos que es una buena manera de que los partidarios de los ETBC contribuyan a la renovación de esa primera hoja de ruta propuesta por Baker (1993).

Y es precisamente Mona Baker, toda una eminencia en el tema que nos ocupa, una parte importante de la contribución final del volumen, "From Text to Data: Mediality in Corpus-based Translation Studies", de Jan Buts y Henry Jones. El artículo final se produce en el marco del proyecto Genealogies of Knowledge (GoK), en el que Baker (profesora emérita en la actualidad) tiene un papel destacado. La propuesta de Buts y Jones es todo un ejemplo del segundo tipo de reflexión, ya que los académicos se atreven a reflexionar sobre los pilares fundamentales de nuestra disciplina, con el objetivo de fortalecerla. Uno de esos pilares dentro de los ETBC son los medios electrónicos "en y a través de los cuales las traducciones se almacenan, se transmiten y -por extensión- se estudian (Armstrong 2020; Pérez-González 2014)" (véase la contribución en este volumen). Estos medios no son meros contenedores

que sirven para preservar y transmitir el significado y el conocimiento de forma impoluta, sino que se erigen como verdaderos entornos transformadores que afectan profundamente a nuestra relación con (y comprensión de) los textos. En general, los ETBC parecen haber dejado de lado cuestiones de medialidad, que han quedado confinadas en algunos de nuestros rincones menos transitados. Sí es cierto, como argumentan Buts y Jones, que las limitaciones y restricciones de la tecnología se han debatido abundantemente. Sin embargo, como ellos mismos defienden, "[w]hatever the cause, the convertibility of the sign and its attachment to the binary standard are yet to be consistently questioned". Ser conscientes de que los medios electrónicos de los que dependen los ETBC y las herramientas informáticas con las que estos realizan sus análisis conducen a la aplicación de determinadas metodologías (de búsqueda de patrones) en lugar de otras (más centradas en las estructuras y las narrativas) es el primer paso para combatir la parcialidad. Diseñar nuevas herramientas de visualización y aplicarlas críticamente a conceptos políticos y científicos -como hacen Buts y Jones en el marco del proyecto GoK- nos parece una poderosa iniciativa para contribuir a la nueva hoja de ruta de los ETBC.

En los albores de la segunda década del siglo XXI, con muchísima gente confinada en todo el mundo aguantando la respiración por lo que pueda pasar y sin tener una idea clara de qué hacer a continuación, la necesidad de pararse a pensar de forma crítica es más acuciante que nunca.

References

BAKER, Mona. (1993) "Corpus Linguistics and Translation Studies — Implications and Applications." En: Baker, Mona; Gill Francis & Elena Tognini-Bonelli (eds.) 1993. *Text and Technology*. Amsterdam: John Benjamins, pp. 233-50.

BAKER, Mona. (1998) "Réexplorer la langue de la traduction: une approche par corpus." In: Sara Laviosa (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 480-485.

BAKER, Mona. (2004) "A Corpus-Based View of Similarity and Difference in Translation." *International Journal of Corpus Linguistics* 9:2, pp.167-93.

BAÑOS, Rocío; Silvia BRUTI & Serenella ZANOTTI (2013) "Corpus Linguistics and Audiovisual Translation: In Search of an Integrated Approach." *Perspectives* 21:4, pp. 483-90.

BAUMAN, Zygmunt. (2000) *Liquid Modernity*. Cambridge: Polity Press.

BERNARDINI, Silvia & Dorothy KENNY. (2020) "Corpora." En: Baker, Mona & Gabriela Saldanha (eds.) 2020. *Routledge Encyclopedia of Translation Studies*. Tercera edición. Londres; Nueva York: Routledge, pp. 112-15.

BOWKER, Lynne. (1998) "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study." En: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 631-651.

DE SUTTER, Gert & Marie-Aude LEFER. (2020) "On the Need for a New Research Agenda for Corpus-Based Translation Studies: A Multi-Methodological, Multifactorial and Interdisciplinary Approach." *Perspectives* 28:1, pp. 1-23.

DELAERE, Isabelle & Gert DE SUTTER. (2013) "Applying a Multidimensional, Register-Sensitive Approach to Visualize Normalization in Translated and Non-Translated Dutch." *Belgian Journal of Linguistics* 27, pp. 43-60.

EBELING, Jarle. (1998) "Contrastive Linguistics, Translation, and Parallel Corpora." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 602-615.

HALVERSON, Sandra. (1998) "Translation Studies and Representative Corpora: Establishing Links between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study." En: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 494-514.

HU, Kaibao. (2016) *Introducing Corpus-Based Translation Studies. New Frontiers in Translation Studies*. Berlin, Heidelberg: Springer Berlin Heidelberg.

HU, Kaibao & Qing TAO. (2014) "The Chinese-English Conference Interpreting Corpus: Uses and Limitations." *Meta* 58:3, pp. 626-42.

HU, Xianyao; Richard XIAO & Andrew HARDIE. (2019) "How Do English Translations Differ from Non-Translated English Writings? A Multi-Feature Statistical Model for Linguistic Variation Analysis." *Corpus Linguistics and Linguistic Theory* 15:2, pp. 347-82.

JIMÉNEZ-CRESPO, Miguel A. (2015) "Translation Quality, Use and Dissemination in an Internet Era: Using Single-Translation and Multi-Translation Parallel Corpora to Research Translation Quality on the Web." *The Journal of Specialised Translation* 23, pp. 39-63.

KRUGER, Haidee. (2019) "*That* Again: A Multivariate Analysis of the Factors Conditioning Syntactic Explicitness in Translated English." *Across Languages and Cultures* 20:1, pp. 1-33.

LAVIOSA, Sara. (1998a) "The English Comparable Corpus: A Resource and a Methodology." En: Lynne Bowker, Lynnen; Michael Cronin; Dorothy Kenny & Jennifer Pearson (eds.) *Unity in Diversity? Current Trends in Translation Studies*. Manchester: St. Jerome, pp. 101-112.

LAVIOSA Sara. (1998b) "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose." En: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 557-570.

LAVIOSA, Sara (ed.) (1998) *Special Issue: L'approche Basée Sur Le corpus/The Corpus Based Approach. Meta. Journal Des Traducteurs/Meta. Translators' Journal*, 43:4.

LAVIOSA, Sara. (2002) *Corpus-Based Translation Studies: Theory, Findings, Applications*. Amsterdam: Rodopi.

MAIA, Belinda. (1998) "Word Order and the First Person Singular in Portuguese and English." En: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/ The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 589-601.

MALMKJÆR, Kirsten. (1998) "Love thy Neighbour: Will Parallel Corpora Endear Linguistics to Translators?" En: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 534-541.

MUNDAY, Jeremy. (1998) "A Computer-Assisted Approach to the Analysis of Translation Shifts." In: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 542-556.

OLOHAN, Maeve. (2004) *Introducing Corpora in Translation Studies*. Londres; Nueva York: Routledge.

OTHMAN, Waleed. (2020) "An SFL-based model for investigating explicitation-related phenomena in translation." En: Calzada Pérez, María & Jeremy Munday (eds.) *Meta: Journal des traducteurs* 65:1, pp. 193-210.

ØVERÅS, Linn. (1998) "In Search of the Third Code: An Investigation of Norms in Literary Translation." En: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 571-588.

RĘDZIOCH-KORKUZ, Anna. (2016) *Opera Surtitling as a Special Case of Audiovisual Translation*. Berna: Peter Lang.

SALDANHA, Gabriela. (2011) "Translator Style: Methodological Considerations." *The Translator* 17:1, pp. 25-50.

SHLESINGER, Miriam. (1998) "Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies." En: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 486-493.

TAYLOR, Charlotte & MARCHI, Anna (eds.) (2018) *Corpus Approaches to Discourse: A Critical Review*. Milton Park, Abingdon, Oxon; New York: Routledge.

TYMOCZKO, Maria. (1998) "Computerized Corpora and the Future of Translation Studies." *Meta* 43:4, pp. 652-60.

VÁRADI, Tamas. (2007) "NP Modification Structures in Parallel Corpora." En: Rogers, Margaret & Gunilla Anderman (eds.) 2007. *Incorporating Corpora. The Linguist and the Translator.* Clevedon: Multilingual Matters.

VENUTI, Lawrence. (2000) "¿Será Útil La Teoría de La Traducción Para Los Traductores?" *Vasos Comunicantes* 16, pp. 26-34.

XIAO, Richard. (2010) "How Different Is Translated Chinese from Native Chinese?: A Corpus-Based Study of Translation Universals." *International Journal of Corpus Linguistics* 15:1, pp. 5-35.

ZANETTIN, Federico. (1998) "Bilingual Comparable Corpora and the Training of Translators". En: Laviosa, Sara (ed.) 1998. *L'Approche Basée sur le Corpus/ The Corpus-based Approach. Special Issue of Meta* 43:4, pp. 616-630.

ZANETTIN, Federico. (2012) *Translation-Driven Corpora Corpus Resources for Descriptive and Applied Translation Studies*. Translation Practices Explained. Manchester; Kinderhook, Nueva York: St. Jerome Pub.

BIONOTE / NOTA BIOGRÁFICA

MARÍA CALZADA PÉREZ is Full Professor of Translation Studies at the Universitat Jaume I (Spain). Her research mainly focuses on corpus-based translation studies, institutional translation (especially at the European Parliament), ideology, and translator-teaching. She is Coordinator of the ECPC (European Comparable and Parallel Corpora of Parliamentary Speeches) research group. She has produced research, such as (i) *Transitivity in Translating: The Interdependence of Texture and Context* (Peter Lang, 2007); (ii) "Five Turns of the Screw. A CADS analysis of the European Parliament" (*Journal of Language and Politics* 16:3, 2017) (iii). "Corpus-based methods for Comparative Translation and Interpreting Studies" (*Translation and*

Interpreting Studies 12:2, 2017); and (iv) "What is kept and what is lost without translation? A corpus-assisted discourse study of the European Parliament's original and translated English" (*Perspectives*. 26:2, 2017). She is also editor of volumes like: *Apropos of Ideology* (St. Jerome, 2003).

María Calzada Pérez es catedrática de Traducción en la Universitat Jaume I (España). Sus líneas de investigación se centran en los estudios de traducción basados en corpus, la traducción institucional (especialmente sobre el Parlamento europeo), la ideología y la didáctica de la traducción. Coordina el grupo de investigación ECPC (European Comparable and Parallel Corpora of Parliamentary Speeches). Entre otros, ha publicado *Transitivity in Translating: The Interdependence of Texture and Context* (Peter Lang, 2007); "Five Turns of the Screw. A CADS analysis of the European Parliament" (*Journal of Language and Politics* 16:3, 2017); "Corpus-based methods for Comparative Translation and Interpreting Studies" (*Translation and Interpreting Studies* 12:2, 2017); y "What is kept and what is lost without translation? A corpus-assisted discourse study of the European Parliament's original and translated English" (*Perspectives*. 26:2, 2017). También ha editado *Apropos of Ideology* (St. Jerome, 2003).

Sara Laviosa is Associate Professor in English Language and Translation Studies at the Department of Lettere Lingue Arti Italianistica e Culture Comparate (LeLiA), Università degli Studi Aldo Moro (Italy). Her research interests are in corpus-based translation studies and pedagogic translation. She is the author of *Corpus-based Translation Studies* (Rodopi/Brill, 2002), *Translation and Language Education* (Routledge, 2014) and *Linking Wor(l)ds* (Liguori, 2020). She is the co-author (with Adriana Pagano, Hannu Kemppanen & Meng Ji) of *Textual and Contextual Analysis in Empirical Translation Studies* (Springer, 2017). She is the guest editor of *L'Approche Basée sur le Corpus/The Corpus-based Approach* (Special Issue of *Meta* 43:4, 1998), *Translation in the Language Classroom* (Special Issue of *The Interpreter and Translator Trainer* 8:1, 2014). She is the co-editor (with Maria González-Davies) of *The Routledge Handbook of Translation and Education* (Routledge, 2020) and (with Meng Ji) of *The Oxford Handbook of Translation and Social Practices* (OUP, 2020). She is the Founder and Editor of the journal *Translation and Translanguaging in Multilingual Contexts* (John Benjamins).

SARA LAVIOSA es profesora titular de inglés y traducción en el Dipartamento di Lettere Lingue Arti Italianistica e Culture Comparate (LeLiA) de la Università degli Studi Aldo Moro, (Italia). Sus líneas de investigación son los estudios de traducción basados en corpus y la didáctica de la traducción. Es autora de *Corpus-based Translation Studies* (Rodopi/Brill, 2002), *Translation and Language Education* (Routledge, 2014) y *Linking Wor(l)ds* (Liguori, 2020). Es coautora (junto a Adriana Pagano, Hannu Kemppanen & Meng Ji) de *Textual and Contextual Analysis in Empirical Translation Studies* (Springer, 2017). Ha sido editora invitada de *L'Approche Basée sur le Corpus/The Corpus-based Approach* (número especial de *Meta* 43:4, 1998), *Translation in the Language Classroom* (número especial de *The Interpreter and Translator Trainer* 8:1, 2014). Es coeditora (junto a María González-Davies) de *The Routledge Handbook of Translation and Education* (Routledge, 2020) y (junto a Meng Ji) de *The Oxford Handbook of Translation and Social Practices* (OUP, 2020). Es miembro fundador y editora de la revista *Translation and Translanguaging in Multilingual Contexts* (John Benjamins).

Recibido / Received: 28/04/2020
Aceptado / Accepted: 25/08/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.02

Para citar este artículo / To cite this article:
Jiménez-Crespo, Miguel Ángel & Maribel Tercedor Sánchez. (2021) "Explicitation and implicitation in translation: Combining comparable and parallel corpus methodologies." In: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection*. *MonTI* 13, pp. 62-92.

EXPLICITATION AND IMPLICITATION IN TRANSLATION: COMBINING COMPARABLE AND PARALLEL CORPUS METHODOLOGIES

MIGUEL ÁNGEL JIMÉNEZ-CRESPO
jimenez.miguel@rutgers.edu
Rutgers University

MARIBEL TERCEDOR SÁNCHEZ
itercedo@ugr.es
Universidad de Granada

Abstract

This paper studies explicitation and implicitation in translated medical texts using a combination of comparable and parallel corpus methodologies. Previous corpus research in this domain has shown common lexical and syntactic shifts between translated and non-translated texts (Askehave & Zethsen 2000; Jensen & Zethsen 2012), including differences in explicitation rates surrounding Latin-Greek (LG) terms (Jiménez-Crespo & Tercedor 2017). A parallel corpus section was compiled in order to identify whether the observed higher explicitation ratios in English to Spanish translations when compared to similar non-translated texts in this last study are due to (1) cross-linguistic interference or replication of source text structures, or (2) to the translational tendency to explicitate. The results point to a possible combination of both, with 21% of cases of explicitation and no implicitation. Higher explicitation ratios mainly support the interference or cross-linguistic influence hypothesis (Kruger 2018). This study also offers support for the risk aversion hypothesis (Pym 2005, 2015; Kruger 2018; De Sutter & Kruger 2018), as translations only show a tendency to include clearer and more explicit formulations.

Keywords: Explicitation; Implicitation; Medical Translation; Medical Terminology; Parallel and Comparable Corpus Studies.

Resumen

En este artículo se estudian las estrategias de explicitación e implicitación en textos médicos traducidos del inglés al español mediante el análisis de un corpus comparable y otro paralelo. Estudios de corpus previos han mostrado que existen diferencias léxicas y sintácticas en los textos traducidos al compararlos con los no traducidos (Askehave & Zethsen 2000; Jensen & Zethsen 2012), entre ellas diferencias en los porcentajes de explicitación de términos médicos de origen grecolatino (Jiménez-Crespo & Tercedor 2017). Se compiló un corpus paralelo con el fin de analizar si el mayor porcentaje de explicitación observado en las traducciones ingles-español con respecto a textos no traducidos en este último estudio se debe (1) a interferencia interlingüística o calcos de estructuras textuales de la lengua origen, o (2) a la tendencia a explicitar, propia de la traducción. Los resultados apuntan a una posible combinación de ambas. La tendencia a explicitar apareció en 21% de posibles casos, sin ningún caso de implicitación. Estos resultados confirman mayoritariamente la hipótesis de interferencia o influencia interlingüística (Kruger 2018). Además, respaldan la hipótesis de la aversión al riesgo (Pym 2005, 2015; Kruger 2018; De Sutter & Kruger 2018), pues las traducciones solo muestran una tendencia a la inclusión de formulaciones más claras y explícitas.

Palabras clave: Explicitación; Implicitación; Traducción médica; Terminología médica; Estudios de corpus paralelos y comparables.

1. Introduction

Explicitation is the general feature, tendency, or universal of translation that has sparked the highest interest in the scholarly community, including the largest body of theoretical research (e.g. Pym 2005; Becher 2011; Hansen-Schirra *et al.* 2007; Murtisari 2013, 2016; Kruger 2014; De Metsenaere & Vandepitte 2017). The present study was motivated by a previous comparable corpus study on explicitation (Jiménez-Crespo & Tercedor 2017). This study showed that medical translations from English into Spanish contained significantly higher lexical explicitation ratios for Latin and Greek terms (LG terms) (e.g. *hypoglycemia* or high blood sugar, *dyspnea* or difficulty breathing) than similar non-translated texts.

For the purposes of the present study, a parallel section was added to the existing comparable corpus. A comparable corpus can be defined as "a structured electronic collection of texts originally written in a particular language, alongside texts translated into that same language" (Baker 1995: 234), while a parallel corpus is a collection of texts alongside their translation(s) (Laviosa 2002: 37). The addition of the parallel section allows for the comparison of source and target texts, and this was used to test whether these higher ratios of explicitation were due to: (1) the mere replication of source text structures, indicative of language, text type and genre-specific differences in lexical use between the language pair in question; (2) the translation process, as a specific type of communicative event (Baker 1995), resulting in explicitation shifts; or (3) a combination of both. The corpus and the compilation process will be described in detail in the methodology section.

From a theoretical and methodological perspective, the motivations of the study are twofold. On the one hand, the existing debate on whether comparable corpora by themselves are an ideal methodology to study explicitation (Hansen-Schirra et al 2007; Murtisari 2016; De Maurentis & Vandepitte 2017). Researchers have argued that this methodology limits the formulation and testing of hypothesis on the causes of explicitation, as source texts are not included (e. g. Kruger 2014: 167, 2018: 23, 25). On the other, it is also motivated by the need to combine the study of explicitation/ explicitness with implicitation / implicitness (Becher 2010; Kathelijne 2012; Kruger 2014; Van Beveren, De Sutter & Colleman 2018). Normally, as indicated a decade ago (Klaudy & Karoly (2005: 14), the number of results revealing explicitation have greatly outnumbered those of implicitation, and this was the foundation of the asymmetry hypothesis. This hypothesis states that, in any given language pair, explicitation and implicitation cannot be considered as symmetric tendencies. In this regard, it is common that "obligatory, optional and pragmatic explicitations tend to be more frequent than the corresponding implicitations regardless of the SL/TL constellation at hand." (Becher 2010: 17).

Following the programmatic direction by several scholars such as Becher (2011) and Klaudy (2004), different studies have explored explicitation and implicitation jointly. Most of them have shown that translators normally

prefer explicitation rather than implicitation strategies in different types of domains, such as legal texts (Hjort-Pedersen & Faber 2010; Faber & Hjort-Pedersen 2013), scientific texts (Kruger 2015), children's literature (Erfieni 2017); back translation (Makkos & Robin 2014), contrasting expert and students' literary translations (Maraeva 2017), and different language specific features in language combinations such as Russian-English (Tao & Jiang 2017). Nevertheless, in a very small number of studies contradictory evidence has also been found (Van Beveren, De Sutter & Colleman 2018).

The textual genre and translation type selected for the study are not casual. The corpus consists of original and translated medical websites with information about health issues for the general public. This type of expert to layperson genre represents an ideal context to study explicitation and implicitation phenomena, because lexical and syntactic usage has to be adjusted to the expectations of end users and to genre conventions. This is so in both intralingual translation through so-called intralinguistic explicitation (Hill-Madsen 2015) taking place in expert to non-expert communication, as seen in the composition of source texts themselves, as well as in explicitation/implicitation shifts resulting from the translation process.

The specific case of English to Spanish directionality also offers an ideal context to study lexical explicitation-implicitation. Both languages display different lexical variation strategies and patterns of use of medical concepts (e.g. Alarcón, López & Tercedor 2016). Latin and Greek (LG) terms are more widely used in Spanish than in English. Therefore, a medical LG term might be adequate in a Spanish text addressed at the non-expert, but this same term would only be found in a more specialized expert-to-expert text in English (Campos Andrés 2013). Support for this fact is also found in the comparable corpus study by Jiménez-Crespo & Tercedor (2017), where the rate of use of LG terms was far less frequent in translations according to normalized frequency lists. These differences were found to be statistically significant. It was also found that even when LG terms were less frequently used in translations, they also showed higher reformulation/explicitation ratios. This means that translators might simply replicate patterns of lexical use in source texts, including cases of intralinguistic explicitation, that is, LG terms with the lay terminological variant (e.g. *myalgia*, also known as pain in the muscles; *myalgia* or muscle pain). In addition, Campos Andrés (2013:

53) indicates the translation of medical patient guides from English into Spanish does not always require to directly transfer intralinguistic explicitation formulations (e.g. *dyspnea* or difficulty breathing) or doublets (e.g. scaring or cicatrization) in source English texts: these LG terms can be perfectly understood by Spanish lay audiences. Again, the previous study offers support for this recommendation, as LG terms showed markedly higher frequency in non-translated texts than in translated ones (Jiménez-Crespo & Tercedor 2017). Oftentimes, the Spanish target reader has a higher chance of being familiar and acquainted with terms of LG origin than the English language reader, since they are common in general language. An example can be *otorrinolaringólogo*: this LG term is commonly used in Spanish while English commonly uses ENT or Ear-Nose-Throat doctor.

The present paper is structured as follows. It first briefly reviews the notions of explicitation and implicitation in Translation Studies, as well as the different hypotheses proposed and recent empirical studies about their causation. It then reviews explicitation and implicitation in translated medical texts, with a focus on the English-Spanish combination. It presents the actual phenomena under study, determinologization and the strategies used to adjust the degree of terminological specialization that involve explicitation/implicitation shifts, as well as shifts in the degree of technicality or specialization of texts. This is followed by an overview of the working hypotheses, the methodology used and the types of variables and shifts identified. The results obtained are then presented with a discussion.

2. Explicitation and implicitation in Translation Studies

Explicitation can be considered as the most popular and widely studied "universal of translation" (Baker 1995), "general feature of translation" (Chesterman 2004), or "general tendency in translation" (Olohan 2004) in the subdiscipline of Corpus-Based Translation Studies (CBTS) (Laviosa 2002). The origins of this research trend can be traced to the seminal studies of Blum-Kulka (1986), Klaudy (1989), Øverås (1998) or Olohan & Baker (2000). Explicitation was initially defined by Vinay & Darbelnet ([1958] 1995: 342), as a "stylistic translation technique that consists of making explicit in the target language what remains implicit in the source language

because it is apparent from either the context or the situation". Another seminal definition in TS is that by Baker, who defines explicitation as "an overall tendency to spell things out rather than leave them implicit" (Baker 1996: 180). Blum-Kulka, on her part, defined her explicitation hypothesis as an "observed cohesive explicitness from SL to TS texts regardless of the increase traceable difference between the two linguistic and textual systems involved" (2001: 30). More recent definitions indicate that explicitation "is observed where a given target text is more explicit than the corresponding source text" (Becher 2011: 19). Explicitation emerges in all instances of language mediation (Blum-Kulka 1986:19), including intralinguistic and interlinguistic communicative processes (Hill-Madsen 2015). Theoretical studies on explicitation often refer back to the Klaudy's (1988, 2009) seminal distinction between obligatory, optional, pragmatic explicitation and translation-inherent explicitation, the last type being highly controversial (e.g. Becher 2011). Obligatory explicitation is "dictated by the differences in the syntactic and semantic structures of languages" (Klaudy 2009: 106), while optional explicitation is motivated by differences between the source and target language systems. That is the case of the studies on optional explicitation of the complementizer *that* in English or *om* in Dutch (Olohan & Baker 2002; Kruger 2018; Beveren De Sutter & Colleman 2018). These particles are compulsory in some languages but optional in others. Other studies have explored, for example, the omission of the subject personal pronoun in Spanish when translating from English (Jiménez-Crespo 2012), showing higher rates of explicitation of the subject pronoun, even though Spanish is a pro-drop language. Pragmatic explicitation is motivated by differences in the cultural or world knowledge of the source and target text audiences. The last type, translation-inherent explicitation, is the most controversial. It is caused by "the nature of the translation process itself" (Klaudy 2008: 107). This type has been the object of several studies that have attempted to dismiss it as harder to identify or justify (e.g. Becher 2011). De Metsenaere & Vandepitte (2017), for example, argue that the notion of translation inherent explicitation should be eliminated and all instances under what they call language systematic categories, or simply in the pragmatic category, as explicitation is seen as a strategy to avoid risk in communicative settings. Becher (2011), on his part, indicates that this category is not clearly defined

by Klaudy, and this is exemplified by the fact that this is the category in which Klaudy herself fails to provide any guiding examples in her work, thus proving that this is a fuzzy and unclear category.

On its part, the opposite tendency, implicitation, has relatively been neglected in TS until the last decade (Kruger 2014: 164). Vinay & Darbelnet defined it as "making what is explicit in the source language implicit in the target language, relying on the context or the situation for conveying the meaning" (Vinay & Darbelnet 1995: 344). Similarly, Becher indicates that implicitation "is observed where a given target text is less explicit (more implicit) than the corresponding source text" (2010: 51). The study of implicitation has regularly been approached as a factor or condition influencing explicitation.

This brief summary of definitions is by no means comprehensive, and many other works offer intense debate on definitions and approaches to explicitation and implicitation (Pym 2005; Becher 2010; Murtisari 2016; De Metsenaere & Vandepitte 2017). This heterogeneity of approaches and epistemological fuzziness has complicated a unified approach that allows for comparison of studies across the board. Often, studies are based on different theoretical frameworks and points of departure, such as the latest studies that attempt to define explicitation based on the notions of explicitness from Relevance Theory (e.g. Murtisari 2016; De Metsenaere & Vandepitte 2017). According to De Metsenaere & Vandepitte (2017: 383), approaches have been:

> [of] very heterogeneous nature, which has made it difficult, if not impossible to compare their findings and to come to conclusive insights into the meanings of explicitation and implicitation in and for translation and translation studies.

The issues surrounding explicitation and implicitation are also complicated by the fact that tendencies are often "tentatively proposed on the basis of empirical results pertaining only to a subset" (Chesterman 2004: 40), that is, limited to a phenomenon observed in a specific corpus or language combination. Similarly, scholars have proposed to incorporate in the study of explicitation multiple variables, such as genre, register, and level of specialization. For example, in a study by Kruger (2016) the degree of technicality or level of specialization of a text was found to correlate to higher explicitation ratios. He found that in an expert-to-expert subcorpus of technical texts there was

a higher rate of explicitation than in an expert-to-semi-expert subcorpus exhibiting a medium degree of technicality.

Nevertheless, without any doubt, the "tendency" to explicitate has been mostly confirmed in most textual subsets or translation contexts in CBTS research, including different language combinations, textual genres, different levels of language, and in a number of translation modalities, such as subtitling (Perego 2003) or localization (Jiménez-Crespo 2011, 2016). It has also been confirmed on different translational phenomena, such as translated texts vs. edited-revised translations (Kruger 2012; Bisiada 2017) or even consecutive (Tang 2018) or simultaneous interpreting (Ewa 2006). However, some studies have also found inconclusive evidence or are cautious about attributing the identified effects to explicitation, such as Puurtinen (2004), Englund Dimitrova (2005), Cheong (2006) or Hansen-Schirra *et al.* (2007).

Recent studies have not only tried to ascertain whether the explicitation hypothesis is empirically confirmed, but have also tackled issues of causation and procedures for testing it, primarily since the advent of advanced multifactorial statistical analysis in Translation Studies (e.g. Oakes and Ji 2012; Kruger & De Sutter 2018; De Sutter & Lefer 2020). In this sense, CBTS has progressed towards a more theoretically robust and methodologically sound treatment of these phenomena (e.g. Becher 2010; Kruger 2013, 2018; Halverson 2016; De Sutter *et al.* 2018). Studies have tested the different proposed hypotheses, more specifically the source-language transfer hypothesis (Becher 2010), also known as the interference or cross-linguistic influence (CLI) effects (Kruger & Van Rooy 2016; Kruger 2018), the risk-aversion hypothesis (Pym 2005, 2015) and the processing-strain hypothesis (Olohan & Baker 2000). According to the cross-linguistic influence effects hypothesis, features or tendencies in translated texts can possibly be traced back to the replication of source language structures when contrastive differences between source and target systems exist (Kruger & Van Rooy 2016; Kruger 2018). Therefore, under certain conditions, source language preferences might be transferred to translations. In the case of explicitation, a source text in a language "with a preference for greater explicitness" might result "in a target text that demonstrates increased explicitness in comparison with non-translated texts in the same language" (Kruger 2018:7). This could be the cause behind the higher explicitation ratios observed in the previous

study (Jiménez-Crespo & Tercedor 2017). If English has a preference for greater lexical explicitness when using LG terms than Spanish, the higher ratios observed in the comparable corpus study would be simply caused by this effect. The risk aversion hypothesis, according to Pym (2005) assumes that, in any translation process, taking a 'risk' might be associated to a partial or complete failure in communication, the main purpose of translation. Therefore, translators, in a receiver-oriented view of communication, make information more easily available to the target audience. This includes the possibility of explicitation, in order to reduce the potential risk of lack of understanding. According to the processing strain hypothesis / cognitive complexity hypothesis (Beveren, De Sutter & Colleman 2018; De Kruger & De Sutter 2018), translators might select the most frequent option in translation to reduce effort. In this case, if the implicit form might be the most frequent option, translators might select it to reduce cognitive load, regardless of other factors.

Recent studies into the causation of explicitation have dismissed the cross-linguistic influence hypothesis (Kruger & De Sutter 2018), and they have offered some support for the processing strain hypothesis / cognitive complexity hypothesis (Beveren, De Sutter & Colleman 2018). Nevertheless, a majority of studies conclude that the risk-aversion hypothesis is the most probable cause of explicitation (Becher 2010; Kruger 2018; Kruger & Van Rooy 2016; Kruger & De Sutter 2018; Kruger 2018), even when risk aversion might not always imply explicitation (Kruger & De Sutter 2018).

3. Explicitation/ Implicitation in the translation of medical texts

According to Montalt, Zethsen & Karwacka (2018: 29), "[a]ppropriate use of medical terminology is one of the core conditions for successful communication in monolingual and multilingual healthcare communities". Nevertheless, one of the main issues that translators face is "adapting their terminological choices to genre-specific and register-specific conventions" (Montalt, Zethsen & Karwacka 2018: 30). In this problem area, medical LG terms play a key role because, even when they are commonplace in all languages, their coexistence with vernacular terminological variants varies from language to language (Askehave & Zethsen 2000; Zethsen 2004; Gutiérrez

Rodilla 2014; (Jiménez-Crespo & Tercedor 2017). Latin was not incorporated to the same extent in all European Languages (Zethsen 2004: 132), and while Spanish, Italian and French medical terminology is eminently Latin and Greek in origin, Northern European languages possess a double-layer medical terminology. In these languages, many scientific LG words find lay or lower register counterparts (e.g. clotting / coagulation, scar / cicatrization), while in Romance languages only the Latin based ones exist ('coagulación' and 'cicatriz' respectively). Therefore, what "in Latin-based languages might sound too low a register is perfectly acceptable as scientific terminology in English" (Montalt-Resurrecció & González Davies 2007: 242). This can result in different issues, because due to historical differences "seemingly identical words may indeed be false friends in an interlinguistic context" and they might also present issues related to the "connotative differences, e.g. at the level of formality" (Zethsen 2004: 131-132). For example, one of the most common ways in which synonymy occurs in medical and scientific domains in English is through the existence of the technical term with its more or less specialized equivalents or doublets, such as with "cephalalgia" and "headache". It is often understood that these cases of synonymy are "a source of translation problems because languages are not symmetrical in their use: for example, what in Spanish is considered to be low register may be perfectly acceptable in English in the same text genre" (Montalt 2011: 80).

Explicitation is part of an intralinguistic communicative process known as reformulation or determinologization. This is one of the most frequent strategies at the lexical level to lower the register and adapt textual genres to non-expert readership. This process appears in medical texts both in intralingual translation such as the case of research articles summarized for lay audiences in the *Annals of Internal Medicine* (Muñoz-Miquel 2012: 200-202), or also in translated texts for general audiences (Tercedor & López Rodríguez 2012). This process involves using general language to communicate the meaning of a specialized term (Meyer & Mackintosh 2000), helping to close the gap between specialized knowledge and lay audiences. Montalt and Shuttleworth (2012: 16) refer to determinologization as:

> a process of recontextualisation and reformulation of specialized terms aiming at making the concepts they designate relevant to and understandable by a lay audience. This process is motivated by specific cognitive, social

> and communicative needs, and takes place as part of a broader process of recontextualisation and reformulation of discourse.

This process involves a large number of potential strategies that are covered under this hypernym such as explanation, definition, reformulation, exemplification, illustration, analogy, comparison and substitution by a more popular term (Campos Andrés 2013; Montalt & González Davies 2007: 252-253). According to Montalt and González Davies (2007), this process can involve a number of strategies to deal with LG terms:

(1) Retain LG term adding an explanation: Poliuria, increase in the volume of urine.
(2) Retain LG term in parenthesis after the explanations: increase in the volume of urine (poliuria).
(3) Retain LG after a popular term: bad breath or halitosis.
(4) LG term is omitted and replaced by explanations or popular terms: patients can experience an increase in the volume of urine.

All these mechanisms can help increase the readability and efficiency of translated medical texts for lay readers, but also relate to explicitation. If reformulation or determinologization represents a natural mechanism in intralingual translation in medical genres, the translation process can potentially increase or decrease the frequency and nature of the explicitation strategies naturally present in these texts. The study of explicitation in translated medical texts therefore has to take into consideration the compounding effect of reformulation as an intralinguistic and intergeneric translation strategy (Ezpeleta 2012), as well as the potential general explicitation tendency in translated texts.

4. Empirical study

The present study focuses on lexical explicitation. Some studies have focused on semantic explicitation such as Puurtinen (2004), who zooms in connectives to identify the type of semantic relation that is conveyed. Øverås (1998:5) includes in her study both grammatical and lexical explicitation, but the former has primarily been the objective of a handful of studies on

translations of literary texts (e.g. Olohan & Baker 2000). The research questions for this study are the following:

1. Research question 1 (RQ1): Do findings from parallel corpora support the explicitation tendency already observed in translated texts through a comparable corpus methodology? This would be expected as previous literature shows that explicitation shifts tend to outnumber implicitation ones, even when implicitation might be appropriate for the English to Spanish combination in medical translation.

2. Research question 2 (RQ2): If higher explicitation ratios are found, do they offer support for the proposed cross-linguistic effect hypothesis and-or the risk aversion hypothesis?[1] This will depend on whether higher explicitation rates are due solely to the replication of source text structures in translation in combination with text type and genre specific differences between the languages under study (EN, ES), or whether translators also make use of explicitation strategies during the translation process.

The hypotheses for the study are:

(1) Hypothesis 1 (H1). Translated texts will show higher ratios of explicitation than implicitation.
(2) Hypothesis 2 (H2). Explicitation in translated texts will mostly be due to a combination of ST structure replication, as well as to translation-inherent explicitation.

4.1. Methodology: The parallel and comparable corpus

This study applies a parallel and comparable corpus methodology, the preferred methodological option to research explicitation and implicitation (Hansen-Schirra *et al*. 2007; Murtisari 2016). The Translational Web Corpus of Medical Spanish (TWCoMS) (Jiménez-Crespo 2014; Jiménez-Crespo & Tercedor 2017) was used, a comparable corpus project conceived as a tool to study variation in translated medical terminology in the United States applied to the VariMed project terminological database (http://varimed.

1. The processing-strain hypothesis (Olohan & Baker 2001) has been discarded for this study. Previous studies have tested it using complexity -related factors related to syntax (Beveren, De Sutter & Colleman 2018) and this study focuses primarily on a noun-phrases related to explicitation of LG terms.

ugr.es) (Tercedor, López Rodríguez & Prieto Velasco 2014). It includes a section of non-translated medical websites addressed at lay readers in Spanish alongside a representative sample of Spanish translations of US medical websites. For the purposes of the present study, a small parallel subcorpus was also added to the TWCoMS, consisting of the largest English websites and their Spanish versions (CDC.gov, MedlinePlus, Womenshealth.gov, Cancer.gov). The comparable portion of the corpus was downloaded and compiled in 2014, and the parallel section in late 2018. The comparable section of the TWCoMS corpus contains medical websites addressed at general audiences in the US and a comparable section of websites localized for Latin America, Mexico and Spain markets.. The translational subcorpus contains 32,330,052 tokens. The comparable non-translational section contains 8,701,867 tokens.

The parallel section was compiled using the Httrack software. Research has shown that localized websites are normally smaller than the original counterparts (Jiménez-Crespo 2012), and therefore, not all source webpages in English in the websites compiled had a Spanish version. For example, once the website for the Center for Disease Control was downloaded there were over 244,000 pages, but only 2,540 were in Spanish. In addition, only 1,834 were direct translations with a corresponding source webpage. The overall parallel corpus thus includes 16,534 source pages with their corresponding localized web pages from the websites Center for Disease Control (CDC.gov), Cancer.gov, WomensHealth.gov and Medlineplus.gov. The corpus was aligned at the page level, and not at the segment level. Since Httrack allows us to download the websites with their internal structures, and each Spanish page has an "English" translation link and vice versa, the searches were carried out in the html version of the corpus by means of identifying the terms used and looking at the localized counterpart.

The terms under study are the original terms used in previous studies (Jiménez-Crespo 2017; Jiménez-Crespo & Tercedor 2017). These terms were originally selected at random with the Excel function from the VariMed terminological variation website from an initial list of 100 concepts:

(1) Dysmenorhea – 'dismenorrea' : painful periods
(2) Dyspnoea – 'disnea' : difficulty breathing
(3) Halitosis – 'halitosis' : bad breath

(4) Hematuria – 'hematuria' : presence of blood in the urine
(5) Hypoglycemia – 'hipoglucemia' : low blood sugar
(6) Hysterectomy – 'histerectomía' : surgical removal of the uterus
(7) Myalgia – 'mialgia' : muscle pain
(8) Polydipsia – 'polidipsia' : excessive thirst
(9) Polyuria – 'poliuria' : excessive volume of urine
(10) Rhinorrhea – 'rinorrea' : nasal dripping
(11) Tachypnea – 'taquipnea' : fast breathing
(12) Xerosis – 'xerosis' : dry skin
(13) Xerostomy – 'xerostomía' : dry mouth

All instances of use of these terms in both the source and target texts were recorded in a database. Only segments using the LG terminological variant in one or both languages were annotated. This means it is possible that cases with two or more lay terms that could also involve explicitation- implicitation shifts were not recorded. Nevertheless, due to the lack of lexicalization of some these lexical units (e.g. hypoglycemia = low blood sugar, low sugar in the blood, abnormally low level of sugar in the blood, etc.), as well as the wide range of possible lay term variants for the different concepts under study, it was decided to focus on LG terms only. The record in the database included both source and target segments, as well as any possible translation shifts (see next section). The two types of recorded phenomena were implicitation - explicitation shifts and shifts in the degree of specialization (see next section). The analyses started from a search in Spanish for the probing medical terms, and once the page opened, the English source webpage was analyzed. In a second search, the direction was the inverse in order to identify LG terms in the source texts that might not have been translated into a LG term in Spanish. The translation pair recorded was the first appearance of the LG term per webpage. There is usually a repetition of the same translation solution for a single concept in a webpage, and therefore only the first instance in the running text when the concept is first introduced or used was recorded.

4.2. Categorization of degree of specialization and explicitation shifts

The analysis conducted included two interrelated steps. The first one involved the identification of shifts in the degree of specialization. This first step was necessary as explicitation and implicitation shifts overall are a type of shift in which the translation displays a higher or lower degree of specialization, for example when an LG term is substituted in the translation by a lay term or vice versa. The use of LG terms in specialized texts contributes to their degree of specialization and terminological density. It is understood to be an indication of the degree of specialization of a text (Cabré 1999: 89). In this context and as previously explained, the translation of any medical concept can include shift in implicitness or explicitness when a reformulation is added or omitted in the translation. For example, the following translation, Myalgia > 'Mialgia (dolor muscular)' [Myalgia (muscle aches)], shows a more explicit formulation with the LG term accompanied by a reformulation or layterm. This simultaneously contributes to lowering slightly the degree of specialization by combining a specialized term with a lay one. Nevertheless, in the corpus it is also possible to identify shifts in the degree of specialization that do not involve explicitation o implicitation. For example, in the following instance from the corpus: air hunger (feeling that you cannot get enough air) > 'Disnea (sensación de que uno no recibe suficiente aire)', the layterm *air hunger* followed by a reformulation that is translated by using the LG term in Spanish, 'disnea', followed by a literal rendition of the intralingual explicitation found in the source text. This represents a clear case of degree of specialization shift that does not involve an explicitation-implicitation shift in the translation.

These shifts in the degree of specialization can include any permutation between LG terms and lay terminological variants. There were three possible categories for degree of specialization shifts: 1. Same degree of specialization, 2. Lower to higher, 3. Higher to lower. Figure 1 shows the possible shifts within each category.

Table 1. Categorization of possible shifts in the degree of specialization

Shift	Degree of Specialization		
	Same 1. S-Ø	Lower to Higher 2. L>H	Higher to Lower 3. H>L
No Shift- Direct translation	1. S-Ø		
Substitution		2a.L>H-Subst	3a. H>L-Subst
Explicitation		2b.L>H-Expl	3b. H>L-Expl
Implicitation		2c.L>H-Impl	3c. H>L-Impl
Doublet Inversion		2d.L>H-Doublet.Inv	3d.H>L-Doublet.Inv

The first category, 1. S-Ø involves no change in the degree of specialization, that is, the LG term stays the same in the translation. This involves a direct or literal translation of the source text, maintaining the same patterns of lexical use as in the source text, both in the single use of the LG term or the combined used of an LG term with its reformulation or intralingual lexical explicitation.

1. Degree of specialization shift: Ø same
   Hypoglycemia > 'Hipoglucemia'
   Hypoglycemia, or low blood sugar > 'Hipoglucemia, o azúcar bajo en la sangre'

Categories 2 (L>H) and 3 (H>L) involve shifts from lower to higher degree of specialization or vice versa. They are subdivided into four possible subcategories that involve a shift by means of using (a) substitution, (b) explicitation, (c) implicitation, or (d) doublet inversion.

Substitution (2a and 3a), involves for example a shift by means of substituting a lay term, such as shortness of breath, for the corresponding LG term in the Spanish translation, 'disnea'. This is a case of increasing the degree of specialization, 2a.L>H-Subs, but this can also happen in the opposite direction, with the LG term substituted entirely by a lay term, 3a.L>H-Subs.

2.a. L>H-Subs. Degree of Specialization/Low to High/ Substitution
   …if you experience *shortness of breath* > '… si tiene *disnea*'.

...has trouble breathing (*shortness of breath*) > … 'tiene dificultad para respirar (*disnea*)'.

Explicitation (subcategories 2b and 3b) involve adding a doublet, either a LG term to a lay term or vice versa. It can involve adding a lay term or reformulation to an existing LG term in the source text, lowering the degree of specialization, 3b. H>L-Expl. This is considered as one of the explicitation categories, as from one term the translation rendering includes two. It is also possible that the translation can involve a LG term added to a lay term, 2b.L>H-Expl. This is also considered as one of the explicitation categories.

2.b. L>H-Expl. Degree of Specialization/Lower to Higher /Explicitation
Shortness of breath > 'disnea (dificultad para respirar)' [dyspnea (difficulty breathing)]

Implicitation (subcategories 2.c. and 3.c.) involve implicitation shifts by eliminating-deleting one part of a doublet, going from LG term plus reformulation in any order to just one. This is a less explicit formulation and therefore identified as implicitation in this study. Depending on which part is deleted the degree of specialization is increased, 2c.L>H-Impl, or lowered 3c.H>L-Impl.

2.c. L>H-Impl Degree of Specialization/Low to High/ Implicitation
- Low blood sugar (hypoglycemia) > 'Hipoglucemia'
3.c. H>L-Impl.
- Myalgia > 'Mialgia (dolor muscular)'

Double inversion (subcategories 2.d. and 3.d.) involves a minimal shift in the degree of specialization within the low to high continuum. Here the LG term and its reformulation are reversed, presenting the LG term first, rather than second, or the other way around. This involves no explicitation shift, but rather an inversion as it might be perceived that the LG term is more transparent to target text users to be presented first.

2.d. Degree of Specialization/Lower to Higher/ Doublet Inversion.
- Low blood sugar (hypoglycemia)' > Hipoglucemia (nivel bajo de azúcar en la sangre)'

After this categorization of the shifts in the degree of specialization, categories 2.b, 2.c., 3.b and 3.c would be representative of more or less explicit formulations, and therefore indicative of the tendencies under study here.

Table 2. Categorization of explicitation and implicitation shifts

Explicitation/Implicitation Shifts		Target Text Shifts	
Source Texts	Intralingual Explicitation (2 or more terms / expressions to render a concept)	Ø Literal translation	Implicitation
	No intralingual explicitation (1 term to render a concept)	Ø Same. Literal translation	Explicitation

The analysis of the data yielded four possible types of explicitation/implicitation shifts. The source text can display either a single medical term variant, or can use two or more. Explicitation shifts will be observed if the translation displays more than one terminological variant or a reformulation, while all instances of source texts that display intralinguistic explicitation (Hill-Madsen 2015) can be rendered as a single terminological or reformulation variant.

The first category would be observed if there is a literal translation of a doublet such as:

(1) Shortness of breath (dyspnea) > 'Falta de aire (disnea)'
(2) A lower-than-normal blood glucose level (hypoglycemia) > 'Un nivel de glucosa en la sangre inferior a lo normal (hipoglucemia)'

Meanwhile, an implicitation shift would emerge if any of the two parts would be deleted.

(1) Shortness of breath (dyspnea) > 'Disnea'

This case is not limited to instances of two terms or lexical units in which one is omitted, since some examples in the corpus include up to three reformulations (e.g. "Dyspnea is the feeling of difficult or uncomfortable breathing or of not getting enough air" (EN)- "La disnea es la sensación de dificultad o incomodidad para respirar, o de no conseguir suficiente aire".

The omission of any of these three lexical units could be considered as an implicitation shift.

On the other hand, any instance of a single LG term or lay variant can be more explicit in the translation if it is accompanied by a second variant or reformulation.

(1) Low blood sugar > 'un nivel bajo de azúcar en la sangre (hipoglucemia)' [a low level of sugar in the blood (hypoglycemia)]
(2) Myalgia > 'mialgia (dolor muscular)' [myalgia (muscle pain)].

This study thus clearly defines what is explicit and implicit (De Metsenaere & Vandepitte 2017: 387), namely LG reformulation or determinologization (or the omission thereof) in the translation. As previously mentioned, given the different level of specialization attributed to many LG terms between Spanish and English, it should be expected that in the English to Spanish direction the reformulation of the LG term could be sometimes deleted in the translation process (Montalt & González Davies 2005; Campos Andrés 2013). Therefore, the decision to insert or delete a doublet or the use/absence of a determinologization strategy with a specialized term in a semi-specialized text is considered indicative of explicitness or implicitness in translation.

5. Results

The results of the parallel corpus analysis point to 316 instances in which the selected LG terms were translated in the corpus from English into Spanish. These cases identified might seem in principle a small volume considering the corpus size, but it should be remembered that only the selected medical terms in the previous study were used, and only the first instance per webpage was selected.

5.1. Results: Shifts in the degree of specialization or technicality

Table 3 shows the results of the contrastive analysis of degree of specialization or technicality. In 80.69% of the instances (255 of the total 316) the translation did not present a shift in the degree of specialization, that is, the translation showed a direct or literal rendering of the source text, as compared to 19.31% of shifts (61 of 316 instances). Within this, in 19% of

instances of medical concepts under study there was a shift from lower to higher degree of specialization, while only 0.31% of cases involved lowering the degree of specialization. This means that translated texts showed a marked tendency to keep the same degree of specialization as the source text by means of replication of source language structures. In those cases where a shift occurs, the most frequent one is to increase the level of specialization by adding a doublet, that is, translating a lay term in English by adding also the LG term, that is explicitation (52.46% of the shifts, 10.12% of the total instances). The next most frequent one involved going from lower to higher degree of specialization by substituting a lay term for a LG term (44.26% of shifts, 8.54% of total instances). Last but not least, another shift that does not involve explicitation or implicitation but that is argued here as a shift, a doublet inversion from higher to lower degree of specialization, is present in 0.31% of cases. This shift often involves an LG term followed by a reformulation or lay term that is reversed in the translated text.

Table 3. Results. Degree of specialization shifts

Degree of Specialization Total instances N=316		
No Shift	**Shift**	
80.69% N=255	**Low to High L>H (19%) N=60**	**High to Low H>L (0.31%) N=1**
	2a.L>H-Subst **44.26% (8.54 of total instances)** N=27	3.a. H>L-Subst **0%** N=0
	2b.L>H-Expl **52.46% (10.12% of total instances)** N=32	3b. H>L-Expl **1.64% (0.31 of total instances)** N:1
	2c.L>H-Impl **0%** N=0	3c. H>L-Impl **0%** N=0
	2d.L>H-Doublet.Inv **1.64%** N=1	3d.H>L-Doublet.Inv **0%** N=0

In sum, if the results are compiled 19% of instances of medical concepts in this study showed a lower to higher degree of specialization shift. The reverse option, a shift from higher to lower degree of specialization only appeared in 0.31% of cases. It should be noted that explicitation shifts in terms of shifts in the degree of specialization are the most common, and that implicitation is not found in the corpus under study. These results therefore also confirm H2, as explicitation found in translated texts is due to a combination of ST structure replication, as well to translation-inherent explicitation.

5.2. Results: Analysis of implicitation- explicitation shifts

In the dataset for this study (316 LG terms), 153 instances in the dataset would allow for implicitation shifts, as the ST includes a doublet or triplet, that is, the medical concept is rendered in the ST through multiple terms. This could be considered cases of intralingual explicitation (Hill-Madsen 2015). Meanwhile, 165 cases would allow for explicitation shifts as the medical concept is rendered through a monolexical terminological unit.

The results show that when explicitation is possible, a shift is identified in 20.6% (34/165) of possible cases, while there are no implicitation shifts (0/147) ($p = < 0.0001$)[2]. Explicitation shifts are therefore quite more prevalent than implicitation ones in the dataset. This higher tendency for explicitation can be even found in instances in which a doublet in the ST has been rendered with three terms or lexical units, such as the following:

(1) This condition is called *shortness of breath*. The medical term for this is *dyspnea*. > 'Esta afección se denomina *falta de aliento* o *dificultad para respirar*. El término médico para esto es *disnea*' [This condition is known as shortness of breath or difficulty breathing. The medical term is dyspnea]

2. Two-tailed Fisher's exact test.

Table 4. Results of explicitation and implicitation shift analysis in parallel corpus

Explicitation/Implicitation Shifts		Target Text Shifts	
Source Texts	Intralingual Explicitation	Ø Literal translation N:147	Implicitation N:0 0%
	No intralingual explicitation	Ø Same. Literal translation N: 131	Explicitation N: 34 20.6%

In this case the lay term, shortness of breath, followed by an explanation that indicates that the medical term for this is *dyspnea*, is translated by using two layterms in Spanish, 'falta de aliento' and 'dificultad para respirar', followed also by the medical term *dyspnea*. It should be mentioned that in the corpus there were instances of deletion, primarily in pronunciation information for English speaking patients that was deleted in the translation. Spanish target readers can easily read any LG term and therefore that aid for source language readers is not necessary and deleted "dyspnea (DISP-nee-uh): Difficult, painful breathing or shortness of breath" > 'disnea: Respiración difícil, dolorosa o deficiencia respiratoria.

6. Discussion

In the previous comparable corpus study (Jiménez-Crespo & Tercedor 2017), LG terms in translations were accompanied by a reformulation in 40.6% of cases, while it was 21.23% for those found in the non-translational corpus. This means that translations contained overall higher rates of explicitation of LG terms, and whether it was the direct result, or not, of the translational process could not be identified with that comparable corpus. These results show that when the concepts under study appear in source or target texts, 10.44% of overall cases can be attributed to translation-specific or translation-inherent explicitation. These results, although not fully comparable, do not match the difference between the translational and non-translational ratios found in the previous study. The remaining cases can therefore be attributed to cross-linguistic effects (Kruger & Van Rooy 2016). This can therefore be indicative that (1) the translation process leave unmistakable

traces of explicitation in translation products, (2) implicitation shifts do not appear in translation products under study, even when it could be possible, (3) higher levels of explicitation in translated texts when compared to non-translated ones are due to cross-linguistic effects, that is, translators simply replicate existing source text structures regardless of possible language specific differences in the genre under examination and (4) combining comparable and parallel corpora can provide a clearer and richer picture of general tendencies and potential causes.

The results therefore offer support for H1, as translated texts show higher ratios of explicitation than implicitation. In addition, this preference appears in a specific case in which implicitation might be suitable or might be a preferred option in the target language.

RQ 2 indicated that if H2 would be confirmed, it would offer support for the interference or cross-linguistic effects hypothesis, as well as for the risk-aversion hypothesis. It is clear that most cases of explicitation are due to cross-linguistic effects, thus offering support for this proposed hypothesis. The translations do show traces of the existing contrastive differences between both languages, as LG terms are less commonly used in Spanish than in English. As far as the risk aversion hypothesis, the picture that emerges is less clear and needs to be explored in future studies. It is clear that explicitation is preferred to implicitation, and translations are more explicit than non-translations. When it is possible to explicitate, 20.6% of cases include explicitation, with no cases of implicitation. Nevertheless, this study has also shown that explicitation strategies go hand in hand with shifts in the degree of specialization, such as changing a layterm in the source text for a LG term (8.54%) (e.g. 'muscle pain' translated as 'myalgia' in Spanish). This raises the question of how explicitation interacts with other lexical shifts in terms of risk avoidance or making the text more clear or easy to understand. Approximately half the cases involve explicitation shifts by adding a LG term to a lay lexical unit, a strategy that might lead to increase the ease of comprehension, while the other half solely involve substituting a lay term for an LG term. It is quite possible that the explicitation ratios here are due to risk avoidance, but it is not fully clear then, as well as a potential question for future studies, whether risk avoidance strategies can be confirmed using only one possible strategy out of many possible factors. What

is clear is that studying lexical explicitation with other lexical shifts might provide a broader and more complete picture.

7. Final remarks

The motivation for this study was to offer a clearer insight into the differences in the frequency of use of LG terms and explicitation ratios between translated and non-translated texts as described in a previous study (Jiménez-Crespo & Tercedor 2017). This paper has shown that the observed differences can primarily be attributed to language and genre specific differences in the language pair involved, that is, they can be attributed to direct transferring of the source text structures, with a percentage of what can be called translation-inherent explicitation (20.6% of possible cases), that is, shifts seem to emerge due to the specific communicative nature of the translation process itself. This general tendency is further supported by the fact that there were no cases of implicitation, even when this strategy was possible and/or appropriate. These results thus add to the body of literature that confirms explicitation as a general tendency in translation. They also offer support for the cross-linguistic effect (Kruger & Van Rooy 2016) as a possible causality of higher explicitation rates in translations than in similar non-translated texts, as translations primarily replicate source text structures in English source texts. The study offers a more lukewarm support for the risk aversion hypothesis, with one fifth of possible cases using explicitation. This study has shown that other types of lexical shifts might interact with these explicitation shifts, the objective of future studies. What is clear is that the combination of the parallel and comparable corpus methodologies has offered strong support for the existence of translation-inherent explicitation, and a combination of both methodologies seems suitable for the study of these phenomena.

References

ALARCÓN, Esperanza; Clara Inés López-Rodríguez & Maribel Tercedor. (2016) "Variation dénominative et familiarité en tant que source d'incertitude en traduction médicale." *Meta* 61:1, pp. 117-145.

ASKEHAVE, Inger & Karen K. Zethsen. (2000) "Medical Texts Made simple – Dream or Reality?" *Hermes: Journal of Linguistics* 23, pp. 63-74.

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

BAKER, Mona. (1996) "Corpus-based Translation Studies: The Challenges that Lie Ahead." In: Somers, Harold (ed.) (1996) *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager.* Amsterdam-Philadelphia: John Benjamins, pp. 175-186.

BECHER, Viktor. (2010) "Abandoning the notion of "translation-inherent" explicitation. Against a dogma of translation studies." *Across Languages and Cultures* 11:1, pp. 1-28.

BECHER, Viktor. (2011) *Explicitation and Implicitation in Translation. A Corpus-based Study of English-German and German-English Translations of Business Texts.* PhD thesis. Hamburg: Universität Hamburg, Fakultät für Geisteswissenschaften.

BISIADA, Mario. (2017) "Universals of editing and translation." In: Hansen-Schirra, Silvia; Oliver Czulo & Sascha Hofmann (eds.) (2017) *Empirical modelling of translation and interpreting.* Berlin: Language Science Press, pp. 241–275. DOI:10.5281/zenodo.1090972

BLUM-KULKA, Shoshana. (1986) "Shifts of Cohesion and Coherence in Translation." In: House, Julianne & Shoshana Blum-Kulka (eds.) (1986) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies.* Tübingen: Gunter Narr, pp. 17-35.

CAMPOS ANDRÉS, Olga. (2013) "Procedimientos de desterminologización: traducción y redacción de guías para pacientes". *Panace@* 14, pp. 48-52.

CHESTERMAN, Andrew. (2004) "Beyond the Particular." In: Mauranen, Anna & Pekka Kujamäki (eds.) (2004) *Translation Universals. Do they Exist?* Amsterdam-Philadelphia: Benjamins, pp. 33-49.

DE METSENARERE, Hinde & Sonia Vandepitte. (2017) "Towards a Theoretical Foundation for Explicitation and Implicitation." *Trans-Komm* 10:3, pp. 385-419.

DE SUTTER, Gert & Patrick Goethals; Torsten Leuschner & Sonia Vandepitte. (2012) "Towards methodologically more rigorous corpus-based translation studies." *Across Languages and Cultures* 13:2, pp. 137-143.

DE SUTTER, Gert & Marie-Aude Lefer. (2020) "On the need for a new research agenda for corpus-based translation studies: a multi-methodological, multifactorial and interdisciplinary approach." *Perspectives* 28:1, pp. 1-23

DENTURK, Kathelenje. (2012) "Explicitation vs. implicitation: a bidirectional corpus-based analysis of causal connectives in French and Dutch translations." *Across Languages and Cultures* 13, pp. 211-227.

ENGLUND-DIMITROVA, Birgitta. (2005) *Expertise and Explicitation in the Translation*. Amsterdam-Philadelphia: John Benjamins.

ERFIENI, Ni Made Diana. (2017) "Explicitation And Implicitation In The Literature Translation For Children." *LITERA: Jurnal Litera Bahasa Dan Sastra* 3:1, pp. 31-39.

EWA, Gumul. (2006) "Explicitation in Simultaneous Interpreting: A Strategy or a By-Product of Language Mediation." *Across Languages and Cultures* 7:2, pp. 171-190.

EZPELETA, Pilar. (2012) "An Example of Genre Shift in the Medicinal Product Information Genre System." *Linguistica Antverpiensia, New Series Themes in Translation Studies* 11, pp. 139-159

FABER, Dorrit & Mette Hjort-Pedersen. (2013) "Expectancy and Professional Norms in Legal Translation: A Study of Explicitation and Implicitation Preferences." *Fachsprache* 2013 1-2, pp. 43-63.

GUTIÉRREZ RODILLA, Bertha. (2014) "El lenguaje de la medicina en español: cómo hemos llegado hasta aquí y qué futuro nos espera". *Panace@* 15, pp. 86-94.

HALVERSON, Sandra. (2016) "Cognitive Translation Studies and the merging of empirical paradigms." *Translation Spaces* 4, pp. 310-340.

HANSEN-SCHIRRA, Silvia; Stella Neumann & Erich Steiner. (2007) "Cohesive Explicitness and Explicitation in an English-German Translation Corpus." *Languages in Contrast* 7:2, pp. 241-266.

HILL-MADSEN, Aage. (2015) "Lexical Strategies in Intralingual Translation between Registers." *Hermes – Journal of Language and Communication in Business* 54, pp. 85-105.

HJORT, Pedersen; Mette Faber & Dorrit Faber. (2010) "Explicitation and Implicitation in Legal Translation. A Process Study of Trainee Translators." *Meta: Translators Journal* 55:2, pp. 237-250.

Jiménez-Crespo, Miguel A. (2011) "The future of general tendencies in translation: Explicitation in web localization." *Target* 23:1, pp. 3–25.

Jiménez-Crespo, Miguel A. (2012) "Loss or lost in localization: A corpus-based study of original and localized non-profit websites." *JoStrans: the Journal of Specialised Translation* 17, pp. 136-165.

Jiménez-Crespo, Miguel A. (2016) "Testing explicitation in translation: triangulating corpus and experimental studies." *Across Languages and Cultures* 16:1, pp. 257-283.

Jiménez-Crespo, Miguel A. & Maribel Tercedor. (2017) "Lexical Variation, Register and Explicitation in Medical Translation: A Comparable Corpus Study of Medical Terminology in US Websites Translated into Spanish." *TIS: Translation and Interpreting Studies* 12: 3, pp. 405-426.

Klaudy, Kinga. (1998) "Explicitation." In: Baker, Mona (ed.) (1998). *Encyclopedia of Translation Studies*. London: Routledge, pp. 80-85.

Klaudy, Kinga. (2001) "The asymmetry hypothesis: Testing the asymmetric relationship between explicitations and implicitations." *Paper presented at the Third EST Congress, Claims, changes and challenges in translation studies,* Copenhagen.

Klaudy, Kinga & Kristina Károly. (2005) "Implicitation in translation: Empirical evidence for operational asymmetry in translation." *Across Languages and Cultures* 6:1, pp. 13-28.

Krüger, Ralph. (2014) *Exploring the interface between scientific and technical translation and cognitive linguistics: The case of explicitation and implicitation*. University of Salford. Unpublished PhD thesis.

Krüger, Ralph. (2015) *The interface between scientific and technical translation studies and cognitive linguistics: with a special emphasis on explicitation and implicitation as indicators of translations text-context interaction*. Berlin: Frank Timme.

Krüger, Ralph. (2016) "The textual degree of technicality as a potential factor influencing the occurrence of explicitation in scientific and technical translation." *Jostrans: The Journal of Specialized Translation* 26, pp. 96-115.

Kruger, Heide & Bertus Van Rooy. (2016) "Syntactic and pragmatic transfer effects in reported-speech constructions in three contact varieties of English influenced by Afrikaans." *Language Sciences* 56, pp. 118-131. https://doi.org/10.1016/j.langsci.2016.04.003

KRUGER, Heide. (2018) "*That* Again: A Multivariate Analysis of the Factors Conditioning Syntactic Explicitness in Translated English." *Across Languages and Cultures* 20:1, pp. 1-33. https://doi.org/10.1556/084.001.

KRUGER, Heide & Gert De Sutter. (2018) "Reconceptualising that-omission in translated and non-translated English using the MuPDAR approach." *Translation, Cognition and Behaviour* 1:2, pp. 251-290.

KRUGER, Heide & Gert De Sutter. (2018) "Alternations in contact and non-contact varieties: Reconceptualising that-omission in translated and non-translated English using the MuPDAR." *Translation, Cognition & Behavior* 1:2, pp. 251-290.

LAVIOSA, Sara. (2002) *Corpus-Based Translation Studies. Theory, Findings, Applications*. Amsterdam-New York: Rodopi.

MAKKOS, Aniko & Edina Robin. (2014) "Explicitation and Implicitation in Back-translation." *Current Trends in Translation Teaching and Learning* 5, pp. 151-182.

MAREVA, Amelia. (2017) *Lexical Explicitation and Implicitation in Experts' and Students' Literary Translations: An Empirical Contrastive Study*. Working Paper. New Bulgarian University Scholar Electronic Repository, Sofia. Electronic version: <https://scinapse.io/papers/2778145166>

MEYER, Ingrid & Kristen Mackintosh. (2000) "When terms move into our everyday lives: an overview of determinologization." *Terminology* 6:1, pp. 111-138.

NISBETH Jensen; Matilde Zethsen & Karen Zethsen. (2012) "Patient Information Leaflets: Trained Translators and Pharmacists-cum-translators – a comparison." *Linguistica Antverpiensia New Series. Themes in Translation Studies* 11, pp. 31-49.

MONTALT, Vicent & María González Davies. (2007) *Medical Translation Step by Step. Translation Practices Explained*. Manchester: St. Jerome Publishing.

MONTALT, Vincent & Mark Shuttleworth. (2012) "Translation and knowledge mediation in medical and health settings." *Linguistica Antverpiensia New Series – Themes in Translation Studies* 11, pp. 95-112.

MONTALT, Vicent; Karen Zethsen & Wioleta Karwacka. (2018) "Medical Translation in the 21st Century: Challenges and Trends." *MonTI* 10, pp. 27-42.

MUÑOZ-MIQUEL, Ana. (2012) "From the Original Article to the Summary for Patients: Reformulation Procedures in Intralingual Translation." *Linguistica Antverpiensia, New Series Themes in Translation Studies* 11, pp. 187-206.

MURTISARI, Elisabeth T. (2013) "A relevance-based framework for explicitation and implicitation: An alternative typology." *Trans-kom* 6:2, pp. 315-344.

MURTISARI, Elisabeth T. (2016) "Explicitation in Translation Studies: The journey of an elusive concept." *Translation & Interpreting* 8:2, pp. 64-81.

OAKES, Michael & Meng Ji. (2012) *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*. New York-London: Routledge.

OLOHAN, Maeve. (2004) *Introducing Corpora in Translation Studies*. London: Routledge.

OLOHAN, Maeve & Mona Baker. (2000) "Reporting that in translated English: Evidence for subconscious processes of explicitation?" *Across Languages and Cultures* 1:2, pp. 141-158.

ØVERVERÅS, Linn. (1998) "In search of the third code: An investigation of norms in literary translation." *Meta* 43:4, pp. 557-570.

PEREGO, Elisa. (2003) "Evidence of Explicitation in Subtitling: Towards a Categorisation." *Across Languages and Cultures* 4:1, pp. 63-88.

PUURTINEN, Tina. (2004) "Explicitation of clausal relations: A corpus-based analysis of clause. Connectives in translated and non-translated Finnish children's literature." In: Mauranen, Anna & Pekka Kujamäki (eds.) (2004) *Translation universals: Do they exist?* Amsterdam-Philadelphia: John Benjamins, pp. 165-76.

PYM, Anthony. (2005) "Explaining explicitation." In: Karoly, Kristina & Agota Foris (eds.) (2005) *New Trends in Translation studies: In Honour of Kinga Klaudy*. Budapest. Akademiai Kiado, pp. 29-34.

PYM, Anthony. (2015) "*Translating as Risk Management*." *Journal of Pragmatics* 85, pp. 67-80.

TANG, Fang. (2018) *Explicitation in Consecutive Interpreting*. Amsterdam-Philadelphia: John Benjamins.

TAO, Yuan & Zanhao Jiang. (2017) "Translation universals of kak structures: a corpus-based approach". *Russian Linguist Linguistics* 41:1, pp. 61-78.

TERCEDOR, Maribel & Clara Inés López Rodríguez. (2012) "Access to health in an intercultural setting: the role of corpora and images in grasping term variation." *Linguistica Antverpiensia, New Series-Themes in Translation Studies* 11, pp. 247-268.

Tercedor, María Isabel; Clara Inés López Rodríguez & Juan A. Prieto Velasco. (2014) "También los pacientes hacen terminología: retos del proyecto VariMed". *Panace@* 15, pp. 95-102.

Van Beveren, Amelie; Gert De Sutter & Thimoty Colleman. (2018) "Questioning explicitation in translation studies: a multifactorial corpus investigation of the om-alternation in translated and original Dutch." *UCCTS 2018 Conference*, University of Louvain, 12-14 Sept. 2018.

Van Dam, Helle; Matilde Brogger & Karen Zethsen. (2018) *Moving Boundaries in Translation Studies*. New York-London: Routledge.

Vinay, Jean Paul & Jean Dalbernet. (1958) *Stylistique comparée du français et de l'anglais*. Paris: Didier.

Zethsen, Karen K. (2004) "Latin –Based Terms: True or False Friends?" *Target* 16:1, pp. 125-142.

BIONOTE / BIONOTA

Miguel A. Jiménez-Crespo holds a PhD in Translation and Interpreting Studies from the University of Granada, Spain. He is a Professor in the Department of Spanish and Portuguese, Rutgers University, and he directs the MA program and the undergraduate certificate in Spanish – English Translation and Interpreting. He is the author of *Crowdsourcing and Online Collaborative Translations: Expanding the Limits of Translation Studies* (translated into Korean) published by John Benjamins in 2017, as well as *Translation and Web Localization* published by Routledge in 2013.

Maribel tercedor Sanchez is full professor at the Department of Translation and Interpreting of the University of Granada, where she teaches Multimedia and Scientific and Technical translation. She has directed the VariMed project on medical terminology (http://varimed.ugr.es) and codirected with Clara I. López Rodríguez the CombiMed project on lexical combinations in Medicine. Her main research interests are in the fields of lexical and cognitive aspects of scientific and technical translation, terminology (variation and phraseology) and accessibility in translation. She is the author of a number of academic papers in these fields.

MIGUEL A. JIMÉNEZ-CRESPO es doctor en Traducción e Interpretación por la Universidad de Granada. Es catedrático en el Departamento de Español y Portugués en la Universidad de Rutgers, EEUU, donde dirige el programa de traducción e interpretación. Es el autor del libro *Crowdsourcing and Online Collaborative Translations: Expanding the Limits of Translation Studies*, publicado en 2017 por John Benjamins y de *Translation and Web Localization* (tradudico al coreano) publicado en 2013 por Routledge.

MARIBEL TERCEDOR SÁNCHEZ es catedrática en el Departamento de Traducción e Interpretación de la Universidad de Granada. Imparte docencia de Traducción Multimedia y Traducción Científica y Técnica. Ha dirigido el proyecto VariMed sobre terminología médica y codirigido, con la Dra. López Rodríguez, el proyecto CombiMed sobre combinaciones léxicas en Medicina. Sus intereses en investigación se centran en los ámbitos del léxico y cognición en la traducción de textos científicos y técnicos, terminología (variación y fraseología) y accesibilidad en traducción, ámbitos en los que ha publicado sus resultados de investigación.

Recibido / Received: 30/05/2020
Aceptado / Accepted: 05/08/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.03

Para citar este artículo / To cite this article:
Arias-Badia, Blanca. (2021) "Using corpus pattern analysis for the study of audiovisual translation: A case study to illustrate advantages and limitations." En: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection*. *MonTI* 13, pp. 93-113.

USING CORPUS PATTERN ANALYSIS FOR THE STUDY OF AUDIOVISUAL TRANSLATION: A CASE STUDY TO ILLUSTRATE ADVANTAGES AND LIMITATIONS

Blanca Arias-Badia
blanca.arias@upf.edu
Universitat Pompeu Fabra

Resumen

La identificación de normas, es decir, de patrones recurrentes de comportamiento, es un objetivo común de la Lingüística, los Estudios de Traducción y los Estudios de Televisión. Este artículo ofrece una revisión metodológica crítica de la aplicación del análisis de patrones de corpus (CPA), una técnica propuesta desde el campo del análisis léxico, al estudio de la traducción audiovisual. El artículo ilustra la aplicación de esta técnica de investigación mediante la presentación de un estudio de caso que busca la identificación de combinatoria léxica anómala en el diálogo en inglés y en los subtítulos en español de tres series de televisión por medio de CPA. Después de ilustrar la metodología, el énfasis de la evaluación recae en la reflexión acerca de las ventajas y limitaciones que el uso de CPA conlleva para el estudio de la traducción audiovisual.

Palabras clave: Análisis de patrones en corpus; Traducción audiovisual; Subtitulación; Combinatoria léxica anómala; Estudios Descriptivos de la Traducción.

Abstract

The identification of norms, i.e. of recurrent patterns of behaviour, is a common concern for Linguistics, Translation Studies, and Television Studies. This paper offers

a critical methodological review after applying Corpus Pattern Analysis (CPA), a technique proposed from the field of lexical analysis, to the study of audiovisual translation. The paper illustrates the application of this research technique by presenting a case study pursuing the identification of anomalous collocations in the English spoken dialogue and in the Spanish subtitles of three television series by means of CPA. After illustration, emphasis of the methodological assessment is placed on the discussion of advantages and limitations entailed in the use of CPA for the study of audiovisual translation.

Keywords: Corpus Pattern Analysis; Audiovisual translation; Subtitling; Anomalous collocates; Descriptive Translation Studies.

1. Introduction[1]

The identification of norms, i.e. of recurrent patterns of behaviour, is a common concern for Linguistics and Translation Studies. Corpus linguistics has been proved to serve as a powerful methodological tool or research methodology (Olohan 2004; Walsh, Morton & O'Keeffe 2011) to account for linguistic and translational norms. Corpus-driven approaches in Translation Studies propose a bottom-up analysis of lexicogrammatical and phraseological patterns in source texts and translations by keeping preconceptions about the texts in hand to a minimum (Corpas 2008: 53).

Television Studies are also interested in norms. The discipline has been nourished by Literary Theory on the notion of genre and reflects on shared conventions governing television text production (Fiske 1987, Feuer 1992, Neale and Turner 2001). It is well known that genre entails recognition (Wolf 1984: 190): audiences of crime TV shows expect to see weapons, uniforms and handcuffs on screen as much as they expect to hear sirens, doors howling or shouts in every episode. Genre and format norms provide the necessary framework for recognition in the television medium.

This paper assesses advantages and limitations of applying a methodology proposed from the field of Linguistics and, more specifically, from

1. This research is framed in the Neómetro project, ref. FFI2016-79129-P (Ministerio de Economía y Competitividad/AEI/FEDER). The author is a member of InfoLex, a research group funded by the Catalan Government, under the SGR funding scheme (2017SGR01366).

lexical analysis, to the study of audiovisual translation. Namely, Corpus Pattern Analysis (CPA) (Hanks 2004, 2013) is the methodological framework proposed in order to determine when screenwriters or subtitlers opt either for normal usage, i.e. "the prototypical syntagmatic patterns with which words in use are associated" (Hanks 2004: 87), or for creativity in television dialogue and its translation.

The following sections are structured as follows: Section 2 starts by framing CPA; Section 3 reports on a case study conducted using CPA; Section 4 summarises the conclusions drawn from applying this technique to a corpus of audiovisual translation; and Section 5 provides closing remarks.

2. Corpus Pattern Analysis (CPA)

The case study reported in Section 3 employs notions from the Theory of Norms and Exploitations (TNE). This approach to the study of lexicon is based on the idea that, when "making meanings", the users of a language either conform to norms, i.e. "pattern[s] of ordinary usage in everyday language with which a particular meaning or implicature is associated" (Hanks 2013: 92), or exploit such norms creatively by means of "a deliberate departure from an established pattern of normal word use, either in order to talk about new or unusual things or in order to say old things in a new, interesting, or unusual way" (*ibid*: 212). Since they are instances of unpredictable uses of language, exploitations may be the result of various linguistic processes. Hanks (*ibid*: 215-250) provides a description of some kinds of exploitations and their link to classical tropes. The present paper focuses on one specific creative device, i.e. anomalous collocates, which will be defined in Section 3.

As a methodological application of TNE, Corpus Pattern Analysis (CPA) aims to build an inventory of normal patterns of word use. Its "objective is to identify, in relation to a given target word, the overt textual clues that activate one or more components of its meaning potential" (Hanks 2004: 6). The analytic procedure to apply CPA is the following: first, concordances for each target word are scanned; then a random sample is scrutinised and annotated by the lexicographer. This methodology is ultimately based on Sinclair's (2003) proposal of a procedure for the analysis of corpus concordances. In his

seven-step guidelines, the author notes that interpretation (i.e. formulation of hypotheses after concordance reading and before consolidation by means of consultation in other sources) is an inextricable part of corpus analysis. On the one hand, CPA annotation focuses on the identification and classification of normal patterns of behaviour; on the other, it accounts for exploitations of each pattern found in the random sample under analysis (*ibid*: 9).

The *Pattern Dictionary of English Verbs* (PDEV) is currently being developed following this methodology. Verbs have been given priority in applying CPA because they are the pivot of the clause and "we make conversation by using clauses" (Hanks n.d.: 12). Each entry in the PDEV shows the patterns of the verb found on a sample of at least 250 corpus occurrences —the more frequent the verb, the larger the sample— and the implicature of each pattern. The pattern displays normal semantic types in combination with the verb and the implicature glosses the meaning of the verb for each pattern. Consider the pattern and implicature of the verb *listen* below:

> **Pattern:** Human | Institution *listen* to Human 2 | Proposition | Sound | Performance
> **Implicature:** [[Human | Institution]] concentrates on hearing or paying attention (to [[Human 2]] | [[Proposition]] | [[Sound]] | [[Performance]])

Thus, a normal use of the verb would be found in a sentence such as *Dolores* [[Human]] *is listening to the New Year's Concert* [[Performance]]. But this pattern may be exploited for pragmatic purposes —consider the imperative sentence *Listen to the chocolate* on the website of a chocolate boutique in Australia.[2] The noun *chocolate* does not fall into any of the lexical type categories described above. The object semantic type has been replaced by the type [[Food]], thus causing pattern exploitation. Even the authors of this creative sentence realize its meaning may not be clear for the reader and specify: "No, this isn't listening to the chocolate saying 'Eat me, Eat me!'. Break your chocolate and listen for a clear 'snap', indicating the chocolate is in a good-tempered state, and cocoa butter is stable". The meaning becomes clearer now. And this is precisely what exploitations are about: they are instances of creative language, but not so creative as to hinder effective

2. <https://josophans.com.au/pages/faqs> [30.05.2020].

communication. Curiously enough, such is also the aim of screenwriters, who teach screenwriting trainees to "make […] words seem lifelike, but not so much that they impede the flow of meaning" (Wolff & Cox 1988: 56).

3. The case study

This section offers a case study to illustrate the application of CPA to the study of audiovisual translations. The study was part of my doctoral dissertation, which pursued a holistic description of the language used in contemporary television series of the crime genre and their translation. The dissertation employed a combination of corpus-based and corpus-driven methodologies to report on more than twenty lexical and morphosyntactic features. The results of the whole study can be found in Arias-Badia (2020), and the initial methodological proposal was presented in Arias-Badia (2015). In this paper, only a synthesis of the approach to anomalous collocation (one of the studied features) is provided to clarify the way in which CPA may be used for the purposes of audiovisual translation research.

3.1. Framing the study

This case study is concerned with anomalous collocation in television dialogue and the way in which it is rendered in interlingual subtitling. To date, the specialized literature has not agreed on a single definition of the notion of collocation. Emphasis of the notion is placed on the idea that words show restrictions when it comes to combining them with one another. Criteria to establish which lexical combinations involve collocation range from educational proposals arguing that combinations prove useful for the language learner (Sinclair 1991, Lewis 1993) to purely statistical measures accounting for lemma cooccurrence in general usage corpora (Kilgarriff *et al*. 2004) (Torner and Arias-Badia, in press). Using the CPA framework, for the purposes of this paper the term will be understood as word combinations that fit the patterns of normal use. In the present study, by contrast, I focus on instances of collocation that are *not* predictable considering the normal patterns of each of the collocates. As defined by Leech (1990/1974: 17), they are "improbable combinations". According to Hanks (2013: 217), anomalous collocates are one type of language exploitation that can be defined

as "noncanonical members of a lexical set", which means that they do not follow the rules of semantic selection.

Within audiovisual translation, subtitling is defined as a translation modality in which a written text is added to the source audiovisual content, to account for the linguistic elements of the source text by keeping synchrony with the moving picture (adapted from Matamala 2019: 127). Interlingual subtitles are hybrid in nature in that they a) transfer semantic content from one language to another; b) involve medium conversion, i.e. spoken to written; and c) are constrained by temporal and spatial parameters well spread in the practice of subtitling —although parameters vary among different clients and companies, subtitles must be typically displayed for 1-6 seconds on screen, have one or two lines, and make use of up to 36-42 characters per line—, thus being well-known for producing condensed translations (Arias-Badia, 2020, p. 27). Those factors have been said to affect lexical selection, since a pragmatic-oriented translation is favoured in which "the speech act is always in focus; intentions and effects are more important than isolated lexical items" (Gottlieb 1998, p. 247). The literature converges on the idea that the use of conventional or standardised lexicon, to the detriment of creative language or language deviating from the norm, is frequently the result of such an approach in subtitling (Zaro 2001; Díaz-Cintas 2003; Bartoll 2012).

Not *all* translation solutions in subtitling, however, are based on a reduction or condensation principle, since sometimes the subtitles are longer than the creative proposals of the ST. Consider Example (1) from our corpus:

(1) DM ST: Pretty fucking please with cheese on top. [D01, 00:09:08][3]
TT: Un jodido favor de los gordos|con queso encima.
BT: A fucking favour of the fat ones|with cheese on top.

In (1a), Dexter's sister Debra produces a sentence in which she assigns the semantic type [[Surface]] to the interjection *please*, thus attributing the

3. The examples in this paper include the following information: transcript of source text utterance (ST), target text subtitle (TT) —when applicable—, and back translation (BT) —when applicable. The initials before each ST stand for the name of the character speaking in the series. Time codes are provided, as well as indication of the episode from which the example is extracted (in Example (1), *Dexter*'s first episode, "D01"). The symbol "|" is used to represent subtitle line breaks.

capacity to hold "cheese on top" to the interjection. This creative phrase is to some extent preserved in the TT, although the translator resorts to the noun *favor* ('favour') instead of using the Spanish interjection *por favor* ('please'). The subtitle lasts 2,5 seconds, which means that, ideally, its length should not exceed 30 characters (Bartoll 2012: 136). However, in this case, the subtitler seems to give priority to the creativity of the character over conventional professional guidelines and offers a 46-character-long two-liner.

Thus, the aim of this case study is twofold: on the one hand, it intends to establish the degree of occurrence of this form of lexical creativity in the source and target texts, namely the first two episodes of the first season of the television series *Castle* (2009), *Dexter* (2006) and *The Mentalist* (2008) and their subtitles published on DVD format —the corpus extension is of 36,995 words (ST) and 34,019 words (TT). On the other, it sets out to describe the types of translation solutions used to render anomalous collocation in subtitling. Unusual collocations have been paid little attention in audiovisual translation. An exception is the work of Teixeira (2015), who looks at the English-Portuguese subtitling of idiomatic collocational patterns. Díaz-Cintas & Remael (2007: 177), for their part, provide advice on collocates segmentation in subtitling. One of the few discussions on anomalous collocation in audiovisual translation can be found in Chaume (2004), when he describes the lexical combination in one of his examples as "odd".

3.2. Methodology: Adapting CPA for the study of lexical exploitation in audiovisual translation

Section 2 has described the analytic procedure of CPA for lexicographic purposes. Beyond the *Pattern Dictionary of English Verbs*, such a procedure has been used in lexicographic practice in other languages, like Spanish, and has been encouraged as a resource for language learning (Renau 2012). The present study proposes to use the CPA language description methodology for a systematic study of translations and, specifically, of audiovisual translation. However, adaptations are needed to pursue this aim.

When conducting corpus-driven research on a specific translation, the researcher does not usually have access to over 250 concordances of each lemma. This hinders a strict application of CPA, which, as stated above,

involves annotation of a relatively large random sample of occurrences of the lemma under analysis to account for its patterns of use.

Therefore, the proposed adaption of CPA involves the following two-step process: a) manually identifying potential creative uses of language in the ST and the TT; and b) checking against general usage corpora and lexicographic tools in order to establish which instances are exploitations. In the present study, the annotation process has been repeated at three different stages of the research in order to foster intra-annotator agreement and minimise human error.

The first step closely follows the first concordance scanning proposed by CPA (see above). In this adaptation, first the transcript and subtitles are read, and potential query terms are identified.

As regards the second step, the corpora and dictionaries employed for the study of the ST in American English have been the *Pattern Dictionary of English Verbs* (PDEV), the *Corpus of Contemporary American English* (COCA), the *Macmillan English Dictionary* (MED), the *Oxford English Dictionary* (OED) and the *Merriam Webster Online Dictionary* (MWD).

For the analysis of the TT in Castilian Spanish, the *Diccionario de la lengua española* (DLE), *Redes: Diccionario combinatorio del español contemporáneo*, *Diccionario combinatorio práctico del español contemporáneo*, Davies' *Corpus del Español* (CDE), and the *Corpus del español del siglo XXI* (CORPES XXI) and have been used to check the exploitation candidates after initial TT scanning. These sources have been complemented with browser queries when needed, although these queries have been kept to a minimum.

To illustrate the annotation process and how the line between conventionalised or true creativity is drawn following this methodology, consider examples (2-4) taken from the ST corpus. The content of these dialogue lines could be understood to fall into one shared conceptual metaphor: LIFE IS A DRAMATIC PERFORMANCE.

(2) DM ST: My sister **puts up a front** so the world won't see how vulnerable she is. [D02, 01:34:41]

(3) PJ ST: It doesn't **play** nearly as well. [M01, 00:24:02]

(4) DM ST: Miami is a great place for me to **play**. [D01, 00:17:43]

In (2), to *put up/on a front* is a lexicalised expression in English —the MED includes it under the following definition of *front*: "behaviour that is not sincere because you want to hide your real feelings". Example (3) may be paraphrased as 'It is not as effective, right?' —which, actually, is the literal subtitling that has been given to this sentence from *The Mentalist* in Spanish. It is difficult to find a dictionary definition of *play* that exactly matches this sense of effectiveness —however, the conceptual metaphor makes it really easy for a native or proficient user of the language to understand the meaning of the sentence, and it can be found in larger corpora, so it would be difficult to classify it as an exploitation: this is a conventionalised metaphor. Finally, in (4), Dexter means that Miami is a great place for him to 'kill people without being noticed or chased by the police.' This sense of *play* is fully understandable in context, like the previous one, but constitutes a clear deviation from the standard sense of *play* —apart from using a linguistic metaphor, Dexter is being sarcastic. Therefore, cases like (4) have been tagged as exploitations in the corpus.

3.3. Summary of main findings

As stated above, a thorough discussion of the case study can be found in Arias-Badia (2020). This section provides a synthesis of the main findings of the study to show the kind of results that may be obtained after application of the proposed methodology.

Anomalous collocations are the most frequent type of lexical exploitation in the studied series, closely followed by creative metaphor. They are traceable in the six studied episodes, unlike other types of exploitation which seem to be favoured by specific writers, but are not a constant to the genre of crime. A total of 43 instances have been annotated as anomalous collocations in the ST.

In TV series, anomalous collocations adopt the two major roles of this kind of exploitation as described by Hanks (2013: 217). On the one hand, they are used for referring to "an abnormal situation" when speaking in a normal way. This is the case of Example (5), in which Dexter, a blood spatter analyst who needs specific phrases to describe the blood stains on a wall

("abnormal situation"), comes up with expressions like *nice, clean sprays of blood*.

(5) DM ST: nice, clean sprays of blood [D01, 00:18:34]
TT: una lluvia de sangre|limpia y ligera
BT: a rain of blood|clean and light

On the other hand, anomalous collocates are used in the studied series as a pure "rhetorical device" (Hanks 2013: 217). This use is illustrated in Example (6), where Castle, a best-selling author characterized by his wit and expressivity, is entitled to use supposedly improvised "rhetorical devices" such as *My lifeless remains cannot sue the city*, where the prototypical semantic type functioning as a subject of the verb *sue* —i.e. [[Human]] or [[Institution]] according to the PDEV— is replaced by the semantic type [[Body Part = Lifeless]].

(6) RC ST: My lifeless remains cannot sue the city. [C02, 00:00:34]
TT: Mis restos no pueden|demandar a la ciudad.
BT: My remains cannot | sue the city.

In crime TV shows, the common denominator to these uses, and one which applies to the 43 anomalous collocates found in the corpus, is that they serve as markers of fictional orality, i.e. a device whereby screenwriters aim to artificially provide their texts with greater authenticity and vivacity (Brumme and Espunya 2012). This is expected if we assume that lexical exploitation is part of naturally occurring conversation and, as stated above, screenwriters strive for making television dialogue lifelike. What we find is a more or less elaborate fictional orality in accordance with each character's profile.

The corpus study following the adapted CPA methodology has revealed the occurrence of domain-specific anomalous collocations, that is, non-canonical collocations directly connected to the prototypical domains of specialisation in the genre of crime (law, police procedure, forensic medicine). An example of these anomalous collocates is *close-case doughnuts* [M01, 00:49:31], translated as *donuts de caso resuelto* ('doughnuts of solved case'). Domain-specific collocations amount to more than a half of the anomalous collocations annotated. Some of them are likely to become norms in the

future, if they become a tendency in the genre. *Gruesome souvenir* [D02, 00:05:51], which occurs once in *Dexter*, stands out in this sense —while there is only one occurrence of this collocation in general usage corpora such as the COCA, it is traceable in some crime novels and war memoir books. The collocation is also present in the press, albeit only rarely —consider headings such as "Cat's head may be 'gruesome souvenir'" from *The Border Mail* (October 9, 2013), referring to the case of a woman who found her cat beheaded, or "A very gruesome 'souvenir'" from *News&Record Greenboro* (July 16, 1990), reporting on the murder of a 15-year-old. In *Dexter*, the collocation has been translated as *un truculento regalo* ('gruesome present'), which is indeed anomalous in Spanish.

This example shows how intertextuality plays a crucial role in the interplay of conventionalised and totally novel uses of language. If a particular exploitation can be found repeatedly in different products belonging to the same genre, for example, that paves the way for it to become a norm (i.e. lexicalised expression) in the future. Likewise, the authors' quoting may be unconscious: they may regard their production as novel, when it actually exists already, or may even have become conventional for a reduced group of people. In the same way, audiences may not share the author's world knowledge and not engage in intertextuality. An interesting example from the corpus in this sense is the expression *There she blows* (M01, 00:23:14), which reproduces a quote from *Moby Dick* referring to a whale, but is used in the series to refer to the face painted by a killer with the victim's blood. The use of the pronoun *she* to refer to a painting strikes as surprising, and it has been labelled as an exploitation in the context of the corpus, but intertextuality has been vital in word selection in this case. In Spanish, the sentence has been translated as *Ahí está* ('There it is'), thus not conveying the indirect literary reference nor making use of an exploitation in Spanish.

Following the ideas posited by the specialized literature regarding subtitling, translation solutions in Spanish have favored standardized renderings of the anomalous collocates found in English in the ST (24 vs. 19 instances). Importantly, no instances of creativity have been traced in the subtitles in segments that were not annotated as creative in the ST. Let us present an example of each major type of solution.

In Example (7), Castle exploits the first pattern of the verb *drink* described in the PDEV: [[Human]] *drink* [[Beverage]] by using an object with the semantic type [[Food]] instead of [[Beverage]], to mean that 'he drank a lot of alcohol' during a specific week. The translation preserves the idea of 'drinking a lot of alcohol' by means of a verbal periphrasis expressing continuity over time (*pasar(se) bebiendo*, 'spend drinking'), which is a standard structure in Spanish.

(7) RC ST: I drank every meal for a week. [C02, 01:14:07]
TT: Me pasé toda la semana bebiendo.
BT: I spent all the week drinking.

In Example (8), the negatively-loaded noun *monster*, which Dexter calls himself because he is a murderer, collocates with the positive adjective *neat*. This is understandable in the context of the series plot because Dexter presents himself as cautious about the remains of his victims. The TT opts for a word-for-word translation of each unit and offers *monstruo pulcro* ('monster neat'). Neither of the options in English or Spanish has matching records in corpora. Instead, in the COCA, *monster* collocates with epithets dealing with physical appearance, such as *green-eyed*, *big* or *two-headed*, as well as with negative adjectives such as *scary* or *evil*. *Neat*, by contrast, collocates with substantives related to the organization or distribution of physical objects, such as *piles*, *lines*, or *rows*. The same kind of collocates are found in Spanish reference corpora.

(8) DM ST: I am a very neat monster. [D01, 00:12:08]
TT: Soy un monstruo muy pulcro.
BT: I am a monster very neat.

The present case study has demonstrated the presence of anomalous collocates in television dialogue. It has also shown how American crime TV series make use of salient, domain-specific anomalous collocates which bear intertextuality with literary products of the crime genre, or the report of crime cases in the press. Although the total occurrence of anomalous collocates is relatively low (43 instances in the whole corpus), the fact that these elements convey information about the characters' personality which should

not be overlooked poses a challenge for audiovisual translators. Therefore, professional subtitlers and trainees could benefit from studies raising awareness of such lexical exploitations present in audiovisual material. TNE and CPA have proved their usefulness as both a theorical and methodological framework for such studies.

A tendency towards normal use has been observed in the subtitling of these creative units. The comparison of these findings with the existing literature on the translation of other instances of creative language, such as metaphor, may lead to the conclusion that using normal, non-exploited language is a preferred solution by professionals when dealing with lexical exploitation. Future studies on the treatment of other kinds of exploitations could result in the identification of a translational norm in translation practice (Toury 1995).

4. Methodological review

This section is devoted to reflecting on the shortcomings of the proposed methodology, as well as on means to tackle them —when possible— and on the advantages entailed in adopting this approach for the study of audiovisual translation. Each disadvantage or advantage is expressed in a sentence and further specified below.

4.1. Limitations

a) CPA and its adaptation entail introspection on the part of the researcher.

Studies dealing with style tend to rely on a first manual, intuitive approach to the text in hand. Leech and Short (1981: 4) pose the question in the following terms:

> [S]tylistics, as the study of the relation between linguistic form and literary function, cannot be reduced to mechanical objectivity. In both the literary and the linguistic spheres much rests on the intuition and personal judgement of the reader, for which a system, however good, is an aid rather than a substitute. There will always remain, as Dylan Thomas says, "the mystery of having been moved by words".

Indeed, it is frequent for translation scholars to undertake corpus studies individually, and the manual annotation proposed for this study does not escape introspection and an inherent bias. As specified by Hanks (2012: 54), however, "introspection of this kind can be confirmed and extended by examination of patterns in corpora". Therefore, corpora and dictionaries are proposed as the "aid", following the term used in the citation above, to minimise the impact of individual bias. Inter-annotator agreement tests undertaken with other researchers in the field are also encouraged when possible. If not possible, as in the case study described, intra-annotator agreement tests are deemed necessary, again, to obtain robust results.

b) The availability of corpora and lexicographic resources varies greatly across different languages.

This study focuses on the English-Spanish language pair, for which extensive resources are available to be used as elements of contrast. Indeed, it would be difficult to conduct this type of research on underrepresented language pairs, since the contrast of personal intuitions with lexicographic sources of information is mandatory to obtain relevant results. Although they are not desirable because they typically produce noise, browser queries could be useful to tackle this limitation.

c) Multimodality must not be neglected in the analysis of audiovisual translations.

When annotating general usage corpora, CPA lexicographers have access to text segments which allow them to assign a specific pattern to each occurrence of the word under analysis. In the case of audiovisual translation, both television dialogue and subtitling take place within the broader audiovisual text. To escape a lexical analysis deprived of contextual information, thus, the audiovisual translation researcher must take the image into consideration in the annotation process. Indeed, "corpus investigations focusing exclusively on the verbal component are at risk of overlooking the importance of the other semiotic codes to the meaning making process in audiovisual products" (Díaz-Cintas 2008: 3). To date, however, audiovisual translation research has approached multimodality in a rather qualitative manner,

indeed constrained by the individual bias of researchers. This is surely the case in the study reported. In this sense, it is hoped that recent research proposing systematic ways to address multimodality in other audiovisual translation modalities, such as Reviers' (2018) account of audio description by means of the integration of a multimodal concordancing system, will enrich the methodology hereby proposed in the near future.

d) Manual annotation is a difficult task: there is a vast "grey area" to consider in the study of the lexicon.

In the corpus annotation process, it is often difficult to draw the line between conventionalised, seemingly creative expressions, and truly creative exploitations. This is because there is no such dividing line, only a "fuzzy grey area" (Hanks 2010, 2013), as noted by lexicographers who have previously faced the CPA annotating task (Renau 2012: 141). In this sense, the above mentioned "aids" become vital for the researcher to provide solid arguments for annotation. However, the difficulty entailed in the manual annotation methodology *per se* must be considered a clear shortcoming of the proposal. The accuracy rates of automatic semantic taggers reported to date, however, are low (Rees 2018: 211) for this kind of task. So, for now there is no potential alternative to manual annotation.

e) Results of this kind of study become obsolete over time: the lexicon changes continuously.

The fact that the lexicon is in constant change is undeniable. Therefore, researchers adopting the approach proposed here must bear in mind that the results obtained will only describe what was (or was not) creative at the time of the research. Corpora and dictionaries reporting on the language used in different time periods may be employed to address the language of a specific television product. In fact, such lexicographic resources are also known for needing constant updating, so the arguments provided for analysis at some point in time may vary if we repeat the annotation process with the same texts after updates in the lexicographic resources have been implemented or general usage corpora have been extended.

It is encouraged that studies using CPA for the study of translation reflect on likely future changes in the results gathered, as done in the case study reported with the phrase *gruesome souvenir*, which can now be labelled as unusual or anomalous, but may become a norm of the genre in the future, since it is already traceable in a number of crime fiction-related products.

4.2. Contributions

a) Adapting a methodology that considers "normal" use of language is a step towards cohesive interdisciplinary research.

As posited above, the identification of "normal" patterns is a focus of interest common to Television Studies, Linguistics, and Translation Studies. Accordingly, visual marks or structural invariants (i.e., invariable patterns) are key to the identification of audiovisual genres and types of product from the standpoint of Television Studies; these patterns provide the "frame of reference" in order for audiences to understand the shows they are exposed to, and for scholars to study them (Richardson 2010: 84). Great efforts are devoted in linguistic research to studying normal patterns of language use. TNE is a clear example of the interest raised by norms in the discipline. In a similar vein, within Descriptive Translation Studies, Toury (1995) postulates the Theory of Norms in Translation, a research framework the object of which is to trace recurrent patterns in translational behaviour. Thus, by establishing a link in the core normative concept in these disciplines, the proposed methodology strives to take the first step towards cohesive research.

b) Adopting a methodology that has proved useful in another field is convenient for Translation Studies.

Just like Television Studies rely on Literary Theory to account for genre norms, i.e. the younger discipline resorts to the older one, it seems natural to suggest that replicating a solidly grounded theory within lexical analysis for the study of translations is a sound practice for the discipline. CPA is fully in accordance with Descriptive Translation Studies principles and provides a systematic approach to the study of translation creativity based on data beyond the researcher's intuition.

c) CPA allows the researcher to escape the hegemony of the source text in their studies.

The independent annotation (and analysis) of the whole source and target texts allows researchers to decide where to place emphasis when conducting their study. Although this has not been the case in the study reported above, the results of this process may reveal creativity in parts of the target text that are not creative in the source text, thus challenging the assumption that translations are usually less creative. CPA allows a true bottom-up, corpus-driven approach to translations as creative texts, deserving as much attention as source texts, and deals with them from a *tabula rasa* perspective.

d) CPA allows a statistical analysis of the results.

As noted by Corpas (2008: 53), corpus-driven studies allow statistical analyses. Hanks (2013: 415) explains that "being able to make predictions is much more useful than speculating about the boundaries of possibility." Only studies undertaken with proved systematic, replicable annotation may lead to the eventual identification of rules, norms and tendencies in Translation Studies. According to Toury (1995), such is the aim and scope of researchers in Descriptive Translation Studies. As put by Laviosa (2002: 79):

> This type of analysis is performed not to evaluate the quality of a given translation, but to understand the decision-making process underlying the product of translation and to infer from it the translational norms adopted by the translator.

Of course, it is difficult to draw the line between tendencies and norms, which typically depend on corpora representativeness (Tognini-Bonelli 2001; Martínez Sierra 2011; Corpas & Seghiri 2016). This aspect, i.e. the degree to which the research intends to extrapolate the results obtained, must always be considered in the light of the specific research questions raised in a study adopting the methodology proposed here.

5. Closing remarks

This paper provides a critical evaluation of a methodological proposal made in my own previous work. It has reported on the methodology used for a

case study aiming to describe the occurrence of anomalous collocations as a creative device in television dialogue and subtitling. Such a methodology entails the adaptation of Corpus Pattern Analysis, which had been hitherto used to study lexicon in corpora for lexicographic and language learning purposes.

The main goal of the paper has been to provide an overview of the limitations and contributions of the proposed methodology for the study of audiovisual translation. Toury (1995: 69) aptly noted that

> […] achievements of actual studies can themselves supply us with clues as to necessary and possible methodological improvements. Besides, if we hold up research until the most systematic methods have been found, we might never get any research done.

Indeed, the proposed adaptation of CPA is lacking in some respects, among which I would highlight the need to further account for multimodality in the case of audiovisual translation. It is, however, a sound methodology for the study of lexicon in translation in the sense that it is based on a well-grounded methodology employed in neighbouring research fields. It has already brought interesting results for language pairs which allow a systematic checking of creative uses of language in extensive corpora and dictionaries.

References

Arias-Badia, Blanca. (2015) "Towards a Methodology for the Analysis of Neutralisation in Spanish Subtitling." In: Corpas, Gloria; Seghiri, Miriam; Gutiérrez, Rut; Urbano, Miriam (eds). *New Horizons in Translation and Interpreting Studies. Proceedings of the 7th AIETI Conference*. Geneva: Tradulex, pp. 513-526.

Arias-Badia, Blanca. (2020) *Subtitling Television Series: A Corpus-Driven Study of Police Procedurals*. Oxford: Peter Lang.

Bartoll, Eduard. (2012) *La subtitulació: aspectes teòrics i pràctics*. Vic: Eumo.

Brumme, Jenny & Anna Espunya (eds.) (2012) *The Translation of Fictive Dialogue*. Amsterdam/New York: Rodopi.

Chaume, Frederic. (2004) "Discourse markers in audiovisual translating". *Meta* 49:4, pp. 843-855.

Corpas, Gloria. (2008) *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt am Main: Peter Lang.

CORPAS, Gloria & S Miriam Seghiri. (2016) *Corpus-based Approaches to Translation and Interpreting*. Frankfurt am Main: Peter Lang.

DÍAZ-CINTAS, Jorge. (2003) *Teoría y práctica de la subtitulación inglés-español*. Barcelona: Ariel.

DÍAZ-CINTAS, Jorge. (2008) "Introduction." In: Díaz-Cintas, Jorge (ed.) *The Didactics of Audiovisual Translation*. Amsterdam/Philadelphia: John Benjamins, pp. 1-18.

DÍAZ-CINTAS, Jorge & Aline Remael. (2007) *Audiovisual translation: Subtitling*. Manchester: St. Jerome Pub.

FEUER, Jane. (1992) "Genre study and television." In: Allen, Robert Clyde (ed.) *Channels of discourse, reassembled*. Chapel Hill: The University of North Carolina Press, pp. 138-160.

FISKE, John. (1987) *Television Culture*. London: Routledge.

GOTTLIEB, Henrik. (1998) "Subtitling." In: Baker, Mona & Kirsten Malmkjær (eds.) *Routledge encyclopedia of translation studies*. London/New York: Routledge, pp. 244-249.

HANKS, Patrick. (n.d.) "Corpus Pattern Analysis: How people use words to make meanings". Electronic version: <http://www.patrickhanks.com/powerpoint.html>.

HANKS, Patrick. (2004) "Corpus Pattern Analysis." In: Williams, Geoffrey & Sandra Vessier (eds.) *Proceedings of the 11th Euralex International Congress*, Lorient: Euralex, pp. 87-97.

HANKS, Patrick. (2012) "How people use words to make meanings: Semantic types meet valencies." In: Thomas, James & Alex Boulton (eds.) *Input, Process and Product: Developments in Teaching and Language Corpora*. Masaryk: Masaryk University Press, pp. 52-67.

HANKS, Patrick. (2013) *Lexical Analysis: Norms and Exploitations*. London: The MIT Press.

KILGARRIFF, Adam; Pavel Rychly; Pavel Smrz & David Tugwell. (2004). "The Sketch Engine." In: Williams, Geoffrey & Sandra Vessier (eds.) *Proceedings of the 11th Euralex International Congress*, Lorient: Euralex, pp. 105-116.

LAVIOSA, Sara. (2002) *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam: Rodopi.

LEECH, Geoffrey. (1990/1974). *Semantics: The Study of Meaning*. London: Penguin.

LEECH, Geoffrey & Mick Short. (1981) *Style in fiction: A linguistic introduction to English fictional prose*. London: Longman.

Lewis, Michael. (1993) *The Lexical Approach*. Hove: Language Teaching Publications.

Matamala, Anna. (2019) *Accessibilitat i traducció audiovisual*. Vic: Eumo.

Martínez Sierra, Juan José. (2011) "De normas, tendencias y otras regularidades en traducción audiovisual." *Estudios de Traducción* 1, pp. 151-170.

Neale, Steve & Graeme Turner. (2001) "Introduction: What is genre?" In: Creeber, Glen (ed.) *The television genre book*. London: British Film Institute, pp. 1-7.

Olohan, Maeve. (2004) *Introducing corpora in translation studies*. New York: Routledge.

Rees, Geraint Paul. (2018) *A Phraseological Multi-Discipline Approach to Vocabulary Selection for English for Academic Purposes*. Universitat Pompeu Fabra (Barcelona). Doctoral dissertation.

Renau, Irene. (2012) *Gramática y diccionario: Las construcciones con* se *en las entradas verbales del diccionario de español como lengua extranjera*. Universitat Pompeu Fabra (Barcelona). Doctoral dissertation.

Reviers, Nina. (2018) *Audio Description in Dutch. A corpus-based study into the linguistic features of a new, multimodal text type*. University of Antwerp (Antwerp). Doctoral dissertation.

Richardson, Kay. (2010) *Television Dramatic Dialogue: A Sociolinguistic Study*. Oxford: Oxford University Press.

Sinclair, John. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. (2003) *Reading Concordances*. London: Longman.

Teixeira, Isadora. (2015) *Subtitling of collocational patterns in children's animated movies: A corpus-based study*. Universidade Federal de Santa Catarina (Florianópolis). MA thesis.

Tognini-Bonelli, Elena. (2001) *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.

Torner, Sergi & Blanca Arias-Badia. (in press) "Los diccionarios de colocaciones." In: Battaner, Paz; Renau, Irene; Torner, Sergi (eds.) *Routledge handbook of Spanish lexicography*. London: Routledge.

Toury, Gideon. (1995) *Descriptive Translation Studies and Beyond*. Amsterdam/Philadelphia: John Benjamins, pp. 53-69.

WALSH, Steve, MORTON, Tom & Anne O'Keeffe. (2011) "Analysing university spoken interaction. A CL/CA approach." *International Journal of Corpus Linguistics* 16:3, pp. 325-344.

WOLF, Mauro. (1984) "Géneros y televisión." *Anàlisi. Quaderns de Comunicació i Cultura* 9, pp. 189-198.

WOLFF, Jurgen & Kerry Cox. (1988) *Successful scriptwriting.* Cincinnati: Writer's Digest Books.

ZARO, Juan Jesús. (2001) "Conceptos traductológicos para el análisis del doblaje y la subtitulación." In: Duro, Miguel (ed.) *La traducción para el doblaje y la subtitulación*. Madrid: Cátedra, pp. 47-64.

NOTA BIOGRÁFICA / BIONOTE

BLANCA ARIAS-BADIA is a tenure-track lecturer at the Universitat Pompeu Fabra in Barcelona, where she teaches translation for general purposes and audiovisual translation. She holds a PhD in Translation and Language Sciences (UPF) and did a postdoc in Audiovisual Translation and Media Accessibility at the Universitat Autònoma de Barcelona, awarded by the Spanish Research Agency under the Juan de la Cierva postdoctoral scheme. She is a member of the research group InfoLex (UPF) and the research coordinator of the Catalan Association for the Promotion of Accessibility (ACPA). She is the author of *Subtitling Television Series*: *A Corpus-Driven Study of Police Procedurals* (Peter Lang, 2020).

BLANCA ARIAS-BADIA es profesora en vías de permanencia en la Universitat Pompeu Fabra de Barcelona, donde imparte asignaturas sobre traducción general y traducción audiovisual. Es doctora en Traducción y Ciencias del Lenguaje (UPF) y realizó un posdoc Juan de la Cierva centrado en la traducción audiovisual y la accesibilidad en la Universitat Autònoma de Barcelona, financiado por la Agencia Estatal de Investigación. Es miembro del grupo InfoLex (UPF) y coordinadora de investigación de la Asociación Catalana para la Promoción de la Accesibilidad (ACPA). Es autora del libro *Subtitling Television Series: A Corpus-Driven Study of Police Procedurals* (Peter Lang, 2020).

Recibido / Received: 11/06/2020
Aceptado / Accepted: 26/10/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.04

Para citar este artículo / To cite this article:
Brett, David Finbar; Barbara Loranc-Paszylk & Antonio Pinna. (2021) "A corpus-driven analysis of adjective/noun collocations in travel journalism in English, Italian and Polish." In: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection*. *MonTI* 13, pp. 114-147.

A CORPUS-DRIVEN ANALYSIS OF ADJECTIVE/ NOUN COLLOCATIONS IN TRAVEL JOURNALISM IN ENGLISH, ITALIAN AND POLISH[1]

David Finbar Brett
dbrett@uniss.it
Università degli studi di Sassari, Italy

Barbara Loranc-Paszylk
bloranc@ath.bielsko.pl
University of Bielsko-Biala, Poland

Antonio Pinna
dedalo@uniss.it
Università degli studi di Sassari, Italy

Abstract

This paper describes the compilation and subsequent analysis of a comparable corpus of travel journalism in three languages (English, Italian, and Polish). By means of a corpus-driven methodology, our study focuses on adjective/noun pairings, extracting a list of statistically significant collocations for each language and observing differences and similarities with those of the other two. Social Networks Analysis tools are used to highlight the most productive collocates. Finally, collocations concerning selected themes are analysed across the three corpora, highlighting how this approach may provide valuable input to the production of reference materials for translators.

1. The research for this article was partly financed by the University of Sassari's "Fondo di Ateneo per la ricerca 2019".

Keywords: Collocation; Travel Journalism; English; Italian; Polish.

Abstract in italiano

L'articolo descrive la compilazione e la successiva analisi di un corpus comparabile plurilingue (inglese, italiano e polacco) di testi giornalistici di viaggio. Nel focalizzarsi sulle collocazioni aggettivo/sostantivo, lo studio affianca la linguistica dei corpora all'analisi delle reti sociali. Mediante la prima, l'esame di liste di collocazioni statisticamente significative per ciascuna lingua ha evidenziato differenze e somiglianze tra i tre corpora, mentre la seconda ha individuato i collocati più produttivi. Infine, l'analisi delle collocazioni relative ad alcuni temi specifici mostra come questo approccio offra un input utile per la produzione di materiale di riferimento per i traduttori.

Parole chiave: Collocazioni; Giornalismo di viaggio; Inglese; Italiano; Polacco.

1. Introduction

Tourism has been widely recognized as a global economic force that contributes significantly to the shaping of contemporary society. In 2018 the sector reached the 1.4 billion mark in terms of international tourist arrivals, while its export earnings grew to 1.7 trillion US dollars (World Tourism Organization, 2019). The travel journalism sector has undergone a similar growth, and its role as a key player in shaping destination images and convincingly conveying them to mass audiences has recently garnered considerable attention on the part of media scholars (e.g. Hanusch 2010; Hanusch & Fürsich 2014a; Pirolli 2019). However, attention to travel reportage paid by linguists so far has been scarce and limited in scope to one or two languages (e.g. Brett 2018; Brett & Pinna 2015; Canals & Liverani 2010; Pinna 2018).

Hence much work remains to be done as regards the language of travel journalism in general, and as Taylor and Marchi (2018: 9) note when discussing under-researched content, "we might also consider languages which are under-researched, often due to lack of resources at the level of both corpora and expertise. Similarly, we might think about the relative lack of studies on multilingual corpora." Our contribution to linguistic research on this genre consists of the compilation and subsequent analysis of a multilingual corpus comprising three languages (English, Italian and Polish) belonging to different sociocultural contexts.

By means of a corpus-driven methodology, our study investigates adjective/noun collocations, a phenomenon that has not yet been investigated in travel reportages and one to which little attention has been paid in the language of tourism.[2] Given their functions of describing and evaluating specific referents, adjectives play a prominent role in constructing destination images and thus provide a vital contribution to the purpose of the genre (Durán-Muñoz 2019: 354). Our decision to study adjective/noun collocations constitutes an attempt to identify recurrent associations of specific descriptions/evaluation and referring expressions by focussing on the following points:

1. the differences and similarities in the frequencies of adjective/noun collocations
2. the differences and similarities in what the most frequent adjective/noun collocations denote
3. connectivity, i.e. the most productive collocates in the three languages, whether these are adjectives or nouns, and whether they are general terms, or closely connected to the subject at hand
4. syntactic variability: in the case of Italian and Polish, adjectives may be placed before and after nouns. This raises the question of whether there are more collocations in one order than in another, and whether there are any collocations that can be found in both orders.

The analysis will then proceed to focus on selected themes that emerge in the results, comparing the related collocations across the three corpora. In this way our study aims to make a contribution to translators and practitioners in travel journalism by highlighting useful information regarding certain collocations in their specific contexts of use and notable cultural differences between the three corpora. The importance of real-world examples in assisting the translator's decision-making process has oft been noted, for instance: "Contextual information is extremely valuable because it shows how the word behaves in a specific communicative setting and also exemplifies how a collocation is used in real language" (Castro & Faber 2014: 232).

2. While collocation is a notoriously difficult term to define (Gries, 2013), the sense in which it is used in the current work is "the tendency of two words to -occur, or as the tendency of one word to attract another" (Hunston, 2002: 68).

1.1. Travel writing and travel reportage

Travel writing constitutes a supra-generic category that includes a wide variety of different (sub)genres, from travel books and tourist guidebooks to maps and itineraries, all sharing a fundamental interest in travel (Witosz 2007). Thompson (2011: 26) maintains that travel writing can only be broadly defined as a constellation of different types of texts sharing some combination of common attributes, the central feature of which is the first-person, non-fictional narrative of travel. Among these, texts belonging to the genre of travel reportage are characterized as factual accounts of travellers' experiences, typically describing and commenting on their trips, usually produced by professional journalists and published in dedicated newspaper sections and magazines, although nowadays travel accounts are increasingly written by amateurs and posted on personal blogs on the Internet.

Scholars of media communication and journalism studies have noted the increasing academic attention paid to this genre (e.g. Fürsich & Kavoori 2001; Hanusch 2010; Hanusch & Fürsich 2014a). In particular, Hanusch and Fürsich (2014b: 5) underline the effects of the expansion of global tourism on the media industry, one that has triggered growing interest in travel-related journalism worldwide and provided an expanding market for travel advertising. In this respect, travel journalism plays a role in the globalized economy by promoting a cosmopolitan identity for the affluent classes worldwide and contributing to the construction of tourist destination images. For Hanusch and Fürsich (2014b: 10), as a type of lifestyle journalism, travel reportage is differentiated from hard news by its commercial orientation, in that it "primarily addresses its audience as consumers, providing them with factual information and advice, often in entertaining ways, about goods and services they can use in their daily lives" (Hanusch 2013: 4). Information, guidance, and entertainment are therefore identified as the main objectives of the genre.

The provision of factual information points to a critical difference between the practices of professional journalism, i.e. the reporting of factual accounts, and travel writing, which allows the inclusion of fictional elements. However, this clear cut distinction is questioned by Thompson (2011: 30) who maintains that "the apparent truthfulness and factuality of a travelogue

is always to some degree a rhetorical effect; and we must remember also that any form of travel text is always a constructed, crafted artefact". Moreover, travel writing has seen various famous authors straddle the divide between travel journalism and literature, such as the British Lawrence Osborne, the Americans Bill Bryson and Paul Theroux, the Italians Tiziano Terzani and Guido Piovene, and the Poles Ryszard Kapuściński and Jacek Hugo-Bader.

Travel reportage may be more protean than the academic taxonomies and definitions would like it to be and various cultural traditions position it along a cline between literary fiction and journalism, as is the case of Polish travel reportage, for instance, a genre that emerged in the Polish literary tradition in the late 19th century (Moroz 2015). Traditionally, travel reportage texts found in Polish newspapers were authored by journalists who focused on reporting their travel experiences (Rajter 2004). The genre evolved in the 20th century from linear, retrospective narratives into polyphonic travels, characterised by a centralised position of narrative persona and an increased use of creative fiction techniques (Moroz 2015).

The proximity between literature and journalism in the Italian travel reportage tradition is summarized by Massimo Bontempelli's (1938: 82) aphorism "one can be a journalist without being a writer, but to be a writer one has to be a journalist." As a matter of fact, contributors to Italian travel reportage have included some of the best writers of the 20th century, for whom travels abroad constituted not only mere visits to foreign destinations, but also pastures new for their imagination, promises of intellectual freedom and personal renewal (De Luca & Scarpa 2012: 812). For others, especially after World War II, travel reportages on their Italian tours allowed the combination of social representation and a search for identity in a world in rapid socioeconomic transformation (Lombardinilo 2016: 76).

A recent survey of travel writing using British and American travelogues is offered by Thompson (2011), who explores how the genre has managed to report the world, reveal the narrating self, and represent the other. In the field of journalism studies, Hanusch and Fürsich (2014a) edited a collection of essays that is not limited to travel journalism in the West, but also takes into consideration India (Raman & Choudary 2014) and China (Bao 2014). Finally, Pirolli (2019) explores important aspects of the practice of travel reporting in the digital age.

1.2. Linguistic studies of tourism discourse and travel reportage

In his seminal work *The Language of Tourism*, Dann (1996) quotes and/or analyzes various linguistic examples taken from travelogues. Moreover, the inclusion of travel reportages in the family of texts involved in the language of tourism is vouchsafed for not only by their topic, but also by their commercial orientation (Hanusch & Fürsich 2014b: 10).

There has been some interest in the linguistic analysis of travel writing in Polish, with studies exploring various elements of tourist discourse in thematic areas such as: urban spaces (Duda 2015), geographical regions (Żarski 2013) and individual countries (Graf 2018). Some studies utilise corpora, such as Kudełko (2016), whose analysis of Polish travel texts and guidebooks about Spain published between 1910-2010 shows how stereotypes of Spanish culture and axiological recommendations of famous places were reflected in these texts. Other studies investigating tourism discourse in Polish guidebooks demonstrate a number of ways in which it has been infused with value-laden information (Podkidacz 2004). These include ekphrastic descriptions of buildings and artworks to attribute positive connotations to history and art (Stanisławek 2013), high frequency of positive evaluative adjectives, the superlatives, metaphorically rich noun phrases and highly formulaic predicate forms to promote persuasive communication and cultural stereotyping (Zarski 2013).

Italian academics' interest in the study of tourism discourse is especially evident in the field of foreign languages, e.g. Calvi (2000) and Gotti (2006). Translation from/into Italian of tourism texts has been studied by Nigro (2006), Margarito *et al.* (2011) and Baumann (2018). However, attention to travel reportage has been scarce and limited to its production in other languages (e.g. Canals & Liverani, 2010, Pinna 2018). In relation to the methodology used here, Brett (2018) employs Social Network Analysis to study the phenomenon of connectivity, extracting networks of collocates from a 1-million-word corpus of travel reports from *The Guardian*, while Brett and Pinna (2015) apply the Part-of-Speech-gram technique to study the inflected superlative adjectives in a 450,000-word corpus of travelogues from the BBC website and not only demonstrate that inflected superlatives

are characteristic of the language of travel writing, but also that they are typically used in a small series of highly frequent constructions with limited lexical variation.

2. Materials and methods

2.1. Corpus compilation

The analysis illustrated in this paper necessitated the compilation of three comparable corpora of travel journalism for the three languages discussed: English, Italian and Polish. An attempt was made to select articles from newspapers of a comparable standing in the three respective speech communities. The authors had already compiled a collection of articles from the 'Travel' section of the British broadsheet *The Guardian* called the Guardian Travel Corpus (GTC). This consisted of a total of 1204 articles, amounting to one million tokens. These articles appeared in the online version of the newspaper (https://www.guardian.co.uk) over a period from 2006-2011. When compiling comparable corpora in Italian and Polish, the choice fell on *La Repubblica* (https://www.repubblica.it/) and *Gazeta* (https://www.gazeta.pl/), respectively, both of which are considered to be quality publications, aimed at an educated middle-class readership. Just as *The Guardian* has a 'Travel' section, *La Repubblica* has a section entitled 'Viaggi' and *Gazeta* has one called 'Podróże'.

The GTC was compiled semi-automatically in the following way:

1. The pages of the archive of the travel section were downloaded using gnu wget, (http://www.gnu.org/software/wget/). This is a free software package for retrieving files using http and other widely-used Internet protocols. In the Window OS it can be used with ms-dos to loop through incrementing addresses (e.g. https://www.theguardian.com/uk/travel?p=1, https://www.theguardian.com/uk/travel?p=2, etc.), retrieving and saving each destination file.
2. Tailor-made perl scripts developed by the authors were used to scan the html of each page of the archive for links to articles. These links were then saved, but only if they met certain criteria, e.g. the link had to contain the word "travel", while "picture", "audio", "gallery" and "video" were filtered out. This process was enacted to make sure

that only samples of travel articles were included in the corpus, as opposed to articles from other sections of the newspaper, or pages presenting multimedia, which cannot be considered to be examples of travel journalism stricto sensu.

3. The result of step 2 is a list of URLs. This was then fed to gnu wget, which proceeded to download the html at each URL.
4. The html of each file was then analysed using another perl script compiled by the authors so as to identify the start and the end of the article proper, and hence eliminate all the 'boilerplate' (advertisements, links to other articles and all other extraneous content). Metadata about data of publication, author and keywords for each article were also collected.
5. The cleaned html was then converted to the txt format using another perl script.

The same procedure was followed for the compilation of the Italian and Polish corpora, with one main difference: while *Gazeta* (like *The Guardian*) allows the reader to browse through all the articles it has ever published by way of a centralised archive (via the URL https://podroze.gazeta.pl/podroze/0,0.html?str=1; one may progress through the archive simply by augmenting the value of str), *La Repubblica* allows access only to a maximum of twelve pages (amounting to approximately the 120 most recent articles). Hence an alternative strategy was necessary in order to gather a large enough sample to allow direct comparison with the corpora in the other languages. This strategy involved searching for archive pages with two variables: the page number and a tag. An example of this is: https://www.repubblica.it/viaggi/ricerca?tags=Irlanda&p=1

In this case the compiler proposes a tag and the script conducts repeated attempts to download the URL with incrementing values of p. At some point the URL will lead to a non-existent page and the script is aborted.

Two sets of tags were provided: the names of all the European countries and, considering the vast amount of internal tourism, those of all the Italian regions. While this strategy did allow the compilation of a corpus for Italian travel journalism of similar dimensions to that for English and Polish, it

is important to note that a certain amount of Euro-centric, and especially Italo-centric, bias has been introduced.

The procedure described above resulted in three 1M-word comparable corpora of travel journalism in English, Italian and Polish. Some variability was noted in the composition of the corpora: the English, Italian and Polish sections were composed of 1204, 725, and 1084 articles. Hence, the Italian section contained articles that were on average longer (1379 tokens), than those of the English (830 tokens) and Polish (922 tokens).

2.2. Annotation for part-of-speech and extraction of collocations

The texts were annotated for Part-of-Speech (PoS) using Tree Tagger,[3] a tool which not only attributes a PoS tag to each token in the text, but also provides its lemma. The parameters used for the tagging process were downloaded and installed separately.[4]

Thereafter, the lemmas constituted the focus of the work. When dealing with adjective/noun pairs in English, there are essentially two variants, that with the singular and plural form of the noun, hence the conversion to lemmas merged just two pairs of word forms into one (e.g. short break, short breaks > SHORT BREAK). Similarly, in order to calculate the total frequency in the corpus (necessary for the test for strength of collocation), using lemmas made no difference to the adjective count, while that of the nouns was the sum of the singular and plural forms. For the other two languages analysed, the impact of using lemmas was far higher. Italian adjectives usually have four forms, masc. sing, masc. plur., fem. sing., and fem. plur. Some have even more, e.g. BELLO > *bello*, *bel*, *belli*, *bei*, *begli*, *bella*, *belle*.

In Polish, both nouns and adjectives decline in case, number and gender. All singular nouns are either masculine, feminine or neuter. Further to that, among masculine nouns there is another differentiation between animate

3. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

4. The English tagset is described at https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf; the Italian tagset developed by Prof. Achim Stein, University of Stuttgart, is described at https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/italian-tagset.txt; and the Polish tagset trained on the Polish National Corpus, developed by Adam Przepiórkowski, is described at http://nkjp.pl/poliqarp/help/ense2.html.

and inanimate nouns. In the plural form, nouns and adjectives distinguish only between personal and non-personal gender. Despite the fact that there are seven cases, endings often overlap, for example, masculine animate adjectives have an identical form for the accusative and genitive case. Consequently, the adjective DOBRY 'good' has the following forms: in the singular: *dobry, dobrego, dobremu, dobrym, dobre, dobra, dobrej, dobrą*, and in the plural: *dobre, dobrych, dobrym, dobrymi, dobrzy*, arriving at 11 different forms in total. (Zagórska-Brooks 1975). Hence, working with lemma pairs, rather than word forms, becomes essential when searching for collocates and compiling wordlists of single items, in order to avoid inputting data to the statistical test that are essentially meaningless.[5]

The collocate pairs were extracted using perl scripts written by the authors. Initially a wordlist for the single lemmas was created. Thereafter, all the immediately adjacent lemma pairs tagged as adjective + noun were extracted. These data were then collapsed into a list composed of type and frequency.[6] The script then took each lemma pair, recorded its frequency, as well as that of each element's frequency on the single lemma wordlist. These data, along with the total number of tokens formed the input for a statistical test to identify pairs whose tendency to co-occur is above that of chance. The statistical procedure adopted was that of mutual information, the cut-off value for significance was 3 and the minimum frequency for collocations was 5.

5. It is important to note that the results reported below are combinations of lemmas, and not word forms. Therefore, all adjectives are in the masculine, regardless of the gender of the modified noun. For example, the Italian collocation *luna piena* 'full moon' is reported as LUNA PIENO.

6. The Polish list had to undergo considerable post-processing to eliminate a) proper nouns, e.g. NOWA ZELANDIA; comprising 4.91% of the original list of collocates b) determiners, including several types of pronouns, such as: possessives (e.g. NASZ KONTYNENT 'our continent'), negatives (e.g. ŻADEN PROBLEM 'no problem'), indefinite particles (e.g. NIEKTÓRY DOM 'a house'), as well as predeterminers (e.g. TAKI OBIEKT 'such a facility'), numerals, e.g. 10 ROK 'tenth year' and dates, e.g. 14 LUTY 'February 14' - deletions of determiners reduced the original wordlist by 30% c) duplicated words, e.g. BARWA BARWA, (2.88%) d) other erroneous inclusions which comprised 0.8% of the original wordlist.

As indicated above, the initial focus was on the prototypical order for combining adjectives and nouns in the three languages, i.e. ADJ+NOUN for English and Polish, and NOUN+ADJ for Italian. However, examination of the concordance lines for Italian and Polish suggested time and time again that the presence of the inverted order variant was not negligible. In fact, on repeating the procedure for the inverted order variant form (ADJ+NOUN for Italian and NOUN+ADJ for Polish) the numbers were indeed substantial (see Section 3.1), and furthermore the vast majority of collocations displayed a preference for one order or the other.

2.3. Networks of collocates

Brezina et al (2015:146) describe the importance of connectivity as a property of collocations, beside those of distance, frequency, exclusivity, directionality, dispersion and type-token distribution among collocates. Bearing in mind the oft-quoted statement of Firth's (1957: 11) "You shall know a word by the company it keeps", this study made use of Gephi (https://gephi.org/), a tool developed in the field of Social Networks Analysis (henceforth SNA). This tool, on importing appropriately formatted data concerning collocations (see Brett 2018), allows the production of detailed graphs highlighting the hub collocates (those which are particularly productive in the formation of collocations) and those which, to the contrary, collocate with only one other word. In SNA terminology, each item is called a node; connections between nodes are called edges; and the number of edges a given node has is called its degree. Just like directionality is a property of collocation (Gries 2013), edges can be directional, or undirected. For the purposes of the present analysis, directionality was not calculated for the dataset, though this may well be incorporated in future studies.

The decisions regarding the formatting of the graphs are the following:

1. Nodes are colour-coded for Part-of-Speech (green for adjectives, red for nouns)
2. Node size reflects frequency of the node word in the corpus
3. Edge size reflects the frequency of the collocation

3. Results and discussion

3.1. Types and tokens

Considerable variation was observed in the number of collocations that the three corpora yielded: English provided 1050 types, with a sum of 11512 tokens, Italian, 765 types, with a sum of 8102 tokens and Polish, 993 types, amounting to 10330 tokens. Therefore, at least as far as concerns the ADJ+NOUN (or NOUN+ADJ for Italian) structure, it would appear that English is the language in which there is greatest formulaicity in travel journalism, and Italian the language which resorts to it the least. Polish would appear to lie somewhere between the two. However, factoring in the variant syntactic pattern (ADJ+NOUN for Italian, and NOUN+ADJ for Polish) overturned these results. Details can be found in Tables 1 and 2. In light of these new data English and Italian appear to be extremely similar, and it is Polish that emerges as the most formulaic, both in terms of types and tokens.

	English	**Italian**	**Polish**
Types (ADJ+NOUN)	1050	365	993
Types (NOUN+ADJ)		765	431
Total	1050	1130	1424

Table 1. Statistically significant collocate pairs in the three corpora: types.

	English	**Italian**	**Polish**
Tokens (ADJ+NOUN)	11512	3404	10330
Tokens (NOUN+ADJ)		8102	5076
Total	11512	11506	15406

Table 2. Statistically significant collocate pairs in the three corpora: tokens.

The pattern is somewhat different if we take into consideration the number of collocations with a frequency greater than or equal to fifty. The types and their frequencies are listed in Table 3. English has 20, Italian has 12 and Polish has 22.

English	LAST YEAR (200), FIRST TIME (116), NEXT DAY (101), YOUNG PEOPLE (99), FULL BOARD (90), MORE INFORMATION (85), NEXT YEAR (79), FURTHER INFORMATION (76), NEXT MORNING (73), NATIONAL PARK (70), GOOD PLACE (70), FEW DAY (60), LOCAL PEOPLE (60), GOOD WAY (58), LAST WEEK (57), WEST COAST (55), FEW YEAR (55), CHIEF EXECUTIVE (53), LIVE MUSIC (53), OPEN AIR (52)
Italian (NOUN+ADJ)	CENTRO STORICO (397), PARCO NAZIONALE (123), PARCO NATURALE (86), RISERVA NATURALE (68), PISTA CICLABILE (61), MUSEO ARCHEOLOGICO (59), METRO QUADRATO (55), GUERRA MONDIALE (53), MACCHIA MEDITERRANEO (52)
Italian (ADJ+NOUN)	GRAN PARTE (106), ULTIMO ANNO (68), GRANDE PARTE (60)
Polish (ADJ+NOUN)	STARY MIASTO (185), DUŻY ATRAKCJA (83), PIĘKNY WIDOK (77), DUŻY MIASTO (73), ŚWIATOWY DZIEDZICTWO (71), PRAWY STRONA (70), LEWY STRONA (67), WYSOKI SZCZYT (64), JEDEN DZIEŃ (57), DOBRY MIEJSCE (53), ŚREDNI TEMPERATURA (50)
Polish (NOUN+ADJ)	PARK NARODOWY (244), WOJNA ŚWIATOWY (120), LINIA LOTNICZY (107), MATKA BOSKI (85), INFORMACJA TURYSTYCZNY (80), BILET LOTNICZY (71), STRONA INTERNETOWY (70), MUR OBRONNY (66), ATRAKCJA TURYSTYCZNY (59), OŚRODEK NARCIARSKI (54)

Table 3. Types and frequencies of the statistically significant ADJ+NOUN (or NOUN+ADJ) collocations found in the three corpora with frequency ≥ 50.

3.2. Syntactic variation

As has been noted above, two of the languages under analysis, Italian and Polish, allow variability in the position of the adjective with respect to the noun. Italian generally prefers to place the adjective after the noun (Serianni 1989), Polish before (Zagórska-Brooks 1975). This study suggests that, at least with regards to the text type taken into consideration, adjective/noun collocations in Italian and Polish occur in the canonical form twice as often

as in the variant form. This proportion remains the same regardless of whether types or tokens are taken into consideration.

A further point of interest is whether collocations display exclusive preferences for a particular order, or whether there are collocations that are statistically significant in both the canonical and the variant forms, and if so, what the proportions involved are. The data present a very clear picture: adjective/noun collocations display a distinct preference for a particular order: just 39 types (2.7% of the total number of types in each form), corresponding to 427 tokens (2.8% of total number of tokens in each form) appeared on both lists in Polish. This separation was even more extreme in the Italian data, as the statistically significant collocate pairs in both the canonical and the variant form consisted of only 11 types (1.0% of total types), corresponding to 87 tokens (0.8% of total tokens). Even when the collocations appeared on both lists, a tendency to occur in one form or the other was still observed in the majority of cases. Figure 1 presents this phenomenon for Italian. It is interesting to note that the collocations that were present in both forms displayed a preference for the variant form, i.e. ADJ+NOUN. For example, the frequency of ANTICO BORGO is 20, whereas that of BORGO ANTICO is 10. Therefore, it displays a preference for the variant form, and hence is plotted one third along the axis spanning a range from -1 to 1. The collocate pairs in the centre of the graph (close to 0) display little or no preference for one form or the other. For obvious reasons of legibility, the collocate pairs that are significant only in one form are not plotted, but if they were, they would all be aligned at -1 or 1 on the x-axis.

The topic of connectivity is discussed in detail in the next section. However, it may be fitting at this point to observe that some collocates are particularly productive in the variant syntactic form. These are all adjectives, and have to do with size (GRANDE, PICCOLO, LUNGO), age (NUOVO, VECCHIO, ANTICO) and positive evaluation (BELLO, SPLENDIDO, SPETTACOLARE).

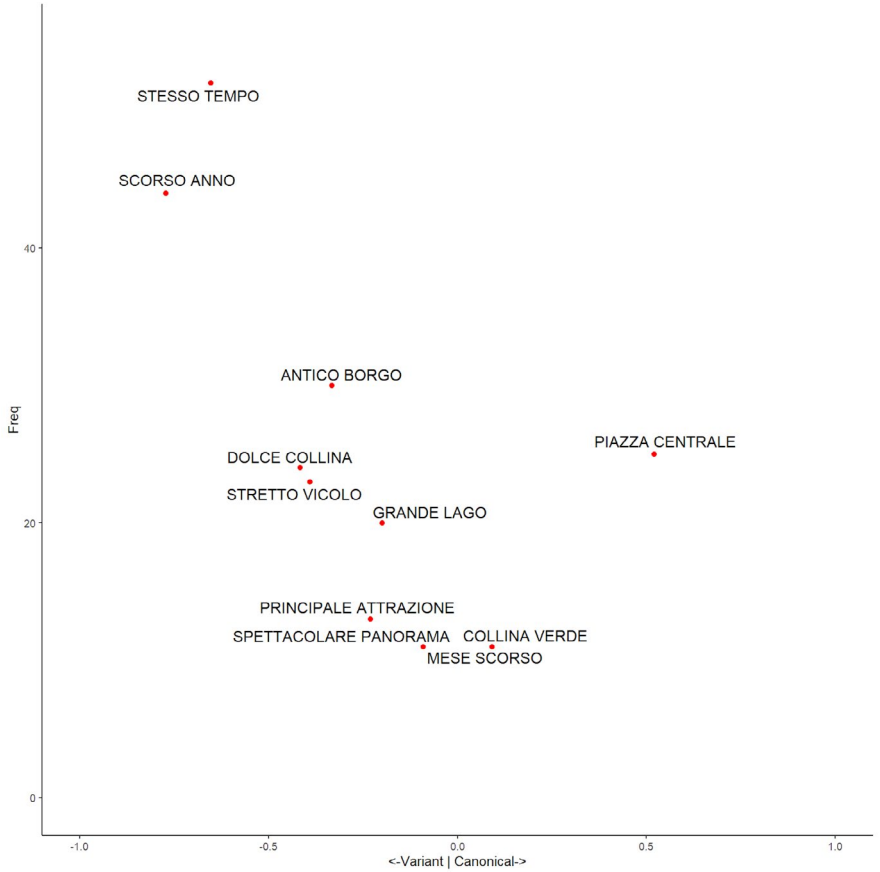

Figure 1. Adjective/noun collocations in the Italian corpus that are statistically significant in both the canonical and the variant order. The x-axis represents the tendency to occur in the canonical (right) and the variant (left) order.

With respect to the Polish data, taken as a whole the collocate pairs that were statistically significant in both orders did not appear to display a preference for a particular order (Fig. 2). On the level of the individual pair, some preferred the canonical (ADJ+NOUN) form, (e.g. CZERWONY SZLAK, WOLNY CZAS), other preferred the variant form (e.g. ATRAKCJA TURYSTYCZNY, TRASA NARCIARSKI, WODA MINERALNY, ŻYCIE NOCNY), while still

more occurred equally in both forms (e.g. TURYSTA INDYWIDUALNY, WODA MORSKI).

The most productive collocates in the variant syntactic form in Polish are, similar to Italian, all adjectives, however they are more specific to the subject matter: TURYSTYCZNY, NARCIARSKI, MIEJSKI. At this point, it is important to note that in Polish slight differences of meaning can be conveyed through noun adjective order. Therefore, if the collocate pair takes on the NOUN+ADJ order, the adjective classifies the noun based on its intrinsic quality. The example of ATRAKCJA TURYSTYCZNY can illustrate this tendency. The canonical form, TURYSTYCZNY ATRAKCJA refers to an entertainment which only potentially can be attractive to tourists, as it is primarily used for other purposes (e.g. *korzystanie z miejskiej kolejki, która sama w sobie może stanowić turystyczną atrakcję* 'taking the urban railway may in itself be a tourist attraction'). On the other hand, the variant form, ATRAKCJA TURYSTYCZNY denotes an entertainment primarily meant for tourists (*na terenie jeziora znajduje się kilka atrakcji turystycznych, dla których warto na kilka godzin podnieść się z łóżka* 'near the lake are a few tourist attractions worth getting up from bed for a few hours') and it tends to be used with a preceding adjective in the superlative form (*Ale to w końcu jedna z najbardziej znanych atrakcji turystycznych świata* 'After all, it is one of the most famous tourist attractions in the world').

With respect to collocate pairs formed with MIEJSKI and NARCIARSKI, the following tendency can be observed: when in the variant, NOUN+ADJ order, the collocations tend to be used in the plural form (*Z placu odjeżdżają autobusy miejskie;* 'City buses depart from this square'*; Wogezy nie słyną ani z narciarskich tras, ani z zapierających dech panoram,* 'The Vosges are not famous for their ski trails, nor for breathtaking views'). The canonical form, on the other hand, especially in the case of MIEJSKI AUTOBUS, is almost exclusively used in the singular form. It is also important to note that, if the collocate pair follows the NOUN+ADJ order, it tends to be pre-modified by another adjective, which does not occur often in the canonical form (*Są tu trzy dość trudne trasy narciarskie,* 'there are three quite difficult ski trails').

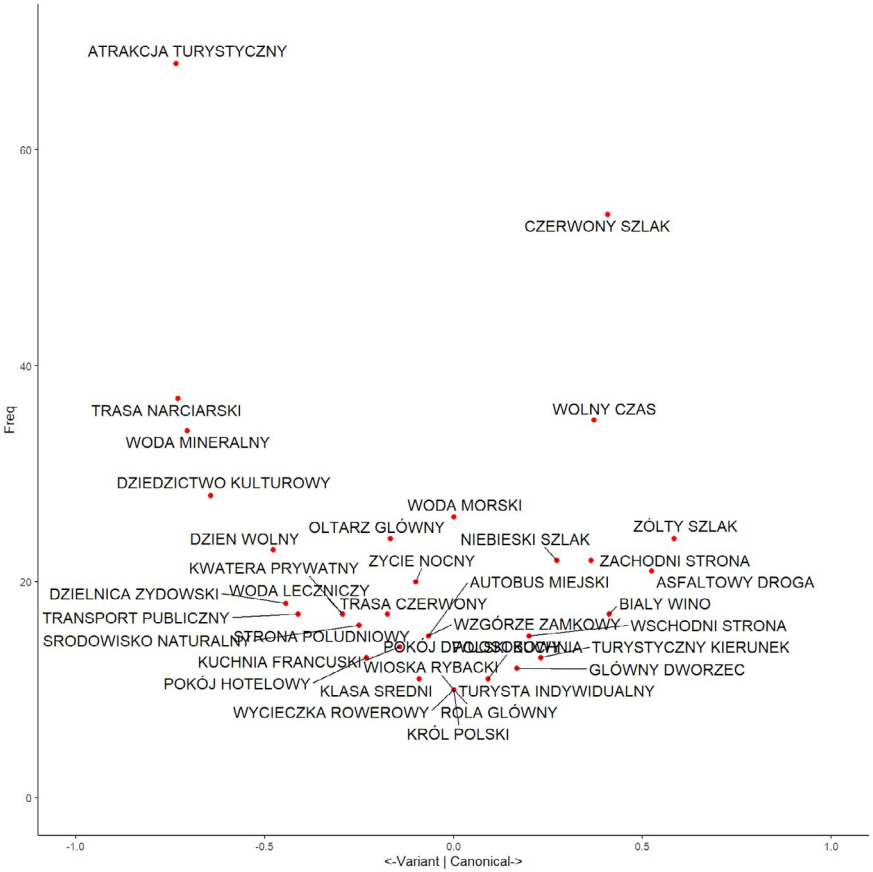

Figure 2. Adjective/noun collocations in the Polish corpus that are statistically significant in both the canonical and the variant order. The x-axis represents the tendency to occur in the canonical (right) and the variant (left) order.

In conclusion, these results may be of interest to translation practitioners and researchers as, when translating from one language to another, it is not sufficient solely to have great familiarity with the collocations in both languages (Baker 2018:53; Taylor 1998:26), but it is also necessary to be aware of the syntactic preferences that these may display in languages that allow such flexibility.

3.3. Connectivity

The lists of collocates were imported to Gephi to facilitate the identification of hub collocates, i.e., those nodes that are most productive in the creation of collocations. The degree of each node (i.e., the number of connections that it has to other nodes) is calculated by running the average degree test. This test provides an indication of the overall connectedness of the network. The average degree results for the English, Italian and Polish data were 2.565, 2.045 and 2.630, respectively. However, these results concerned only the canonical syntactic form. When the results were integrated with the data regarding the variant form, the average degree results for Italian and Polish were updated to 2.503 and 2.711, respectively.

Therefore, the overall connectedness figures are quite similar across the three corpora. Had substantial differences emerged, for example with the data for one language having a particularly low value, it would suggest that the dataset in question was composed of a greater proportion of isolates (i.e., pairs of words that collocate only with each other), as opposed to hubs.

Naturally, such isolate pairs can be found in all three datasets. They generally tend to be either technical terms, such as TIDAL BORE, INLAND WATERWAY and RENEWABLE ENERGY or low-frequency collocations such as BEATEN TRACK, HIDDEN GEM and PLAIN SAILING, to provide examples from the English dataset. A similar pattern was found in the other two languages. The isolates constituting technical terms in Italian included: SCARTAMENTO RIDOTTO 'narrow gauge', SET CINEMATOGRAFICO 'film set', and RITO PROPIZIATORIO 'propitiatory rite', whereas the low-frequency collocations included PIEDE NUDO 'bare foot', MANTO NEVOSO 'snow cover' and LUNA PIENO 'full moon'. The isolates in the variant syntactic form data were on the whole of the latter type, examples include CONTINUO EVOLUZIONE 'continuous development', TARDO POMERIGGIO 'late afternoon' and LARGO ANTICIPO 'well in advance'.

A few representative examples of technical terms found in Polish include OBFITY OPAD 'heavy precipitation', PRZEDNIA SZYBA 'windscreen', and SWOBODNY PRZEPŁYW 'free flow' and low-frequency collocations can be exemplified by the following: PRZYJEMNY CHŁÓD 'pleasant chill',

SZKLANA KOPUŁA 'glass dome', and AKTYWNY WYPOCZYNEK 'active recreation'.

Greater differences are to be seen when observing the nodes that are most connected: there are two aspects that are striking. Firstly, here the Italian corpus appears to be the outlier, with less than half the number of collocates that act as hubs, in comparison to English and Polish. A second point to be made is that the lemmas that do appear to be particularly productive in the formation of collocations are by and large specifically connected with the subject at hand in the Italian corpus, these include terms relating to culture (e.g. STORICO, MEDIEVALE, CULTURALE 'historical, medieval, cultural') and places (e.g. CENTRO, LOCALE, CAPITALE 'centre/town, local, capital'). The hub collocates found for English are mostly lemmas that could pertain to any domain. In fact, all of the noun lemmas appear within the top 100 most frequent nouns in the BNC, with the exception of BEACH. Similarly, the adjective collocates are all on the list of the top 100 most frequent adjectives in the BNC, with the exception of GAY and NEXT.[7] Interestingly, Polish appears to lie in the middle of these two extremes also with respect to the nature of its hub collocates. While quite a few are general, all-purpose words (e.g. DOBRY, DUŻY, MIEJSCE, INNY, CZĘŚĆ 'good, big, a place, the other, a part'), many more relate to the specific subject matters dealt with in travel journalism (e.g. MIASTO, WODA, DROGA, ATRAKCJA, BRZEG 'a town, water, a road, an attraction, a river bank').

3.4. Themes

The analysis will now focus on a number of themes that emerge as being recurrent in the collocations extracted from the corpora in the three languages. Differences and similarities will be highlighted, while one caveat must always be borne in mind: the current analysis is an observation solely of adjective/noun collocations. The absence of a given collocation in a particular corpus, corresponding to collocations in one or two of the other corpora, does not necessarily mean that this entity or concept is not widely dealt with or discussed in that corpus. Its absence from the list of statistically

7. BNC wordlists available at http://www.kilgarriff.co.uk/bnc-readme.html#raw

significant collocations could be due to the fact that it is expressed with a different pattern (e.g. a compound noun, or even a sole noun). Alternatively, it could be expressed by way of a number of adjective/noun expressions that are not frequent enough, or do not display a strong enough attraction, to reach statistical significance.

3.4.1. Theme 1: Human settlements

By far the collocation with the highest frequency across the three corpora is the Italian CENTRO STORICO (397). As noted above, CENTRO constitutes a hub, with 11 collocates. This is partially due to the polysemy of the word itself, and the different senses of the word can be observed in the list of collocates:

1. the core of a larger entity (STORICO, CITTADINO). The first would translate into English as 'old town' (i.e. the historical nucleus of a town/city), the second 'town/city centre'.
2. a medium/large settlement (ABITATO, URBANO, MEDIOVALE). The first two would translate simply as 'town' or 'city', the last as 'medieval town'.
3. a place of great importance (ARTISTICO, CULTURALE, SPIRITUALE). These would translate as 'artistic/cultural/spiritual centre'.
4. a place with the facilities for a specific activity (COMMERCIALE, TERMALE, BALNEARE). These would translate as 'shopping centre/mall, spa resort, seaside resort'.

STORICO is also a hub collocate, contributing to the formation of no fewer than 21 pairings. Amongst these we find QUARTIERE STORICO (5), NUCLEO STORICO (5), both of which have very similar meanings to CENTRO STORICO. Another collocate of STORICO is BORGO (6), which in turn collocates with ANTICO (10), hence providing two collocations with practically the same meaning, 'old/historic village', hinging on the noun BORGO. Of interest is the fact that the latter collocation is present also in the variant order list. In fact, ANTICO BORGO (20) is actually twice as frequent as the collocation in the canonical form.

There is another near synonym of STORICO that appears in the list of collocates: VECCHIO. This is present in a sole collocate pair, one that to all intents and purposes would again appear to have exactly the same meaning as CENTRO STORICO, CITTA' VECCHIO (8). However, an examination of the concordance lines reveals that it is used almost exclusively in non-Italian contexts.

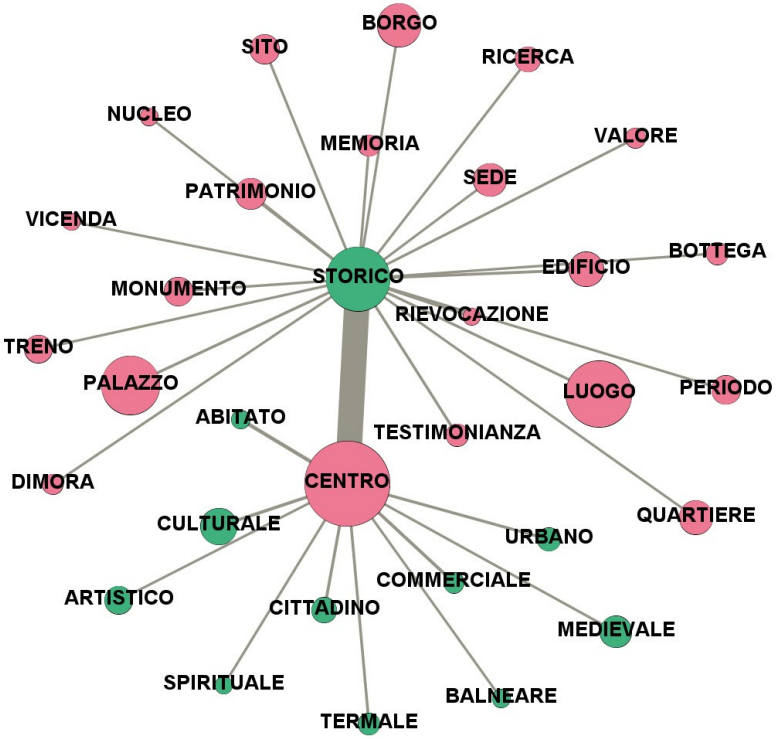

Figure 3. Network graph showing the collocates of CENTRO and STORICO

In Polish, MIASTO 'town' has 11 collocates, and that with the highest frequency is with STARY 'old'.

The following senses of the word MIASTO can be illustrated by looking at its collocates:

1. The oldest, most picturesque part of a town where most of the historic sites are located, would be represented by the collocation STARY MIASTO. This would translate into English as 'old town'. Other collocations that express the concept of STARY MIASTO are: STARY CENTRUM, HISTORYCZNY CENTRUM, and ZABYTKOWY CENTRUM - a closer look at the concordance lines shows that all three collocations are used predominantly in non-Polish contexts (*znajdują się oczywiście w centrum miejscowości, przy starym mieście Główna plaża Lloret de Mar,* 'they are obviously located in the town centre, near the Old Town, where the main beach in Lloret de Mar can be found'). It is important to note, however, that the same concept as STARY MIASTO, CENTRO STORICO and OLD TOWN can be expressed using a single noun, STARÓWKA (a derivative of STARY), of which there are 91 occurrences in the corpus.
2. A large area inhabited by a number of inhabitants where facilities are located and services are provided (DUŻE MIASTO) This would translate into English as 'city, big town';
3. A place the town or city in which a (famous) person used to live (RODZINNE MIASTO). This would translate into English as 'hometown';
4. A tourist destination, a foreign place worth visiting (EUROPEJSKI, WŁOSKI). This would translate into English simply as 'town' or 'city';
5. A place of great (religious or spiritual) importance, similar to Italian: ANTYCZNY, STAROŻYTNY, ŚWIĘTY. This sense is also expressed by the collocation WAŻNY CENTRUM. English equivalents would be 'ancient city', 'religious centre', 'important centre'.

The concept of TOWN CENTRE is expressed in Polish through the collocation ŚCISŁY CENTRUM that denotes the most central area of the town, in which all the major sites are located. An examination of concordance lines reveals that this collocation is used when referring to conveniently located places, especially accommodation.

There are two near synonyms of STARY that can be found in the list of collocations: DAWNY and ZABYTKOWY. While STARY is a more generic adjective that denotes old age and tends to form collocations with inanimate

nouns that refer to buildings (RATUSZ, 'townhall') or constructions erected as a unified community (STARY CMENTARZ', 'old cemetery'), as well as animate nouns (STARY DRZEWO, 'old tree'), its synonym ZABYTKOWY forms pairings only with inanimate nouns that refer to buildings or human settlements. DAWNY, on the other hand, collocates with both inanimate (DAWNY DZIELNICA, 'old district') and animate nouns (DAWNY MIESZKAŃCY, 'former inhabitants'), as well as nouns denoting abstract concepts (DAWNY CZAS, 'old time', DAWNY ŚWIETNOŚĆ, 'past glory') and carries the meaning of past state/activity that is no longer part of the present.

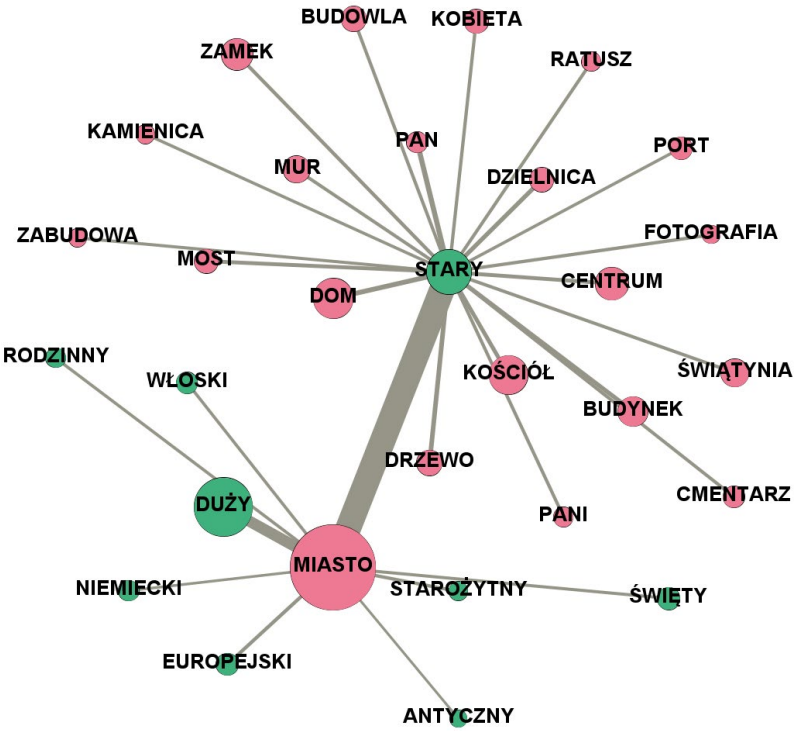

Figure 4. Network graph showing the collocates of STARY and MIASTO

The presence in the English corpus of equivalents of CENTRO STORICO/ STARY MIASTO is rather underwhelming. The only option for expressing this would appear to be OLD TOWN (27). Similarly, collocations referring to the age of human settlements are limited to MEDIEVAL TOWN (11) and HISTORIC TOWN (8).

3.4.2. Theme 2: Destination Appeal

Another theme of interest is that of evaluation, specifically the notion of attraction. Tourism, based as it is on personal and group preferences, which in turn are influenced by fads and fashions, is a particularly fickle and unpredictable market, prone to mass vagaries and whims. It is perhaps no mere coincidence that places to be visited are imbued with animation and construed as being active sentient beings that attract tourists, in the same way that humans and animals attract potential mates. This metaphor is persistent across the three corpora, as statistically significant collocations have been found featuring ATTRACTION, ATTRAZIONE and ATRAKCJA. In Italian the noun ATTRAZIONE collocates with three adjective lemmas TURISTICO (13), NATURALE (6) and PRINCIPALE (5). The tokens of the most frequent collocation, that with TURISTICO, are more or less in equal measure singular and plural. In some instances, one may detect a slightly negative semantic prosody, as if denoting places/sites etc. that are very much *on* the beaten path, (e.g. *Ma la vera Brac è molto di più di queste attrazioni turistiche e per chi vuole scoprirla davvero* 'But there is much more to Brac than these tourist attractions, and for those who really want to discover it'). In other cases, the prosody is decidedly positive (e.g. *Non altrettanto scontato è invece il fatto che siano considerate attrazioni turistiche a pieno titolo, di quelle che, per intenderci, valgono una deviazione, se non il viaggio* 'The fact is not so obvious, however, that they are considered fully-fledged tourist attractions, those which, to make things clear, are worth taking a detour for, if not the whole trip').

The tokens of the collocation with PRINCIPALE are almost all plural, and it is of interest to note that three out of the five form part of an identical string: *Una delle attrazioni principali* 'One of the main attractions'. Similarly, the tokens of ATTRAZIONE NATURALE are all plural, and three out of

the six instances are *una delle attrazioni naturali* 'One of the natural attractions'. Here the semantic prosody is markedly positive, and the superlative is present in the co-text in four instances, one example being *tra le attrazioni naturali più affascinanti e spettacolari del nostro Paese* 'Amongst the most fascinating and spectacular natural attractions in our country'.

The noun ATRAKCJA is particularly productive in Polish forming pairs with 9 adjectives, of which two, such as WIELKI and DUŻY are close synonyms. The most frequent collocation is formed with the adjective DUŻY (83) 'large/big attraction'. The tokens of the collocation do not reveal a clear preference for either singular or plural, but almost all use the superlative form of the adjective, therefore conveying positive semantic prosody. In contrast to this, all the tokens of the collocate pair WIELKI ATRAKCJA form part of a three-item string, WIELKA ATRAKCJA TURYSTYCZNA, and, interestingly, only occur in the singular.

The tokens of the collocation with GŁÓWNY (40), 'main attraction', bear some resemblance to the tokens of Italian PRINCIPALE, namely, almost all occur in the plural, and if they are in the singular, they tend to form part of an identical string: *Jedną z głównych atrakcji*, 'One of the main attractions'. The instances of two collocations, DODATKOWY ATRAKCJA (24), 'additional attraction', and CIEKAWY ATRAKCJA (11), 'interesting attraction', generally follow the same pattern. The tokens of these collocations occur both in the singular and the plural. In the case of the latter, they tend to be preceded by quantifiers of amount, for example [*w programie bardzo dużo ciekawych atrakcji z lokalnej flory i fauny,* 'there are lots of interesting attractions related to local flora and fauna in the program']; suggesting a large number of additional or interesting attractions. A similar sense to the abovementioned is conveyed through another pair of collocates, LICZNY ATRAKCJA (7), which exclusively occurs in the plural form and translates into English as 'numerous attractions'. Interestingly, these tokens of DODATKOWY ATRAKCJA and CIEKAWY ATRAKCJA, which are not preceded by the quantifiers of amount, tend to occur at the very beginning of the sentence signalling a novel aspect of the information being conveyed, (e.g. *Ciekawą atrakcją jest też stojący tuż obok kolumny św Trójcy miejski ratusz,* 'An interesting tourist attraction is the townhall located next to the column of the holy Trinity').

The instances of the collocation WAŻNA ATRAKCJA (6), 'important attraction', follows a pattern similar to DUŻA ATRAKCJA, i.e. the tokens of the collocation occur both in the singular and in the plural form, and interestingly they all use the superlative form of the adjective (e.g. *Do najważniejszych atrakcji regionu należą: spływ Dunajcem na drewnianych tratwach,* 'the Dunajec river rafting ride is the most important tourist attraction of the region').

The tokens of the collocation NOWY ATRAKCJA (6) are almost exclusively in the singular form, and similar to English, they are used with respect to theme or aqua parks, and there is also an occasional instance of the superlative NAJNOWSZY, 'the newest', referring to attractions for children (e.g. *także dla rodzin podróżujących z dziećmi. Najnowszą atrakcją dla tych ostatnich* , 'also for the families travelling with kids. The newest attraction for the latter').

Polish also avails of the adjective ATRAKCYJNY which derives from ATRAKCJA. In our dataset it collocates with the following three nouns: MIEJSCE (6) 'place', OFERTA (6) 'offer' and CENA (5) 'price'. In the first case, all the tokens of ATRAKCYJNY MIEJSCE are in the plural and denote a place which tourists would find appealing and worth visiting (e.g. *posiadłości z atrakcyjnymi miejscami odpoczynku dla turystów,* 'estates offering attractive places for tourists to relax'). On the other hand, the two other collocates: ATRAKCYJNY CENA and ATRAKCYJNY OFERTA are used in the sense of there being a bargain, i.e. affordable and not overpriced. These would translate into English as 'reasonable price' and 'good offer', respectively, (e.g. *To była naprawdę wyjątkowo atrakcyjna oferta, tzw . last minute,* 'It was really a good offer, so-called last minute'; *Ich zaletą prócz atrakcyjnej ceny jest znakomita lokalizacja,* 'Their advantage is, apart from a reasonable price, a perfect location').

The collocates of ATTRACTION in the English corpus all have the function of evaluating the reasons for visiting a particular locality and/or event. MAIN (10) would appear at first sight to be a direct counterpart of the Italian PRINCIPALE. However, only half of the instances are in the plural, and none concern *one of the main attractions*. Most of the examples are equative constructions with the collocation being either before or after the copula verb, e.g. *but the main attraction is the food*; *The main attractions are the people and*

the stunning scenery plains with volcanic hills. NEW (8) is again sometimes singular, sometimes plural, with two instances of the superlative NEWEST, implying there are also other 'new' attractions. The pattern is almost invariably 'the new(est) attraction(s) at + PROPER NOUN'. It is of interest to note that the proper noun in question is always that of a theme park (e.g. Alton Towers), a zoo or an aquarium (e.g. Amazon World/Blue Reef), or an educational museum (e.g. National Space Centre). Therefore, NEW would appear to collocate with one of the senses of ATTRACTION, that denoting a sight or an activity which is exciting and novel, particularly appealing to children.

One glaring absence in the English list is a counterpart of ATTRAZIONE TURISTICO and TURYSTYCZNY ATRAKCJA: an equivalent does exist in the English corpus, but to the contrary of the other two languages, it does not adhere to the ADJ+NOUN pattern, illustrating the caveat mentioned at the start of the section. The collocation in question is TOURIST ATTRACTION, a compound noun, that is correctly annotated in the corpus as being NOUN + NOUN. Remarkably, its frequency, 11, is almost identical to that of its Italian and Polish counterparts. En passant, it was noted above that concepts requiring ADJ+NOUN in one language could be expressed as single words in another. One strong candidate as an equivalent of ATTRAZIONE PRINCIPALE and DUŻY/GŁÓWNY ATRAKCJA would be HIGHLIGHT. There are 99 instances of this in the Guardian Travel Corpus, with the singular and plural forms in roughly equal proportions. In comparison, in the BNC, the two forms sum to a frequency of 930.[8]

4. Final remarks

The aim of the current paper is twofold. Firstly, it constitutes an attempt to demonstrate how corpus linguistics techniques can contribute to a better understanding of collocation across languages. Secondly, it aims to explore differences and similarities in collocation patterns in different languages, especially with the aim of garnering information useful to translators. With regards to the first objective, Baker (2018: 56) remarks:

8. Applying the chi-squared test, the presence of HIGHLIGHT in the Guardian Travel Corpus is statistically significant at $p \leq 0.001$, with a result of 92.79.

Every word in a language can be said to have a range of items with which it is compatible, to a greater or lesser degree. Range here refers to the set of collocates, that is other words, which are typically associated with the word in question. Some words have a much broader collocational range than others.

While this observation is extremely insightful, it was destined to remain anecdotal, until the phenomenon of connectivity became a topic of interest in corpus linguistics, and methodologies started to appear that allowed the objective measurement of the feature (see Brezina *et al.* 2015). This study demonstrates how tools developed outside the field of corpus linguistics can be harnessed to highlight the presence of hubs, what we can consider super-collocates, those terms that are particularly productive in the creation of collocate pairs. These are no other than the terms with a "much broader collocational range than others" referred to above. Similarly, isolate pairs, words that only collocate with each other can be identified with ease.

In terms of the second point, an overall observation of the data would suggest that the language of Italian travel journalism is slightly less formulaic than that of English and Polish, at least with regard to adjective/noun pairs.

During the analysis an interesting point emerged which would seem to constitute an exception to the view that the "most important difference between grammatical and lexical choices, as far as translation is concerned, is that grammatical choices are largely obligatory while lexical choices are largely optional" (Baker 2018: 96).

In Italian and Polish, certain adjective/noun collocations, admittedly limited in number, were seen to be statistically significant in both the canonical order and in a variant, marked, order. Hence, what is essentially a grammatical choice, allows an option. Since this phenomenon concerned only certain lemmas, almost exclusively adjectives, it would appear to constitute an important intersection between grammar and lexis. For example, while ANTICO BORGO and BORGO ANTICO are both attested, at such frequencies to be both statistically significant collocations, the first, a variant order form, is twice as frequent. This type of knowledge is of use to the translator who is translating a travel journalism text into Italian and aims to reproduce similar lexical/grammatical choices to those enacted by a travel journalist writing in the target language.

In fact, Castro & Faber (2014: 205), in a paper describing the most representative English and Spanish collocation dictionaries for general language with the aim of evaluating how useful they may be for translators, observe that:

There is a general consensus among translators that phraseological information in lexicographic resources is crucial, especially in the final production of the target language text. In this phase, the translator may need grammatical and syntactic information related to terms, including collocations in the target language.

It is hard to imagine how objective data concerning the lexical/syntactic behaviour of collocations may be gleaned without recourse to the corpus-driven methodologies described in this paper, especially where special domains of human endeavour are concerned.

References

BAKER, Mona. (2018) *In Other Words. A Coursebook on Translation* (3rd ed.), London: Routledge.

BAO, Jiannu. (2014) "Going with the flow: Chinese travel journalism in transition." In: Hanusch, Folker & Elfriede Fürsich (eds.) 2014. *Travel Journalism. Exploring Production, Impact and Culture*. Basingstoke: Palgrave Macmillan, pp.134-151.

BAUMANN, Tania. (2018) *Reiseführer – Sprach- und Kulturmittlung im Tourismus / Le guide turistiche – mediazione linguistica e culturale in ambito turistico.* Bern: Peter Lang.

BONTEMPELLI, Massimo. (1938) *L'avventura novecentista. Selva polemica (1926-1938)*. Firenze: Vallecchi.

BRETT, David. (2018) "Social Network Analysis and the Analysis of Collocations in the Language of Travel Journalism." In Baumann, Tania (ed.) 2018. *Reiseführer – Sprach- und Kulturmittlung im Tourismus / Le guide turistiche – mediazione linguistica e culturale in ambito turistico.* Bern: Peter Lang, pp.183-207.

BRETT, David & Pinna, Antonio. (2015) "Patterns, fixedness and variability: using PoS-grams to find phraseologies in the language of travel journalism." *Procedia - Social and Behavioral Sciences* 198, pp. 52 – 57.

BREZINA, Vaclav; McEnery, Tony & Wattam, Stephen. (2015) "Collocations in context: A new perspective on collocation networks." *International Journal of Corpus Linguistics* 20:2, pp. 139-173.

CALVI, Maria Vittoria. (2000) *Il linguaggio spagnolo del turismo*. Viareggio: Baroni.

CANALS, Jordi & Elena Liverani (eds.) (2010) *Viaggiare con la parola*. Milano: Franco Angeli.

CASTRO, Miriam Buendía, & Pamela Faber. (2014) "Collocation dictionaries: a comparative analysis." *MonTI. Monografías de Traducción e Interpretación* 6, pp. 203-235.

DANN, Graham. (1996) *The Language of Tourism. A Sociolinguistic Perspective*. Wallingford: CAB International.

DE LUCA, Bernardo & Domenico Scarpa. (2012) "Gli scrittori in viaggio". In Scarpa, Domenico (ed.) 2012. *Atlante della Letteratura Italiana. Dal Romanticismo a oggi*. Vol. III. Torino: Einaudi, pp. 812-821.

DUDA, Beata. (2015) *Dyskursywne i tekstowe reprezentacje współczesnej przestrzeni miejskiej* [Representations of contemporary urban spaces through discourse and text]. Katowice: Uniwersytet Śląski. Unpublished doctoral dissertation.

DURÁN-MUÑOZ, Isabel. (2019) "Adjectives and their keyness: a corpus-based analysis of tourism discourse in English." *Corpora* 14: 3, pp. 351–378.

FIRTH, John R. (1957) *Papers in Linguistics 1934–1951*. London: Oxford University Press.

FÜRSICH, Elfriede & Anandam P. Kavoori (2001) "Mapping a critical framework for the study of travel journalism." *International Journal of Cultural Studies* 4:2, pp. 149-171.

GRIES, Stefan Th. (2013) "50-something years of work on collocations: what is or should be next …" *International Journal of Corpus Linguistics* 181, pp. 137–165.

GOTTI, Maurizio. (2006) "The language of tourism as specialized discourse." In: Palusci, Oriana & Sabrina Francesconi (eds.) 2006. *Translating Tourism. Linguistic/Cultural Representations*. Trento: Editrice Università degli studi di Trento, pp.15-34.

GRAF, Paweł. (2018)" Teoria i praktyka dyskursu turystycznego (na przykładzie Słowacji i Rzymu wpisanych w przewodniki)." [Theory and Praxis of the Tourist Discourse (on the Example of Slovakia and Rome as Seen in Guidebooks)]. In: Rejter, Artur; Ewa Biłas-Pleszak; Joanna Przyklenk & Katarzyna Sujkowska-Sobisz (eds.) 2018. *Wędrówka, podróż, migracja w*

języku i kulturze. Katowice: Wydawnictwo Uniwersytetu Śląskiego, pp. 433-448.

HANUSCH, Folker. (2010) "The dimensions of travel journalism: Exploring new fields for journalism research beyond the news." *Journalism Studies* 11:1, pp. 68-82.

HANUSCH, Folker. (2013) "Broadening the focus: the case of lifestyle journalism as a field of scholarly enquiry." In: Hanusch, Folker (ed.) 2013. *Lifestyle Journalism*. Abingdon: Routledge, pp.1-10.

HANUSCH, Folker & Fürsich, Elfriede (eds.) (2014a) *Travel Journalism. Exploring Production, Impact and Culture*. Basingstoke: Palgrave Macmillan.

HANUSCH, Folker & Fürsich, Elfriede. (2014b) "On the relevance of travel journalism: An introduction." In: Hanusch, Folker & Fürsich, Elfriede (eds.) 2014. *Travel Journalism. Exploring Production, Impact and Culture*. Basingstoke: Palgrave Macmillan, pp.1-17.

HUNSTON, Susan. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

JAWORSKA, Sylvia. (2013) "The quest for the 'local' and 'authentic': Corpus-based explorations into the discursive constructions of tourist destinations in British and German commercial travel advertising." In: Höhmann, Doris (ed.) 2013. *Tourismuskommunikation. Im Spannungsfeld von Sprach- und Kulturkontakt*. Frankfurt am Main: Peter Lang, pp. 75-100.

KUDEŁKO, Joanna. (2016) *Językowo-kulturowa rekreacja wizerunku Hiszpanii w polskich tekstach turystycznych w XX wieku* [Linguistic and cultural re-creation of an image of Spain in Polish tourist texts published in the 20th century]. Lublin: Katolicki Uniwersytet Lubelski Jana Pawła II. Unpublished doctoral dissertation.

LOMBARDINILO, Andrea. (2016) "Metafisica quotidiana. Viaggio e società nel giornalismo culturale di Flaiano e Manganelli." *Prospettiva persona* 95-96, pp. 76-81.

MANCA, Elena. (2008) "From phraseology to culture. Qualifying adjectives in the language of tourism." *International Journal of Corpus Linguistics* 13:3, pp. 368–385.

MARCHI, Anna & Taylor, Charlotte (2018) "Introduction: Partiality and reflexivity." In: Taylor, Charlotte & Anna Marchi (eds.) 2018. *Corpus approaches to discourse: A critical review*. London: Routledge, pp. 1-15.

MARGARITO, Mariagrazia; Marie Hediard & Nadine Celotti. (2011) *La comunicazione turistica - Lingue culture istituzioni a confronto / La communication touristique - Langues cultures institutions en face-à-face*. Torino: Libreria Cortina.

MOROZ, Grzegorz. (2015) "Ryszard Kapuściński: between Polish and Anglophone travel writing." *Studies in Travel Writing* 19:2, pp. 169-183.

NIGRO, Maria Giovanna. (2006) *Il linguaggio specialistico del turismo. Aspetti storici, teorici e traduttivi*. Roma: Aracne.

PINNA, Antonio. (2018) "Affect in the language of travel journalism". In: Baumann, Tania (ed.) 2018. *Reiseführer – Sprach- und Kulturmittlung im Tourismus / Le guide turistiche – mediazione linguistica e culturale in ambito turistico*. Bern: Peter Lang, pp. 151-182.

PIROLLI, Brian. (2019) *Travel Journalism. Informing Tourists in the Digital Age*. London: Routledge.

PODKIDACZ, Waldemar. (2004) "Aksjologizacja obrazu świata w tekstach przewodników turystycznych [Using axiology to depict the world image in the guidebooks]." *Poradnik Językowy* 7, pp.45-58.

RAJTER, Artur. (2004) "Wzorzec tekstowy reportażu podróżniczego w aspekcie ewolucji gatunku mowy: próba syntezy." [The textual model of travel reportage in the context of the evolution of speech genres: An attempt at synthethis]. In: Malinowska, Elżbieta & Dariusz Rott (eds.) 2004. *Wokół reportażu podróżniczego, tom 2* [Of travel reportage, vol. 2]. Katowice: Wydawnictwo Uniwersytetu Śląskiego, pp. 8–17.

RAMAN, Usha & Choudary, Divya. (2014) "Have travelled, will write: User-generated content and new travel journalism." In: Hanusch, Folker & Elfriede Fürsich (eds.) 2014. *Travel Journalism. Exploring Production, Impact and Culture*. Basingstoke: Palgrave Macmillan, pp. 116-133.

SERIANNI, Luca. (1989) *Grammatica italiana*. Torino: UTET Libreria.

STANISŁAWEK, Joanna. (2013) "Ekfraza w tekstach użytkowych na przykładzie ofert turystycznych [Ekphrasis in the texts of tourist offers] ." *Rozprawy Komisji Językowej* 59, pp. 253-262.

TAYLOR, Charlotte & Marchi, Anna. (eds.) (2018) *Corpus approaches to discourse: A critical review*. London: Routledge.

TAYLOR, Christopher. (1998) *Language to Language: A Practical and Theoretical Guide for Italian/English Translators*. Cambridge: Cambridge University Press.

THOMPSON, Carl. (2011) *Travel Writing*. London: Routledge.

WITOSZ, Bożena. (2007) "Gatunki podróżnicze w typologicznym ujęciu genologii lingwistycznej." [Travel genres in the typology of linguistic geneaology]. In: Rott, Dariusz (ed.) 2007. *Wokół reportażu podróżniczego* [Of travel reportage]. Katowice: Wydawnictwo Uniwersytetu Śląskiego, pp. 11–29.

WORLD TOURISM ORGANIZATION UNWTO (2019) *International Tourism Highlights 2019*. Online version: <https://www.e-unwto.org/doi/pdf/10.18111/9789284421152> (last accessed 30th May 2020)

ZAGORSKA Brooks, Maria. (1975) *Polish reference grammar.* The Hague: Mouton.

ŻARSKI, Waldemar. (2013) "Językowo-kulturowy obraz Śląska i Ślązaków w przewodnikach turystycznych." [Linguistic and cultural portraying of Silesia and the Silesians in tourist guides]. In: Ursel, Marian & Olga Taranek-Wolańska (eds.) 2013. *Śląskie pogranicza kultur, t. 2*,[Silesian borders of culture, vol.2] Wrocław: Oficyna Wydawnicza ATUT, pp.159-184.

BIONOTE / BIONOTA

DAVID FINBAR BRETT is a full-time researcher at the University of Sassari and has been working in Italy for over 25 years: initially in the field of teaching English as a foreign language, and more recently in that of research in the sector of English language and translation. His main research interests include corpus linguistics, e-learning and foreign language learning, and computer assisted pronunciation training. He has given numerous presentations on these topics in international conferences and has held workshops on CALL, corpus linguistics and EFL materials development in Italy, France, Slovenia, Spain, Poland and Cyprus.

BARBARA LORANC-PASZYLK holds a PhD in Applied Linguistics from University of Silesia, Poland. She works as assistant professor at University of Bielsko-Biała, Poland. Her research interests focus on exploring various linguistic aspects of telecollaboration as well as innovative uses of digital resources in foreign language teaching and learning. She has published in international journals and edited volumes in the field of second language acquisition.

ANTONIO PINNA has an MPhil in Corpus Linguistics from the University of Birmingham (UK). He works as associate professor of English Language at the University of Sassari (Italy) where he teaches Pragmatics, (Critical)

Discourse Analysis, and English for Tourism Studies at both undergraduate and postgraduate level. His research interests include U.S. Presidential discourse, News discourse, and applications of Corpus Linguistics to various discourse types.

David Finbar Brett è ricercatore a tempo indeterminato presso l'Università di Sassari e lavora in Italia da oltre 25 anni: inizialmente nel campo dell'insegnamento dell'inglese come lingua straniera, successivamente nella ricerca nel settore scientifico Lingua Inglese e Traduzione. I suoi interessi di ricerca principali includono la linguistica dei corpora, l'e-learning e l'apprendimento delle lingue straniere, gli strumenti digitali per l'apprendimento della pronuncia. Ha presentato le sue ricerche su questi argomenti in numerosi congressi internazionali e ha tenuto workshop sul CALL, la linguistica dei corpora e lo sviluppo dei materiali in Italia, Francia, Slovenia, Spagna, Polonia e Cipro.

Barbara Loranc-Paszylk ha conseguito un dottorato di ricerca in Linguistica Applicata presso l'Università della Slesia, Polonia. Attualmente è Professore Associato all'Università di Bielsko-Biała, Polonia. I suoi interessi di ricerca riguardano l'esplorazione di vari aspetti linguistici della tele-collaborazione oltre agli utilizzi innovativi delle risorse digitali per l'insegnamento e l'apprendimento delle lingue straniere. Ha pubblicato su riviste internazionali e ha curato diversi volumi nel campo dell'acquisizione della seconda lingua.

Antonio Pinna ha un MPhil in Linguistica dei Corpora presso l'Università di Birmingham (UK). Attualmente è professore associato di Lingua Inglese all'Università di Sassari (Italia) dove insegna Pragmatica, Analisi Critica del Discorso e Lingua Inglese per il Turismo nei corsi di studio triennali e magistrali. Tra i suoi interessi di ricerca ci sono il discorso presidenziale statunitense, il linguaggio della stampa britannica e le applicazioni della linguistica dei corpora a varie tipologie di discorso.

Recibido / Received: 24/05/2020
Aceptado / Accepted: 10/08/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.05

Para citar este artículo / To cite this article:
Kajzer-Wietrzny, Marta & Łukasz Grabowski. (2021) "Formulaicity in constrained communication: An intermodal approach" In: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection. MonTI* 13, pp. 148-183.

FORMULAICITY IN CONSTRAINED COMMUNICATION: AN INTERMODAL APPROACH

MARTA KAJZER-WIETRZNY
kajzer@amu.edu.pl
Adam Mickiewicz University, Poland

ŁUKASZ GRABOWSKI
lukasz@uni.opole.pl
University of Opole, Poland

Abstract

In this exploratory study bordering on corpus linguistics, formulaic language and studies on constrained communication (focusing on translation, interpreting, and L2) we aim to verify whether constrained texts found in the Polish-English component of an intermodal EPTIC corpus differ from native texts in terms of use of adjacent word combinations commonly known as bigrams and whether similar patterns can be found across spoken and written registers. To that end, we fit a Poisson regression model with fixed and random effects. The results show that the translated language variety contributes to the higher number of the most frequent bigram types in both spoken and written registers, and that the number of frequent bigrams in texts generally increases when the speech/source speech is delivered impromptu, but the effect is significant only for the written register. The findings reveal the considerable impact of individual variation on formulaicity as most of the bigram variation within both models is explained by text-specific random variables rather than fixed variables.

Keywords: Formulaic language; Interpreting; Translation; Constrained language; Corpus linguistics.

Zusammenfassung

In dieser an Korpuslinguistik, Formelsprache und Studien über eingeschränkte Kommunikation grenzende Forschungsstudie, die sich hier auf Übersetzung, Dolmetschen und L2 konzentriert, wollen wir überprüfen, ob die in der polnisch-englischen Komponente eines intermodalen EPTIC-Korpus gefundenen eingeschränkten Texte sich von den einheimischen Texten unterscheiden in Bezug auf die Verwendung benachbarter Wortkombinationen, oft Bigrams genannt, und ob ähnliche Muster in gesprochenen und geschriebenen Registern gefunden werden können. Dazu erarbeiteten wir das Poisson-Regressionsmodell mit festen und zufälligen Effekten. Die Ergebnisse zeigen, dass die übersetzte Sprache zur höheren Anzahl der häufigsten Bigram-Typen sowohl in gesprochenen als auch in geschriebenen Registern beiträgt und dass die Anzahl der häufigen Bigrams in Texten generell zunimmt, wenn die Sprache / Quellsprache spontan geliefert ist, aber der Effekt ist lediglich für das schriftliche Register statistisch signifikant. Abschließend zeigen die Ergebnisse einen signifikanten Einfluss der individuellen Variation auf die Formelsprache, da man den größten Teil der Bigram-Variation in beiden Modellen eher durch die textspezifischen zufälligen Variablen als durch die festen Variablen erklärt.

Schlagwörter: Formelsprache; Dolmetschen; Übersetzung; Eingeschränkte Sprache; Korpuslinguistik.

1. Introduction

In the last two decades research on translation and interpreting has provided ample support for the assertion that there is no unified way or method of approaching translational and non-translational texts. Consequently, contemporary Translation/Interpreting Studies resemble a cluster of overlapping perspectives, e.g. formal, pragmatic, psycholinguistic, neurolinguistic or corpus linguistic, etc. Following interest in corpus-based and corpus-driven research on translation and interpreting universals (Baker 1993; Laviosa 1998, 2002; Mauranen 2000; Olohan 2004; Chesterman 2004; Ulrych & Murphy 2008; Kajzer-Wietrzny 2012; De Sutter et al. 2013; Grabowski 2013; Biel 2014; Szymor 2018), i.e. repeatedly observed characteristics of translations, more attention has been paid recently to the concept of 'constrained communication' (Kruger 2012, Kruger & Van Rooy 2016a, Kotze 2019), where language use is constrained by mediation (translation/interpreting), foreign language use or both. For example, Lanstyák and Heltai (2012)

hypothesize that both translation and non-native production share the main constraint, i.e. the need to manage two languages and the ensuing "linguistic uncertainty resulting from the parallel activation of two languages". At the same time, they point out that constrained varieties differ in that non-native language/text production involves descriptive language use (i.e. it does not depend on any other text), translation being additionally constrained by interpretive language use (i.e. it is dependent on the source text).

Current research on translated English and non-native English appears to validate the view that there are similar linguistic tendencies with respect to "features resulting from processing strain" (Kruger & Van Rooy 2016a: 26). Among the constrained varieties, translation is usually viewed as the extreme case of bilingual activation and perceived as particularly constrained at the psycholinguistic level due to rapid bi-directional switching between languages and activation both at the level of language in general as well as the specific linguistic variants of the source text (Kruger & van Rooy 2016b: 121). On these grounds, we can argue that simultaneous interpreting is an even more extreme case due to the time constraint, which makes the entire process more rapid than written translation. Thus, it is imperative that the analysis be expanded to include interpreting, as in many respects it shows different linguistic patterns compared to written translation (cf. Sandrelli & Bendazzoli 2005; Shlesinger & Ordan 2012; Defrancq et al. 2015; Kajzer-Wietrzny 2015; Bernardini et al. 2016; Ferraresi et al. 2019). For the same reason, spoken non-native texts should also be included in this paradigm because, like interpreting, such texts are not subject to intermediate intervention (e.g. editing). That is why they may also reveal peculiar linguistic patterns.

The rapidly growing literature on constrained communication also points to shared cognitive limitations in the production of non-native and translated texts and, as pointed by Aston (2018: 84-85, after Forster 2001), cognitive resources seem to be liberated by the use of formulae which are also believed to be used in greater proportions in settings requiring more processing effort. In an exploratory study of interpreter discourse in the European Parliament, Aston (2018: 83) looks at the frequency of n-grams with 5 words or longer found in transcripts of simultaneous interpretations and argues that "the language of fluent interpreters relies heavily on recurrent formulaic phraseologies." As the formulaic repertoire of second language speakers is

supposed to be smaller than that of native speakers, Aston (2018: 83) points "to the need for interpreters working into their second language to enlarge this repertoire as far as possible", especially that linguistic preferences of translators and interpreters do not always reflect native speakers' preferences manifested, among others, in the use of the so-called formulaic language.

Although the concept of 'formulaic language' (or 'formulaicity') has been explored by linguists of various schools and research traditions as well as with various purposes in mind (descriptive, applied or otherwise), the debate about its theoretical status has been rather inconclusive and there has been little agreement as to its precise definition and operationalization (Wray 2002, 2007; Schmitt & Carter 2004; Wood 2015; Forsyth & Grabowski 2015; Buerki 2016, 2020; Myles & Cordier 2017; Pęzik 2018; Nelson 2018; Siyanova-Chanturia & Omidian 2019; Szerszunowicz 2020). Consequently, a wide variety of criteria is used in the identification and classification of various manifestations of formulaic language in texts, e.g. distributional (frequency, distribution range, collocational strength measures), syntactic (fixed versus flexible word order, substitutability), semantic (non-/compositionality of meaning), pragmatic (genres, registers etc.), to name but a few. That is why 'formulaic language' acts as an umbrella term for the many different types of linguistic items or operationalizations of recurrent patterns of language use, such as collocations, bigrams, binomials, multi-word verbs, speech formulae, routine formulae, pragmatic routines, pragmatemes, lexical bundles, idioms, winged words, proverbs, sayings, clichés etc.

As in this paper we adopt a textual, quantitative corpus linguistic perspective on constrained communication, frequency and repetition become the main criteria for us to identify formulaic language. As such, frequency constitutes a statistical property of multi-word combinations because language users, be it in translation, interpreting or native language use, generally give priority to the linguistic items that are frequently used in their discourse communities. Moreover, since formulaic phrasings are inherently repetitive, we believe that focusing on frequent bigram types will provide a cursory insight into the amount of formulaic language in the study corpus, similar to Altenberg's (2018) research on recurrent n-grams. Furthermore, the frequency-driven approach to study formulaic language is particularly attractive for the analyses of routinized or clichéd texts because such texts

rely more on restricted sets of prefabricated text chunks, notably when compared with more creative texts (literary or otherwise) (Forsyth & Grabowski 2015). Hence, the frequency-driven approach focusing on the use of contiguous sequences of words (e.g. bigrams, trigrams) seems to be well justified when exploring the properties of somewhat restricted and clichéd European Parliament discourse (Kajzer-Wietrzny 2012).

Thus, in this exploratory study, which interfaces corpus linguistics, formulaic language and studies on translation, interpreting as well as L2, we aim to verify whether constrained spoken texts (read out and delivered impromptu) differ from native spoken texts in terms of use of adjacent word combinations (bigrams). We look at the formulaicity of texts produced in English by native English speakers and native speakers of Polish as well as that from interpreters at the European Parliament working into their B (L2) language and Polish-English translations of the European Parliament debates. The study aims to verify whether such constrained texts differ from native texts regarding the number of most frequent bigram types and whether similar patterns can be found across spoken and written registers. More precisely, we put forward the claim that, due to increased processing constraints, interpreters, translators and non-native speakers rely more on the use of formulaic language (operationalized as the number of bigram types among the most frequent bigrams in the registers under study) than native speakers, and that the mode of delivery of the text and delivery rate, particularly in the case of spoken production, might impact the number of distinct bigram types, which is our working hypothesis. In other words, the discussion presented in this paper focuses on the factors that impact the use of formulaic language in constrained communication with the European Parliament discourse as a case in point. In what follows, we describe the research material and methodology of our study in greater detail.

2. Translation, interpreting and non-native language as forms of constrained communication

Constrained language is an umbrella concept marrying two independent research directions focusing on translation and foreign language. It recognizes the shared cognitive constraints in those two communicative situations

involving bilingual activation, which may help identify their shared linguistic characteristics. As already mentioned, while translation is source text dependent, non-native production is not at the same level. Both translation and non-native language use are constrained by parallel bilingual activation and the ensuing linguistic uncertainty (Lanstyák & Heltai 2012). Kruger and van Rooy (2016a) suggest also that the common denominator of the constrained language varieties is the "transfer or cross-linguistic influence (CLI)". It can therefore be expected that patterns observed in one form of language contact may be reflected in other language contact conditions.

Even though the existence of such links was suggested over a decade ago (e.g. Halverson 2003; Chesterman 2004), it has only recently been addressed in empirical investigations of different instances of constrained communication together with factors such as "processing complexity and cognitive effort, (communicative) risk avoidance, and cross-linguistic influence (CLI)" (Kruger & De Sutter 2018: 252). Kruger (2018: 10) argues that

> "constrained varieties may be seen as probabilistically conditioned by five overarching and interacting constraint dimensions (conceived as continua rather than binaries), enabling us to model the similarities and differences between varieties:
> (1) Language activation (monolingual—bilingual)
> (2) Modality and register (spoken—written—multimodal) (3) Text production (independent/unmediated dependent/mediated)
> (4) Proficiency (native/proficient—non-native/learner)
> (5) Task expertise (expert—non-expert)".

As this research direction is relatively new, the studies addressing those five constraint dimensions are still relatively scarce and mostly limited to written register. Also, the very few studies conducted so far focus on the comparisons of texts written in the English language. For example, Kruger and Van Rooy (2016a: 26) showed that translated English and non-native written English show similar tendencies with respect to "increased formality, explicitation of information through elaboration and specification, and features resulting from processing strain". Expertise and proficiency also play a role as "less advanced non-native varieties and translated texts avoid informality features in written registers to a much larger degree than more advanced

non-native varieties and native varieties", and that this tendency, which is likely to be caused by a risk-avoidance strategy, diminishes with greater proficiency (Kruger & Van Rooy 2018: 237). A similar conclusion seems to transpire from the study by De Sutter and Lefer (2020), who examined the use of explicit variant (*that* vs. zero complementizer) and observed that it is most often chosen by learners with little writing experience, followed by less experienced native writers, only then by translators and non-translators. What is more, the "the two groups of professionals hardly differ, although in some very specific contexts translators use explicit that somewhat more often than non-translators" (De Sutter & Lefer 2020). Not only is explicitation more frequent in constrained communication, but also certain structures indicating implicit syntactic relationships are underused when compared to original native texts (Ivaska et al. submitted). In other words, it seems that non-native authors use less implicit relationships than translations and the pattern is consistent across different registers. In a similar vein, Rabinovich et al. (2016: 1871) show that lexical richness of constrained varieties is lower, idiomatic expressions and pronouns are differently distributed and the proportion of more frequent words is much higher as well as that of cohesive devices.

Studies showing multimodal approaches to constrained communication are still few and far between, but they seem to confirm that non-native and translated texts share a common ground also in the spoken register. A small-scale study of non-native and interpreted texts (Kajzer-Wietrzny 2018: 111) shows that a tendency to an increased frequency of optional connective *that* can be observed in both spoken varieties of constrained communication. Kajzer-Wietrzny et al. (2019) observe that mediation has an equalizing effect on formality differences causing the mediated written and spoken varieties to be closer to each other on the formality spectrum than the native non-mediated written and spoken varieties.

Another study on lexical diversity in constrained language examined through the lens of lexical density, variability, evenness, dispersion, rarity and semantic disparity (Kajzer-Wietrzny & Ivaska 2020) confirms that both spoken and written constrained texts show a tendency similar to the "equalizing effect". First observed by Shlesinger (1989) with reference to

interpreting, it is supposed to diminish "the orality of markedly oral texts and the literateness of markedly literate ones". Moreover, constrained texts in general tend to shift towards the middle of the involved vs. informational speech production continuum. On the other hand, interpreted and translated texts show a greater uniformity, i.e. are "more like each other" (as observed by Baker 1996 and Laviosa 1998 with reference to translations), which hints at the possibility of translation-specific levelling-out effect.

It is also important to note that, especially in the context of the European Parliament, the mode of speech delivery also affects the patterns of language use in mediated discourse, both written and spoken. While orthographic transcripts[1] are considered source texts of spoken mediated texts (English interpretations), verbatim reports drafted in Polish and available at the EP website constitute the sources of written mediated texts (English translations). Moreover, it is hypothesized here that the mode of delivery of a source event (i.e. the original speaker delivering a speech at the European Parliament) impacts characteristics of both spoken (i.e. orthographic transcripts of the speech) and written texts (i.e. verbatim reports in all language versions). The impact of mode of delivery on interpreting seems to be more direct, but it is plausible that at least a selection of typically oral or typically written features that can be attributed to the mode of delivery of source events are transferred also to the target texts of translations of the verbatim reports of these. This is reflected, for example, in the lexical diversity of both simultaneous interpretations of speeches delivered at the EP as well as translations of the verbatim reports of these speeches (Kajzer-Wietrzny & Ivaska, 2020) and in cohesion patterns in these texts (Kajzer-Wietrzny, accepted). The impact of this factor seems worth investigating also in the context of formulaicity.

1. While orthographic transcripts of the source and target speeches were manually produced in the compilation process of the EPTIC corpus (Section 3), verbatim reports of the source speeches and their translations into the EU official languages have been published at the EP website (translations into EU official languages are available for all the speeches given at the EP until mid-2011).

In another study (Kajzer-Wietrzny, accepted) we also show the tendency towards increased cohesion in constrained varieties, which is, however, realized in different ways. The overall frequency of cohesive devices (excluding phrase-level coordinators) points to a significantly higher number of cohesive devices in translations when compared to native and non-native texts. A similar significant effect, albeit slightly weaker, is visible in interpretations. Non-natives do increase the overall level of cohesion of their utterances in the spoken register with an overuse of phrase-level coordinators. All those findings encouraged us to undertake a further study, this time focusing on formulaicity in spoken and written unconstrained and constrained language varieties. We believe that the patterns of use of bigrams (as we operationalize formulaic language) will cast more light on the specificity of constrained communication.

3. Methodology

3.1. Research material

The research material includes the Polish-English components of the European Parliament Translation and Interpreting Corpus[2] (henceforth EPTIC), which is an intermodal corpus rich in contextual information (e.g. speaker, delivery rate, mode of delivery of the text/source text). The texts compiled in EPTIC include speeches first delivered at the plenary sittings of the European Parliament by MEPs or Commissioners and simultaneously interpreted into official EU languages. Subsequently, verbatim reports were drawn and until 2011 they were also translated and published on the EU Parliament website (Figure 1).

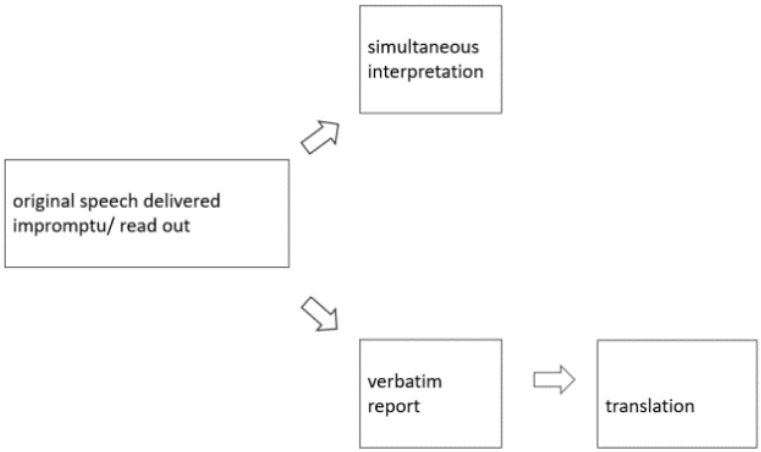

Figure 1. Text cycle at the European Parliament (adapted from Defrancq et al. 2015: 202)

EPTIC sub-corpora, which enable one to conduct a number of different comparisons (e.g. interpretations vs translations, interpreted vs non-interpreted language, native English vs non-native English etc.), include the following:

- sources – spoken: orthographic transcripts of the original speeches;
- sources – written: official verbatim reports of the source speeches;
- targets – interpreted: orthographic transcripts of the interpretations;
- targets – translated: translations of the verbatim reports.

The present study was carried out on a dataset comprising English speeches as well as English interpretations and translations from Polish selected from EPTIC and augmented with two corpora of non-native speeches delivered by Polish representatives (MEPs and Commissioners) delivering speeches in English at the European Parliament[3]. In total, the study corpus comprises 250 texts with 59,540 words (tokens), which are divided into two sub-corpora representing spoken and written registers. These, in turn, are further

3. While the core EPTIC files are based on speeches delivered at the European Parliament in 2011, the two additional corpora contain speeches delivered in 2010 and 2011.

subdivided into native English-originals, non-native English originals and interpretations[4]/translations from Polish into English (Table 1).

Spoken	Written				
Native English Originals*	Non-native English Originals**	Interpretations from Polish into English*	Native English Originals*	Non-native English Originals**	Translations from Polish into English*
9,487 w 34 texts	9,869 w 33 texts	9,567 w 58 texts	9,200 w 34 texts	9,703 w 33 texts	11,714 w 58 texts

** Components of EPTIC*
*** Corpora compiled according to EPTIC guidelines*

Table 1. Analysed dataset

3.2. Unit of analysis, methods and procedures

Apart from the fact that n-gram models, i.e. models based on contiguous sequences of *n* words, have been effective in general in modelling language data in various statistical natural language processing applications, we used bigrams (2-word sequences) as the unit of analysis because they have also been used as indicators of formulaic language in texts (Altenberg 1998). Although not all bigrams represent neat form-and-meaning pairings, they nevertheless tap into the most important aspects of formulaic language (from the corpus linguistic perspective seen primarily as recurrent use of fixed or semi-fixed multi-word units in texts), such as frequency and fixedness (Schmitt & Carter 2004; Wood 2015; Pęzik 2018; Siyanova-Chanturia & Omidian 2019). Also, the frequency-driven approach to study formulaic language is particularly useful for the analyses of clichéd texts because such texts rely more on limited stocks of prefabricated text chunks or boilerplate conventional formulas (Forsyth & Grabowski 2015). Furthermore, Nesi (2012: 422) claims that "n-grams in spoken and written texts tend to be constituted differently [...], and some genres are more formulaic than others". Another rationale behind focusing on bigrams rather than longer sequences of *n* words (e.g. trigrams, fourgrams) is the limited size of the study corpus and, consequently, the problem of data scarcity. Our preliminary inspection

4. Interpretations were carried out from Polish into English by native Polish interpreters.

of the lists of trigrams and fourgrams showed that their number was not sufficient for a large-scale statistical analysis, which would be feasible only with a larger study corpus.

Although in recent years corpus-based research on formulaic language in translation has been flourishing, most studies have been primarily descriptive rather than explanatory and pertained to native versus non-native distinction (e.g. Hu et al. 2016, Ebeling & Ebeling 2018). In this study, we aim to also address the distinction between constrained and unconstrained language as well as attempt to identify those text-related factors that condition the degree of formulaicity (operationalized as the number of bigram types) in spoken and written constrained texts under scrutiny. As mentioned earlier, in this study we investigate formulaicity only within the most frequent bigram types used in the registers under scrutiny. The tools used in the study include Formulib software package (Forsyth 2015), R (2013) and ad hoc scripts written in Python.

We explored formulaicity by identifying the 400 most frequent bigrams in spoken and written sub-corpus, which – given the small size of the sub-corpora – provides a sufficient number for an analysis. In order to avoid a topic bias, we decided to remove from the list all the bigrams that perform referential functions, such as proper names (e.g. *Lady Ashton, of Congo, in Poland, Mr Lukashenko*) or bigrams related to topics of particular speeches (e.g. *construction products, cohesion policy, foreign policy, Christians in*) as these were bound to be more dictated by the discussed problem than the hypothesized cognitive processes that might constrain the investigated forms of bilingual communication. The manual filtering procedure resulted in the selection of 354 and 352 bigrams in spoken and written registers, respectively. From those two samples, we selected those bigrams that were found in all the sub-corpora in the spoken (215 bigram types) and written (237 bigram types) dataset under scrutiny. Next, using ad hoc scripts written in Python, we checked whether each bigram type occurred in each text, which eventually enabled us to count the number of these highly frequent bigram types in each text in each sub-corpus.

As in this paper we focus on identification of factors/predictors that impact the number of bigram types (count data, i.e. non-negative integer

numbers), we used a Poisson regression model[5]. In short, Poisson regression is a type of a Generalized Linear Mixed Model (GLMM) that is typically employed to model count data and contingency tables (Winter 2019: 247), which in this study are matrices of bigram counts. We hypothesize that these counts depend on multiple independent variables (predictors), e.g. mode of delivery or delivery rate, which are our fixed effects. Importantly, predictors in Poisson regression models can be a mixture of numeric and categorical variables. As in any GLMM model, an individual slope in Poisson regression models provides an estimate of the multiplicative change in the response variable (e.g. the number of bigram types) for a one-unit change in the corresponding predictor (e.g. a delivery rate) (Scherber 2019b). For example, if the slope equals -0.12 then for a one-unit change (1 word per minute) in delivery rate the number of bigram types decreases $e^{-0.12}$ fold. Since we have a potentially large pool of speakers, translators, interpreters and topics of the texts under scrutiny (due to the number of observations we cannot include all of them within our model), we decided to include Text IDs as random effects into our model[6]. Without them we would risk having loads of unaccounted variation. Our analysis is thus based on a mixed-effects model and in order to fit it in R (2013), we used *lme4* package (Bates et al. 2015).

Bentz and Winter (2013) describe assumptions to be met in this type of analysis. For example, random effects in mixed models should have 5 to 6 levels at a minimum, which is a criterion that has been met in all models analyzed here[7]. Similarly, an "important assumption of the Poisson distribution is that the sample mean and the sample variance are identical" (Bentz & Winter 2013) applying to distribution of a response variable, which in this study is a count variable. If sample variance exceeds the mean it indicates overdispersion. In the current analysis, none of datasets (neither spoken

5. For more on statistical modeling (linear models, generalized linear models and mixed models), see, Hastie et al. (2016), Kuhn & Johnson (2018), Scherber (2017, 2019a, 2019b), Winter (2019), the latter focusing primarily on linguistic data.
6. We have only included random intercepts, as the inclusion of random slopes was impossible due to an insufficient number of observations.
7. More precisely, our random effects (Text ID) constitute a factor with *n* levels (particular Text IDs) which come from a probability distribution because, potentially, we had infinite number of levels from which our texts could have come (although the EPTIC corpus is restricted in size and composition).

nor written) showed signs of overdispersion as shown by the overdisp_fun[8]. Another important issue involves zero-inflation, which is the case when there is an excessive number of zero-occurrences in the dataset. None of the reported regression models showed any sign of significant zero-inflation and none of the fixed or random effects were highly correlated.

Model summaries and R2m and R2c point to how much of the variation in the data is explained by the fixed effects accounted for in the model and how much can be explained by the full model including random intercepts (see Appendix 2 and Appendix 3). Marginal and conditional R2 were calculated in R with the MUMIN package (Barton 2019). It is also worthwhile emphasizing that no likelihood ratio tests aiming at establishing the contribution of single effects to the model were carried out as, according to Bolker et al. (2009: 132), "the LR test is not recommended for testing fixed effects in GLMMs, because it is unreliable for small to moderate sample sizes." In such cases, Bolker et al. (2009: 132) "recommend against using the LR test for fixed effects unless the total sample size is and number of blocks are very large", which is not the case in the reported study. Additionally, the variables included in the analysis were theoretically motivated and therefore we did not conduct any model comparison.

3.3. Research questions, hypotheses and study stages

This paper is an attempt to explore formulaicity – operationalized as the number of most frequent bigram types – in constrained communication using the European Parliament discourse as a case in point. The study aims to provide answers to the following research questions:

1. What is the most important factor/predictor (language variety, mode of delivery, delivery rate[9]) that impacts the degree of formulaic language in constrained communication versus native texts?
2. Are the observed patterns the same across spoken and written registers in the case of constrained and native texts?

8. Bolker et al. (2020). GLMM FAQs. (URL: https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#overdispersion)
9. Delivery rate as a variable will only be examined in the models regarding the spoken register.

We expect that, due to increased processing constraints related to bilingual processing and/or interpretive language use, interpreters, translators and non-native speakers rely more on the use of formulaic language – measured by the number of bigram types among the most frequent bigrams in the registers under study – than native speakers, and that the mode of delivery of the text and delivery rate, particularly in the case of spoken production, might impact the number of distinct bigram types. Additionally, we hypothesize that increased speed of delivery may contribute to a greater processing effort in the spoken form of constrained communication, which has already been proved in interpreting (Plevoets & Defrancq 2016).

The study will be conducted in a number of stages. First, we will fit Poisson regression models with fixed and random effects to the data obtained from the spoken register[10]. Next, we will repeat the same procedure as applied to the written register. In the last stage, we will compare the results and discuss their implications, paying attention to limitations of this study. In what follows, we present the study findings.

3. Results

As mentioned earlier, in order to provide answers to the research questions, we fitted Poisson regression models with fixed and random effects. The total number of bigram types was modelled as a function of the following predictor variables: text variety, mode of delivery of the source (and speed of delivery in the case of spoken register) adjusted by an exposure variable, which is in this case the number of bigrams in individual text (z-scored). Text-specific random intercepts were also included for the effect of text variety and mode of delivery (and speed of delivery in the case of spoken register) on the number of bigram types (only random intercepts, as the inclusion of random slopes was impossible due to an insufficient number of observations). The source texts of interpretations and translations were in many respects identical[11] and therefore the models for spoken and written registers

10. The dataset, together with the statistical analyses, can be accessed at: https://osf.io/7ktm8/
11. When compared to orthographic transcripts, the verbatim reports analyzed here lack the typical features of orality e.g., repetitions, truncated words etc.; syntactic adjustments are mostly made in the case of discontinued sentences and lexical

were fitted separately. In all models, native English speeches (spoken or written) are used as intercepts.

3.1. Number of bigram types in spoken register

We start by looking at the number of the most frequent bigram types in constrained and non-constrained spoken registers. The first model[12] estimates (1) how the number of bigram types changes as a function of the fixed predictor variables, i.e. text variety, mode of delivery and speed of delivery of the original speech (expressed in words per minute) adjusted by an exposure variable: the number of bigrams in individual text (z-scored) and (2) the variability among the levels of the random effect, i.e. individual texts.

Effect plots (Figure 2) illustrate the general tendencies that can be inferred from the GLMM modelling the patterns in spoken register. First, as visible in the Text Variety effect plot, the number of most frequent distinct bigram types increases with the number of constraints that the users of language have to handle. Thus, said number is the lowest in the spoken native variety, where the speakers are not constrained by either bilingual processing or by the message of the source text; it is higher among the speakers of a foreign language and the highest among interpreters, who transfer someone else's message into a non-native tongue. Second, delivery effect plot shows that when the speakers deliver their speech impromptu, or interpreters interpret a speech that was delivered by the original speaker impromptu, the use of distinct most frequent bigram types in a text increases. The tendency is reverse when the speeches are read out. Finally, the faster the speakers deliver their speeches or the faster the original speakers deliver the speeches that interpreters interpret, the greater the use of distinct

changes are rare. The exact scope of the changes from the spoken parliamentary discourse to the written representation in the analyzed dataset was not measured. It is likely, however, that as in the case of the Hansard, the written representation of the EP debates is not a "hazard" (Mollin 2007) for many linguistic features of interest (Kotze et al. in review).

12. Bigramsspoken <-glmer(CommonBigramTypesNumber~TextVariety+Delivery+ST-WPM+offset(TotalBigramsInText)+(1|TextID)

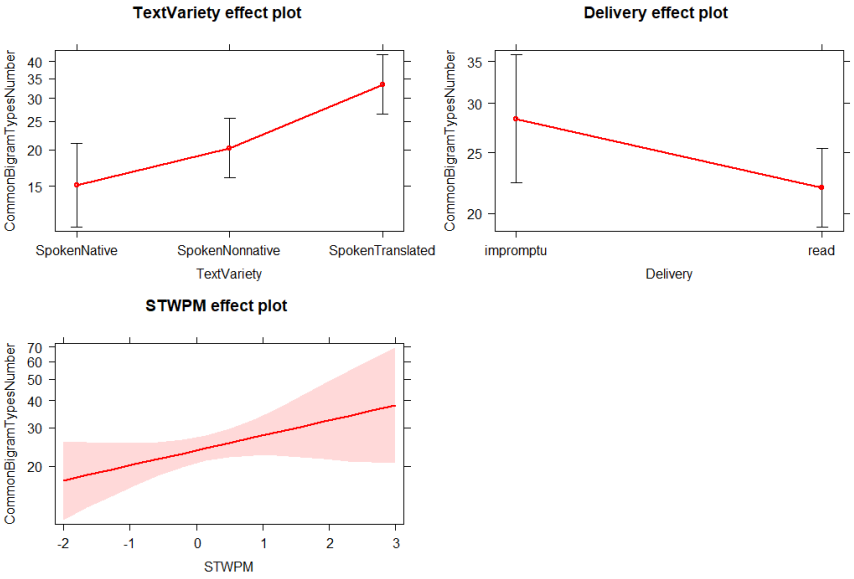

Figure 2. The number of bigram types as a function of fixed and random effects in spoken register

most frequent bigram types, which in this study are used to operationalize formulaicity.

It is worth noticing, however, that not all these trends are statistically significant, which can be inferred from the model summary (see Appendix 2). The outputs of the analysis based on a generalized linear mixed model with Poisson distribution show that within the spoken register both constrained varieties, i.e. non-native language speech and simultaneous interpretations, are characterized by a higher number of distinct most frequent bigram types than spoken native English texts (as illustrated in Figure 2). Still, with estimates at 0.2900 (p=0.14823) in the case of non-native English speeches and 0.7896 (p=0.00127) in the case of interpretations into English, only the difference between the intercept and the latter is statistically significant. The impact of the mode of delivery of the source (i.e. whether the source speech was read out or delivered impromptu) on the number of different bigram

types in a text approaches statistical significance (p=0.07497). Read out speeches, though, seem to contain, in general, a smaller number of different bigram types (estimate -0.2530) than impromptu speeches. In general, the speed of delivery of the (original) speech increases the number of bigram types in a text but its impact is not significant (estimate 0.1579, p=0.11763). It needs to be noted that in the reported regression analysis a large share of variation within the data was explained by the full model including both fixed effects and random intercepts, while fixed effects account only for almost 15% of the variation (as indicated by R^2m). This means that individual text-related effects contributed most to the variation of the number of bigram types (see Appendix 2 for full results). This observation accords with our decision to include the random effects (text ID) into the model as, without it, we would not have been able to capture loads of variation in the model.

3.2. Number of bigram types in written register

Let us now inspect the number of most frequent bigram types in constrained and non-constrained written registers. The second model[13] estimates (1) how the number of bigram types changes as a function of the fixed predictor variables, i.e. text variety and mode of delovery of the original speech adjusted by an exposure variable: number of bigrams in individual text (z-scored) and (2) the variability among the levels of the random effects, i.e. individual texts.

Effect plots (Figure 3) illustrate the general tendencies that can be inferred from the second GLMM. First, the TextVariety effect plot shows that the number of distinct most frequent bigram types increases with the number of constraints that the users of language have to deal with. It can be seen that said number is the lowest in the written native variety, where the authors are constrained by neither bilingual processing nor the message of the source text; it is higher among the authors using a foreign language, and it is the highest among translators, who transfer someone else's message (with the caveat that their native tongue is unknown). Second, the Delivery effect plot shows that when the text represents a verbatim report of a speech that was originally delivered impromptu or when translators translate a speech

13. Bigramswritten <- glmer(CommonBigramTypesNumber~TextVariety+Delivery+offset(TotalBigramsInText)+(1|TextID)

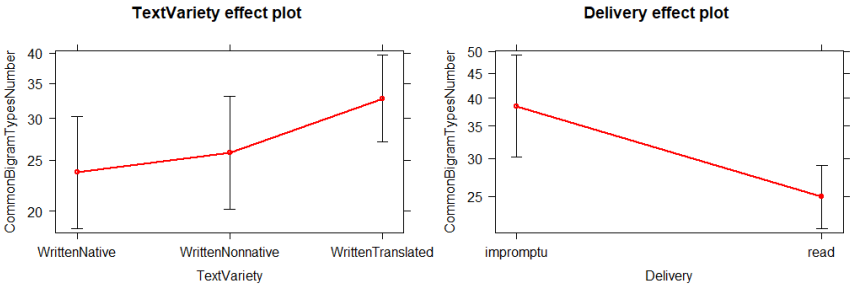

Figure 3. The number of bigram types as a function of fixed and random effects in written register

that was delivered by the original speaker impromptu, the use of distinct most frequent bigram types in a text increases. The observed tendency is reverse when the speeches are read out.

Similarly to the previous model, not all the tendencies are statistically significant. The outputs of the analysis based on a generalized linear mixed model with Poisson distribution (see Appendix 3) show that within the written register both constrained varieties, i.e. non-native language speech and simultaneous interpretations, are characterized by a higher number of different bigram types than written native English texts. Again, with estimates at 0.08581 (p=0.62790) in the case of non-native English texts and 0.32417 (p=0.04154) in the case of translations into English, only the difference between the intercept and the latter is statistically significant. The impact of the mode of delivery of the source (i.e. whether the source speech was read out or delivered impromptu) on the number of different bigram types in a text is statistically significant (p=0.00329) with the read out speeches containing, in general, a smaller number of different bigram types (estimate -0.43241) than impromptu speeches. As in the case of the earlier model, a large share of variation within the data was explained by the full model including both fixed effects and random intercepts, whereas fixed effects account only for approximately 11.5% of the variation (as indicated by R^2m). This means that individual text-related effects contributed most to the variation of the number of bigram types (see Appendix 3 for full results).

Similar to the model described in Section 3.1, the decision to include random effects into the model has been justified.

4. Discussion

In this study we were primarily interested in, first, the identification of predictors of formulaic language in constrained versus native texts and, second, the verification whether the same patterns were observed in native and constrained texts, both spoken and written. Formulaicity was operationalized as the number of the most frequent bigram types, which was our response count variable. As for the potential predictors, we focused on text variety, delivery rate, mode of delivery (fixed effects) and text ID (random effects), which we fit in a mixed-effects model using Poisson regression.

It transpires from both regression models applied to spoken and written registers, respectively, that the translated variety is the main predictor of the number of most frequent bigram types in both registers. A similar trend can be observed in the spoken and written non-native variety, but the estimates do not diverge significantly from the intercept, i.e. the native texts. It may also be argued that the constrained varieties across registers seem to pattern together in a similar way, yet they do not differ from the native varieties to the same extent.

As observed by Kotze (2019: 339), the patterns setting translated language apart from non-translated language, in particular the tendencies relating to "cross-linguistic influence, priming or transfer [are] often of subtle and indirect type". Additionally, written text production in the EU setting is heavily standardized, which can further filter out the nuances which are subtle even in genres not subject to such standardization (e.g. literature or journalistic texts). This is also reflected in the results of the present analysis. Fixed effects in the spoken register account for more variation within the data than in the model fitted for the written register (as indicated by a slightly higher value of R2m and lower value of R2c in the spoken model). One of the potential reasons for such tendencies may be the standardizing effect of the editing and proofreading process at the EU institutions, which might render written texts in the studied constrained and non-constrained varieties more similar. Furthermore, the fixed effects in the spoken model

may be augmented by the "double" cognitive constraint imposed on interpreters relating to the process of language mediation and speaking a foreign language, as all interpretations were carried out into L2. As regards the expectations that speakers might attempt to decrease the higher cognitive load related to higher delivery rate with the use of more formulaic language, these have been only partially confirmed. The rate of delivery, indeed, contributes to a higher number of most frequent bigram types in a text, but the effect is not significant.

We also found that the mode of delivery of the (original) speech is a significant predictor in the written register, and it approaches statistical significance in the regression modelling of the spoken one. In general, the impromptu mode of delivery seems to consistently point to an increased use of most frequent bigram types across all varieties. It is clear that the effect of the mode of delivery in spoken register is significant albeit weaker than in written register. This observation ties in with the one made by Shlesinger (1989, cited in Pym 2007: 178) about the equalizing effect of interpreting, which affects "the position of a text on the oral-literate continuum" and ultimately leads to the reduction of the range of this continuum in simultaneous interpreting. This renders markedly oral texts less oral and markedly literate texts less literal. Such tendencies have also been hinted at in other corpus studies on simultaneous interpreting (Dayter 2018, Kajzer-Wietrzny & Ivaska, 2020). Our findings show that the effect of delivery is weaker in the spoken register, meaning that the "equalizing effect" is stronger in the spoken register than in the written one. This is an important implication for interpreter and translator training: register-specific formulaicity features are transferred with varying degree of difficulty across registers and as such they may require additional attention.

5. Final remarks

The results of a quantitative corpus study like this one should be interpreted with caution. It has to be emphasized that the texts included in the EPTIC corpus are, by their very nature, quite short (100-300 words) and, more importantly, the corpus used in this study contains slightly less than 60,000 words (although it is representative of the registers under scrutiny). Also, it

is noteworthy that the very form of translation and interpreting, and hence the study results, could be influenced by the translators' or interpreters' idiolects. This can be taken up in further research on corpora annotated with such metadata. Another confounding variable could be the effect of L1: in the case of written translations of the MEPs' speeches, it is impossible to establish whether they were produced by native speakers.

There are many ways in which this study could be continued further in order to provide more comprehensive answers to the question of whether constrained communication is by its very nature more formulaic than unconstrained communication. Apart from focusing on count variables such as the number of bigram types, it is possible to adopt other units of analysis that have been used in research on formulaicity. For example, it is possible to focus on frequencies rather than counts of recurrent multi-word items in texts (bigrams, trigrams etc.). Apart from n-grams, one can also explore formulaicity by exploring phrase frames (Fletcher 2002), which are sequences of *n* words identical except for one and which provide a neat generalization of recurrent sequences of words in texts). Hence, it is possible to explore the predictors of formulaic language by focusing on measures of pattern variability applied to phrase frames, e.g. VPR (variant-to-phrase frame ratio proposed by Römer (2010: 105)), Hapaxity, Haprate etc. (Forsyth & Grabowski 2015). As such, these metrics constitute continuous response variables and require the use of linear regression models to identify their predictors. In this study we used a single unit of analysis (bigram types), yet it might be necessary in the future to combine multiple units of analysis to obtain more comprehensive findings since formulaicity is a multi-faceted phenomenon and it cannot be fixed at a single level of analysis only. Also, it might be possible to further explore the causal relation between formulaic language in interpreting, on the one hand, and other text-external variables (e.g. interpreter's status, direction of interpreting), on the other.

Furthermore, in this study we have explored formulaicity in constrained communication using English language material only. However, as pointed out by Buerki (2020), the degree to which languages feature formulaic material remains unclear, notably in the rather underexplored translation/interpreting context, which invites further cross-linguistic (e.g. English-French or

English-Spanish) corpus linguistic research using topically matched corpora with texts representing constrained communication.

Another unexplored avenue of future research on formulaicity in constrained communication, notably in translation/interpreting, is the transfer of discourse functions from the source texts to translations/interpretations, which has implications on how the message of the translation/interpretation is comprehended as compared with the source text. Preliminary exploratory research into this matter, conducted with the use of inter-rater agreement metrics and focusing on recurrent phrases with specific discoursal functions (stance expressions, discourse organizers, including polyfunctional items, e.g. *at the end of the day*), revealed that the discoursal functions are often not conveyed in a fixed and stable way (Grabowski & Groom, accepted). This implies that oftentimes the source and target texts (be it written translations or interpretations) are pragmatically understood differently by respective readers. It seems, however, that further research is required to study the rationale behind the modification of the discoursal functions[14] of recurrent formulas in translation as compared with the original. As this study accounts for an early step in research on formulaicity in constrained communication, we hope that it will pave the way to more comprehensive empirical research into this matter in the future.

References

ALTENBERG, Berndt. (1998) "On the phraseology of spoken English: The evidence of recurrent word combinations". In: Cowie, Anthony (ed.), *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 101-122.

ASTON, Guy. (2018) "Acquiring the language of interpreters: A Corpus-based Approach". In: Russo, Mariachiara, Claudio Bendazzoli & Bart Defrancq (eds.), *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer, pp. 83-96.

14. This is related to more general questions of interest to translation/interpreting practice, namely why translators/interpreters often fail to recognize phrases as a holistic unit and translate them literally, etc.

BAKER, Mona. (1993) "Corpus linguistics and translation studies: Implications and applications". In Baker, Mona, Francis, Gill & Toginini-Bonelli, Elena (eds.), *Text and Technology. In Honor of John Sinclair.* Amsterdam: John Benjamins, pp. 233-250.

BARTON, Kamil. (2019) "MuMIn: Multi-Model Inference". Online version: <https://CRAN.R-project.org/package=MuMIn>

BATES, Douglas, Martin Mächler, Ben Bolker & Steve Walker. (2015) "Fitting linear mixed-effects models using lme4". *Journal of Statistical Software* 67:1, pp. 1-48. Online version: <https://arxiv.org/abs/1406.582>.

BENTZ, Christian & Bodo Winter. (2013) "Languages with More Second Language Learners Tend to Lose Nominal Case." In Wichmann, Soren & Jeff Good (eds.), *Quantifying Language Dynamics: On the Cutting edge of Areal and Phylogenetic Linguistics*. Leiden: Brill, pp. 96-124.

BERNARDINI, Silvia, Adriano Ferraresi & Maja Miličević. (2016) "From EPIC to EPTIC—Exploring simplification in interpreting and translation from an intermodal perspective." *Target. International Journal of Translation Studies* 28:1, p. 61-86.

BIEL, Łucja. (2014) *Lost in the Eurofog: The Textual Fit of Translated Law*. Frankfurt am Main: Peter Lang Verlag.

BOLKER, Benjamin, Mollie Brooks, Connie Clark, Shane Geange, John Poulsen, M. Henry Stevens & Jada-Simone White (2009) "Generalized linear mixed models: A practical guide for ecology and evolution". *Trends in Ecology & Evolution* 24:3, pp. 127-135.

BUERKI, Andreas. (2016) "Formulaic sequences: a drop in the ocean of constructions or something more significant?" *European Journal of English Studies* 20:1, pp. 15-34.

BUERKI, Andreas. (2020) "(How) is Formulaic Language Universal? Insights from Korean, German and English". In Piirainen, Elisabeth, Natalia Filatkina, Sören Stumpf & Christian Pfeiffer (eds.), *Formulaic Language and New Data Theoretical and Methodological Implications*. Berlin: Degruyter, pp. 103-134.

CHESTERMAN, Andrew. (2004) "Hypothesis about translation universals". In: Hansen, Gyde, Kirsten Malmkjaer & Daniel Gile (eds.), *Claims, Changes and Challenges in Translation Studies*. Amsterdam: John Benjamins, pp. 1-13.

DAYTER, Daria. (2018) "Describing Lexical Patterns in Simultaneously Interpreted Discourse in a Parallel Aligned Corpus of Russian-English Interpreting

(SIREN)". *FORUM. Revue Internationale d'interprétation et de Traduction / International Journal of Interpretation and Translation* 16:2, pp. 241-264.

DEFRANCQ, Bart, Koen Plevoets & Cedric Magnifico. (2015) "Connective Items in Interpreting and Translation: Where Do They Come From?". In: Romero-Trillo, Jesus (ed.), *Yearbook of Corpus Linguistics and Pragmatics*. Bern: Springer, pp. 195-222.

DE SUTTER, Gert & Lefer, Marie-Aude (2020) "On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multi-factorial and interdisciplinary approach". *Perspectives*, 28:1, pp. 1-23. https://doi.org/10.1080/0907676X.2019.1611891

EBELING, Jarle & Signe Oksefjell Ebeling. (2018) "Comparing n-gram-based functional categories in original versus translated texts". *Corpora* 13:3, pp. 347-370.

FERRARESI, Adriano, Silvia Bernardini, Maja Miličević & Marie-Aude Lefer. (2019) "Simplified or Not Simplified? The Different Guises of Mediated English at the European Parliament." *Meta: Journal Des Traducteurs / Translators' Journal* 63:3, pp. 717-738.

FORSYTH, Richard. (2015) "Formulib: Formulaic Language Software Library". Online version: http://www.richardsandesforsyth.net/zips/formulib.zip

FORSYTH, Richard & Łukasz Grabowski. (2015) "Is there a formula for formulaic language?" *Poznań Studies in Contemporary Linguistics* 54:1, pp. 511-549.

FOSTER, Pauline. (2001) "Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers". In Bygate, Martin, Peter Skehan & Merill Swain (eds.), *Researching pedagogic tasks: Second language learning, teaching and testing*. London: Longman, pp. 75-93.

FOX, John, Sanford Weisberg, Michael Friendly, Jangman Hong, Robert Andersen, David Firth & Steve Taylor. (2019) Package 'effects'. Online version: https://cran.r-project.org/web/packages/effects/effects.pdf

GRABOWSKI, Łukasz. (2013) "Interfacing corpus linguistics and computational stylistics: translation universals in translational literary Polish". *International Journal of Corpus Linguistics*, 18:2, pp. 254-280.

GRABOWSKI, Łukasz & Nicholas Groom (accepted) "Functionally-defined recurrent multi-word units in English-to-Polish translation: a corpus-based study". *Revista Española de Linguistica Aplicada/Spanish Journal of Applied Linguistics*.

HALVERSON, Sandra. (2003) "The cognitive basis of translation universals". *Target. International Journal of Translation Studies* 15:2, pp. 197-241.

HASTIE, Trevor, Robert Tibshirani & Jerome Friedman. (2016) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. Berlin: Springer.

HU, Xianyao, Richard Xiao & Andrew Hardie. (2016) "How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis". *Corpus Linguistics and Linguistic Theory* 15:2, pp. 347-382.

IVASKA, Ilmari, Adriano Ferraresi & Silvia Bernardini. (Under Review) "Syntactic properties of constrained English: A corpus-driven approach".

KAJZER-WIETRZNY, Marta (2021) "Intermodal approach to cohesion in constrained and unconstrained language" *Target*. https://doi.org/10.1075/target.19186.kaj

KAJZER-WIETRZNY, Marta & Ilmari Ivaska (2020) "A multivariate approach to lexical diversity in constrained language". *Across Languages and Cultures* 21:2, pp. 169-194.

KAJZER-WIETRZNY, Marta. (2012) *Interpreting Universals and Interpreting Style*. Unpublished PhD dissertation. Adam Mickiewicz University, Poznań, Poland.

KAJZER-WIETRZNY, Marta. (2015) "Simplification in interpreting and translation". *Across Languages and Cultures* 16:2, pp. 233-255.

KAJZER-WIETRZNY, Marta. (2018) "Interpretese vs. Non-native Language Use: The Case of Optional That". In Russo, Mariachiara, Claudio Bendazzoli & Bart Defrancq (eds.), *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer, pp. 97-113.

KAJZER-WIETRZNY, Marta, Ilmari Ivaska, Adriano Ferraresi & Silvia Bernardini. (2019) "Thanks very much President…" or "Thank you Mr President..." Investigating formality in mediated and non-mediated discourse". Paper delivered at 49th Poznań Linguistics Meeting, 16-19 September, 2019 in Poznań, Poland.

KOTZE, Haidee, Minna Korhonen, Adam Smith and Bertus van Rooy. (under review) "Salient differences between oral parliamentary discourse and its official written records: A comparison of "close" and "distant" analysis methods" In Korhonen, Minna, Kotze Haidee & Tyrkkö Jukka (eds.), *Parliamentary discourse across time and space: Using big data to study language and society. Studies in Corpus Linguistics*. Amsterdam: John Benjamins.

KOTZE, Haidee. (2019) "Converging what and how to find out why: An outlook on empirical translation studies". In Vandevoorde, Lore, Joke Daems & Bart Defranq (eds.), *New Empirical Perspectives on Translation and Interpreting*. London: Routledge, pp. 333-371.

KRUGER, Haidee & Bertus Van Rooy. (2016a) "Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English". *English World-Wide* 37:1, pp. 26-57.

KRUGER, Haidee & Bertus Van Rooy. (2016b) "Syntactic and pragmatic transfer effects in reported-speech constructions in three contact varieties of English influenced by Afrikaans". *Language Sciences* 56, pp. 118-131.

KRUGER, Haidee, & Bertus Van Rooy. (2018) "Register Variation in Written Contact Varieties of English". *English World-Wide* 39:2, pp/ 214–242. doi:10.1075/eww.00011.kru.

KRUGER, Haidee. (2012) "A corpus-based study of the mediation effect in translated and edited language". *Target* 24:2, pp. 355-388.

KRUGER, Haidee. (2018) "Expanding the third code: Corpus-based studies of constrained communication and language mediation." In Granger, Sylviane, Lefer, Marie-Aude & Penha-Marion, Laura (eds.), *Book of abstracts. Using corpora in contrastive and translation studies conference* (5th edition) CECL papers 1. Louvain-la-Neuve: Centre for English Corpus Linguistics/ Université Catholique de Louvain, pp. 9-12.

KUHN, Max & Kjell Johnson. (2013) *Applied Predictive Modeling*. Berlin: Springer.

LANSTYAK, Istvan & Pal Heltai. (2012) "Universals in Language Contact and Translation". *Across Languages and Cultures* 13:1, pp. 99-121.

LAVIOSA, Sara. (1998) "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose". *Meta* 43:4, pp. 557-570.

LAVIOSA, Sara. (2002) *Corpus-based translation studies: theory, findings, applications*. Amsterdam: Rodopi.

MAURANEN, Anna. (2000) "Strange strings in translated language: A study on corpora". In Olohan, Meave (ed.), *Intercultural Faultlines. Research Models in Translation Studies 1:Textual and Cognitive Aspects*. Manchester: St. Jerome Publishing, pp. 119-141.

MOLLIN, Sandra. (2007) "The Hansard hazard: Gauging the accuracy of British parliamentary transcripts". *Corpora* 2:2, pp. 187-210.

MYLES, Florence & Caroline Cordier. (2017) "Formulaic Sequence(fs) Cannot be an Umbrella Term in SLA: Focusing on Psycholinguistic FSs and Their Identification". *Studies in Second Language Acquisition* 39, pp. 3-28.

NELSON, Robert (2018) "How 'chunky' is language? Some estimates based on Sinclair's Idiom Principle". *Corpora* 13:3, pp. 431-460.

NESI, Hillary. (2012) "ESP and Corpus Studies". In: Paltridge, Brian & Sue Starfield (eds.), *The Handbook of English for Specific Purposes*. London: Wiley, pp. 407-426.

OLOHAN, Meave. (2004) *Introducing Corpora in Translation Studies*. Routledge: London.

PĘZIK, Piotr. (2018) *Facets of prefabrication. Perspectives on modelling and detecting phraseological units*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

PLEVOETS, Koen & Bart Defrancq. (2016) "The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis". *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association,* 11:2, pp. 202-224.

PYM, Anthony. (2007) "On Shlesinger 's proposed equalizing universal for interpreting". In Pochhacker, Franz, Jakobsen, Arnt Lykke & Mees, Inger M. (eds.), *Interpreting studies and beyond: A tribute to Miriam Shlesinger.* Copenhagen: Samfundslitteratur Press, pp. 175-190.

RABINOVICH, Ella, Sergiu Nisioi, Noam Ordan & Shuly Wintner. (2016) "On the Similarities between Native, Non-Native and Translated Texts". In van den Bosch, Antal (ed.) *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 7-12 August. Stroudsburg, PA: Association for Computing Machinery.

RÖMER, Ute. (2010) "Establishing the phraseological profile of a text type. The construction of meaning in academic book reviews". *English Text Construction* 3:1, pp. 95-119.

SANDRELLI, Annalisa. & Claudio Bendazzoli. (2005) "Lexical Patterns in Simultaneous Interpreting: a Preliminary Investigation of EPIC (European Parliament Interpreting Corpus)". *Proceedings from the Corpus Linguistics Conference Series*. Online version: < https://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx>

SCHERBER, Christoph. (2019a) "An introduction to mixed-effects models". Online version: <http://www.christoph-scherber.de/content/PDF%20Files/Mixed%20effects%20models.pdf>

SCHERBER, Christoph. (2019b) "An introduction to generalized linear models". Online version: <http://www.christoph-scherber.de/content/PDF%20Files/Generalized%20linear%20models.pdf>

SCHERBER, Christoph. (2017) "Using R to Interpret Interaction Effects in Statistical Models". *Software Developer´s Journal*. Online version: <https://www.researchgate.net/profile/Christoph_Scherber/publication/312093784_Using_R_to_Interpret_Interaction_Effects_in_Statistical_Models/links/586f67ad08ae329d6215fc4c/Using-R-to-Interpret-Interaction-Effects-in-Statistical-Models.pdf>

SCHMITT, Norbert & Ronald Carter. (2004) "Formulaic sequences in action: An introduction". In: Schmitt, Norbert (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, pp. 1-22.

SHLESINGER, Miriam. (1989) *Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-Literate Continuum*. MA thesis, Tel Aviv University.

SHLESINGER, Miriam & Noam Ordan. (2012) "More Spoken or More Translated?: Exploring a Known Unknown of Simultaneous Interpreting". *Target* 24:1, pp. 43-60.

SIYANOVA-CHANTURIA, Anna & Omidian, Taha. (2019) "Key issues in researching multi-word items". In: Webb, Stewart (ed.), *The Handbook of Vocabulary Studies*. London: Routledge, pp. 511-524.

SZERSZUNOWICZ, Joanna. (2020) "New Pragmatic Idioms in Polish: An Integrated Approach in Pragmateme Research". In: Piirainen, Elisabeth, Natalia Filatkina, Sören Stumpf & Christian Pfeiffer (eds.), *Formulaic Language and New Data Theoretical and Methodological Implications*. Berlin: De Gruyter, pp. 173-196.

SZYMOR, Nina. (2018) "Translation: universals or cognition?". *Target* 30:1, pp. 53-86.

TEAM, R.C. (2013) "R: A language and environment for statistical computing". Online version: <https://www.r-project.org/>

ULRYCH, Margherita & Amanda Murphy. (2008) "Descriptive Translation Studies and the Use of Corpora: Investigating Mediation Universals". In:

Torsello, Carol Taylor, Katherine Ackerley & Erik Castello (eds.), *Corpora for University Language Teachers*. Frankfurt am Main: Peter Lang, pp. 141-166.
WINTER, Bodo. (2019) *Statistics for Linguists: An Introduction Using R*. London: Routledge.
WOOD, David. (2015) *Fundamentals of Formulaic Language*. London: Bloomsbury.
WRAY, Alison. (2002) *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
WRAY, Alison. (2008) *Formulaic language. Pushing the boundaries*. Oxford: Oxford University Press.

Appendix 1. List of frequent bigrams in spoken and written registers

Spoken register (353 bigrams)

of the, in the, it is, the european, to the, and the, on the, for the, we are, we have, i would, mr president, like to, that we, is a, that the, european union, there is, would like, this is, is the, is not, we should, in this, the commission, member states, at the, should be, by the, madam president, with the, will be, has been, the eu, there are, to make, to be, i think, all the, and a, as a, the future, to ensure, the same, from the, we need, of this, i have, i am, president i, this house, the union, that is, and we, and i, need to, must be, the world, not only, does not, but also, that it, do not, and to, and in, european parliament, is important, ensure that, the report, those who, in europe, have been, about the, which is, for a, of a, president the, that this, they are, want to, we do, if we, the situation, that there, it should, the case, what we, we must, it will, is that, have to, we can, in our, a very, in a, in particular, the crisis, and that, who are, which i, what is, i hope, as the, to do, talking about, of european, which will, we cannot, fact that, also like, to thank, the fact, that are, is still, in which, a policy, we will, that in, on this, not the, to ask, let us, have a, but we, but i, the country, between the, would also, think that, the moment, because it, will not, which we, the most, level of, into the, part of, it must, are not, and not, with a, to say, of our, can be, and it, of eu, is to, is no, in my, as we, a new, european commission, the commissioner, this agreement, important to, believe that, of national, for example, are talking, the policy, the people, the common, should not, policy and, which are, to follow, the other, policy of, of course, who have, the time, the need, the euro, say that,

have the, has done, when we, to take, such as, role in, is also, and its, not be, make a, it has, be the, a year, a good, is an, the opposition, important that, and political, the internal, into account, according to, this matter, the council, the second, the recent, the global, percent of, means that, market and, will have, which has, union and, these are, said that, policy we, of people, make sure, future of, for their, cannot be, and there, access to, a certain, union is, to which, the role, the risk, the last, terms of, room for, needs to, jobs and, is about, in terms, in other, based on, and this, who has, we want, we know, we also, was not, to this, to lend, this in, the way, role of, make it, lack of, kind of, is very, a major, a clear, that i, such a, it was, is why, i want, i know, but it, as to, be a, implementation of, this parliament, the possibility, the independent, european budget, cooperation and, very important, to participate, the resolution, parliament has, national level, for innovation, commission has, the president, the elections, the countries, situation and, single market, government of, commission to, agreement and, thousands of, the southern, the republic, the question, the national, the external, research and, president we, member state, in countries, countries in, a resolution, a compromise, where there, welcome the, the subject, the present, the interim, of economic, involved in, included in, in addition, has already, continue to, within the, which were, the number, the budget, states and, social and, for europe, during the, context of, both sides, across the, years ago, under the, today the, to create, the visit, that they, thanks to, thank you, states in, source of, same time, rights in, report on, report is, reform of, policy is, people of, number of, my report, know that, in future, i believe, hope that, have said, for those, first and, debate on, crisis in, case that, are still, and their, after all, you have, years in, when the, were not, visit of, trade in, to avoid, the very, the next, order to

Written register (352 bigrams)

of the, in the, it is, the european, to the, and the, on the, for the, we are, we have, i would, mr president, like to, that we, is a, that the, european union, there is, would like, this is, is the, is not, we should, in this, the commission, member states, at the, should be, by the, madam president, with the, will be, has been, the eu, there are, to make, to be, i think, all the, and a, as

a, the future, to ensure, the same, from the, we need, of this, i have, i am, president i, this house, the union, that is, and we, and i, need to, must be, the world, not only, does not, but also, that it, do not, and to, and in, european parliament, the construction, is important, ensure that, the report, those who, in europe, have been, about the, which is, for a, of a, president the, that this, they are, want to, we do, if we, the situation, that there, it should, the case, what we, we must, it will, is that, have to, we can, in our, a very, in a, in particular, and that, who are, which i, what is, i hope, as the, to do, talking about, of european, which will, we cannot, fact that, also like, to thank, the fact, that are, is still, in which, a policy, we will, that in, on this, not the, to ask, let us, have a, but we, but i, the country, between the, would also, think that, the moment, because it, will not, which we, the most, level of, into the, part of, it must, are not, and not, with a, to say, of our, can be, and it, of eu, is to, is no, in my, as we, a new, european commission, the commissioner, this agreement, important to, believe that, of national, for example, are talking, the policy, the people, the common, should not, policy and, which are, to follow, the other, policy of, of course, who have, the time, the need, the euro, say that, have the, has done, when we, to take, such as, role in, is also, and its, not be, make a, it has, be the, a year, a good, is an, the opposition, important that, and political, into account, according to, this matter, the council, the second, the recent, the global, percent of, means that, market and, will have, which has, union and, these are, said that, policy we, of people, make sure, future of, for their, cannot be, and there, access to, a certain, union is, to which, the role, the risk, the last, terms of, room for, of human, needs to, is about, in terms, in other, based on, and this, who has, we want, we know, we also, was not, to this, to lend, this in, the way, role of, make it, lack of, kind of, is very, a major, a clear, that i, such a, it was, is why, i want, i know, but it, as to, be a, implementation of, this parliament, the possibility, the independent, cooperation and, very important, to participate, the resolution, parliament has, national level, for innovation, commission has, the president, the elections, the countries, situation and, single market, government of, commission to, christians in, agreement and, thousands of, the republic, the question, the national, the external, research and, president we, member state, in countries, countries in, a resolution, a compromise, where there, welcome the, the subject, the present, the interim, republic of,

of economic, involved in, included in, in addition, has already, continue to, within the, which were, the single, the number, states and, social and, for europe, during the, context of, both sides, across the, years ago, under the, today the, to create, the visit, that they, thanks to, thank you, states in, source of, same time, rights in, report on, report is, reform of, policy is, people of, number of, my report, know that, in future, i believe, hope that, have said, for those, first and, debate on, crisis in, case that, are still, and their, after all, you have, years in, when the, were not, visit of, trade in, to avoid, the very, the next, order to

Appendix 2

Details of the model reported in section 3.1 (number of bigram types as a function of predictor variables in spoken register). Marginal and conditional R2 has been calculated in R with the MUMIN package (Barton 2019).

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: CommonBigramTypesNumber ~ TextVariety + Delivery + STWPM + offset(TotalBigramsInText) +      (1 | TextID)
   Data: df
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))

     AIC      BIC   logLik deviance df.resid
    1068     1085     -528     1056      119

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.12001 -0.06356  0.06704  0.15995  0.46832

Random effects:
 Groups Name        Variance Std.Dev.
 TextID (Intercept) 0.4378   0.6616
Number of obs: 125, groups:  TextID, 125

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  2.9014     0.2098  13.830  < 2e-16 ***
TextVarietySpokenNonnative   0.2900     0.2006   1.446  0.14823
TextVarietySpokenTranslated  0.7896     0.2450   3.222  0.00127 **
Deliveryread                -0.2530     0.1421  -1.781  0.07497 .
STWPM                        0.1579     0.1009   1.565  0.11763
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) TxtVSN TxtVST Dlvryr
TxtVrtySpkN -0.678
TxtVrtySpkT -0.815  0.714
Deliveryred -0.612  0.075  0.193
STWPM       -0.626  0.542  0.784  0.130
> r.squaredGLMM(Bigramsspoken)
                R2m       R2c
delta     0.1491673 0.9325990
lognormal 0.1493523 0.9337552
trigamma  0.1489759 0.9314020
```

Appendix 3

Details of the model reported in section 3.2 (number of bigram types as a function of predictor variables in written register). Marginal and conditional R2 has been calculated in R with the MUMIN package (Barton 2019).

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: CommonBigramTypesNumber ~ TextVariety + Delivery + offset(TotalBigramsInText) +      (1 | TextID)
   Data: df
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))

     AIC      BIC   logLik deviance df.resid
  1122.1   1136.3   -556.1   1112.1      120

Scaled residuals:
     Min       1Q   Median       3Q      Max
-0.94026 -0.04801  0.06200  0.15205  0.38361

Random effects:
 Groups Name        Variance Std.Dev.
 TextID (Intercept) 0.4912   0.7008
Number of obs: 125, groups:  TextID, 125

Fixed effects:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   3.47663    0.17108  20.322  < 2e-16 ***
TextVarietyWrittenNonnative   0.08581    0.17705   0.485  0.62790
TextVarietyWrittenTranslated  0.32417    0.15905   2.038  0.04154 *
Deliveryread                 -0.43241    0.14711  -2.939  0.00329 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) TxtVWN TxtVWT
TxtVrtyWrtN -0.518
TxtVrtyWrtT -0.673  0.554
Deliveryred -0.684  0.005  0.147
> r.squaredGLMM(BigramsWritten)
                R2m       R2c
delta     0.1154960 0.9455923
lognormal 0.1155951 0.9464038
trigamma  0.1153939 0.9447561
```

BIONOTE

MARTA KAJZER-WIETRZNY is an Assistant Professor in the Department of Translation Studies at the Faculty of English, Adam Mickiewicz University in Poznań. Following her PhD dissertation on *Interpreting universals and interpreting style* (2012) she continues with empirical investigations of interpreted, translated and non-native language use, e.g. within the recent TRINFO project carried out in part during an over year-long research stay at the University of Bologna. At times she attempts to combine corpus methods with translation process research such as key-logging and eye-tracking, in

particular while looking into the traits and the process of inter- and intralingual translation.

KURZBIOGRAFIE

MARTA KAJZER-WIETRZNY ist Assistenzprofessorin am Institut für Übersetzungswissenschaft der Fakultät für Englisch der Adam-Mickiewicz-Universität in Posen. Nach ihrer Dissertation über die Universalien und den Stil des Dolmetschens (2012) setzt sie ihre empirischen Forschungen des gedolmetchten, übersetzten und nicht-muttersprachlichen Sprachgebrauchs fort, z.B. im Rahmen des TRINFO-Projekts, das teilweise während eines über einjährigen Forschungsaufenthaltes an der Universität Bologna durchgeführt wurde. Zuweilen versucht sie, Korpusmethoden mit Übersetzungsprozessforschung wie Key-Logging und Eye-Tracking zu kombinieren, insbesondere indem sie sich mit den Merkmalen und dem Prozess der inter- und intralingualen Übersetzung befasst.

ŁUKASZ GRABOWSKI is an Associate Professor at the Institute of Linguistics, University of Opole, Poland. In 2013, he was a post-doctoral research fellow at the University of Birmingham (UK). His main research interests include corpus linguistics, formulaic language and translation studies. He has published internationally in such journals as *International Journal of Corpus Linguistics, International Journal of Lexicography, Across Languages and Cultures* and *English for Specific Purposes*; he has also authored a number of chapters in edited volumes published by John Benjamins, Springer and Emerald, among others.

KURZBIOGRAFIE

ŁUKASZ GRABOWSKI ist außerordentlicher Professor am Institut für Linguistik der Universität Opole, Polen. 2013 war er Postdoktorand an der University of Birmingham (UK). Seine Forschungsschwerpunkte umfasssen Korpuslinguistik, Formelsprache und Übersetzungswissenschaft. Er hat in internationalen Zeitschriften wie *International Journal of Corpus Linguistics, International Journal of Lexicography, Across Languages and Cultures* und

English for Specific Purposes veröffentlicht. Er hat auch eine Reihe von Kapiteln in Sammelbänden verfasst, die unter anderem von John Benjamins, Springer und Emerald veröffentlicht wurden.

Recibido / Received: 31/05/2020
Aceptado / Accepted: 28/09/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.06

Para citar este artículo / To cite this article:
Hermosa-Ramírez, Irene. (2021) "The hierarchisation of operatic signs through the lens of audio description: A corpus study." In: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection. MonTI* 13, pp. 184-219.

THE HIERARCHISATION OF OPERATIC SIGNS THROUGH THE LENS OF AUDIO DESCRIPTION: A CORPUS STUDY[1]

Irene Hermosa-Ramírez
Irene.hermosa@uab.cat
Autonomous University of Barcelona

Abstract

In opera, a multimodal art form by nature, meaning is constructed by the synthesis of its musical, verbal, visual and dramatic components. Audio description (AD) is an audiovisual translation modality that provides blind and visually impaired patrons with access to the visual elements of the play. The first aim of this pilot study is to conduct a corpus analysis to define the lexico-grammatical patterns of opera AD in the Gran Teatre del Liceu in Barcelona and the Teatro Real in Madrid. The second aim is to perform a semiotic analysis to elucidate the hierarchisation of the action, the *parole* and the visual aesthetics of the production in the AD scripts. The conclusions suggest a number of linguistic and semiotic idiosyncrasies that are shared with other AD modalities, as well as some key divergences.

Keywords: Audio description; Corpus analysis; Opera; Semiotics; Accessibility.

1. This study is part of the RAD project (Researching Audio Description: Translation, Delivery and New Scenarios), reference code PGC2018-096566-B-I00 (MCIU/AEI/FEDER, UE). The author is a member of the TransMedia Catalonia research group (2017SGR113). She has been awarded a PhD grant from the Catalan Government (2019FI_B 00327).

Resumen

En la ópera, un arte multimodal por naturaleza, el significado se cimenta en la síntesis de los elementos musicales, verbales, visuales y dramáticos. La audiodescripción (AD) es una modalidad de la traducción audiovisual que proporciona el acceso a los elementos visuales para las personas con pérdida de visión. El objetivo de este estudio piloto es, en primer lugar, realizar un análisis de corpus para definir las características léxico-gramaticales de la AD operística en el Gran Teatre del Liceu de Barcelona y en el Teatro Real de Madrid. El segundo objetivo es realizar un análisis semiótico para esclarecer la jerarquización de la acción, la *parole* y la estética visual de la puesta en escena en los guiones de AD. Las conclusiones sugieren varios patrones lingüísticos y semióticos compartidos con otras modalidades de la AD, pero también algunas discrepancias.

Palabras clave: Audiodescripción; Análisis de corpus; Ópera; Semiótica; Accesibilidad.

1. Introduction

Audio description (AD), the "verbal commentary providing visual information for those unable to perceive it themselves" (Fryer 2016: 1), is an audiovisual translation modality and accessibility service that began in the context of the performing arts. For a brief historical review, AD provision was first implemented in theatre, in 1981, at the Arena Stage Theatre in Washington DC (Pfanstiehl & Pfanstiehl 1985; ITC 2000: 4). Accessibility was later on imported to Europe, first at the Robin Hood Theatre in Averham, Nottinghamshire, and then expanded into other countries and modalities such as film and television during the 1990s (Arma 2011: 43-44). As for opera AD, its beginnings date back to the early 2000s (York 2007; Cabeza-Cáceres & Matamala 2008). Since then, opera AD and opera accessibility services as a whole – including surtitles, touch tours and audio subtitles – have witnessed steady interest both in practice and research: from early descriptive and practice-based studies (Matamala 2007; York 2007; Orero & Matamala 2007; Cabeza-Cáceres 2010; Corral & Lladó 2011; Eardley-Weaver 2010) to an increasing interest in user-centred approaches and participatory accessibility (Di Giovanni 2018a, 2018b). The rationale for this

pilot study is that opera AD has not yet been studied from a corpus-based perspective. We therefore aim to fill this gap and apply the methodology offered by Corpus Linguistics to identify lexico-grammatical patterns as well as conduct a semiotic analysis.

Though recent, the application of Corpus Linguistics studies in Audiovisual Translation (AVT) research has been fruitful (cf. Baños *et al*. 2013). Corpus research in AD, however, has been scarcer. For the purposes of this study, previous AD corpora can be classified into two categories: monomodal (text-based) corpora and multimodal corpora. Text-based corpora confirm hypotheses regarding linguistic patterns by producing frequency lists and allowing for concordance analysis (Evison 2010: 122). On the other hand, multimodal corpora involve the manual or semi-automatic tagging of the different meaning-making semiotic modes (Tuominen *et al*. 2018: 9), thereby addressing the multimedia nature of audiovisual products. In both cases, previous corpus studies on AD have widely been devoted to film and, to a lesser extent, museum AD (Perego 2019; Soler Gallego 2018).

Within the monomodal category, Salway (2007) led the TIWO project, where a corpus of 91 film AD scripts in English was compiled and analysed in order to study and define the idiosyncratic language of AD. Unusual frequently occurring phrases were found regarding shifts in the characters' focus of attention, change of scene, non-verbal communication and shift in the situation (Salway 2007: 160-161). Later on, Arma (2011) used part of the TIWO corpus to study the use of adjectives in filmic AD. In other languages, Reviers (2018) compiled a corpus of 39 Dutch films and series and conducted a lexico-grammatical analysis combined with a systemic-functional approach. All previously mentioned studies confirm the common hypothesis that AD language presents distinctive lexical, grammatical and syntactical patterns when compared to reference corpora in their matching language. Hence, this paper further questions: Is the idiosyncratic language of film AD shared by opera AD?

As for multimodal corpora, the TRACCE project compiled a corpus of over 300 film AD scripts in Spanish, which was later on complemented with 50 films in English, German and French. They developed a tagging system

comprising three levels. The narratology level included tags referring to space, time and character identification and emotion. The cinematography level referred to camera language, and the grammar level encompassed semantic domains and the syntactic and discourse structure (Jiménez Hurtado 2010: 70; Jiménez Hurtado & Seibel 2012). The films were annotated with an ad-hoc software program (Tagetti), which facilitated the arduous process of manual segmentation that often hinders multimodal corpus analysis.

Lastly, Matamala (2018, 2019) led the VIW project, an open-access, comparable corpus of student and professional AD of the same short film in Catalan, Spanish and English, where two main tagging tiers were developed and applied. On the one hand, the linguistics tiers were developed to analyse part-of-speech (PoS) frequencies and semantic tagging of open-class words. The filmic and cinematic tiers, on the other hand, "were used to carry out the visual tagging taking into account relevant elements in film construction" (Matamala 2019: 527). Within them, the scene tier referred to the location and setting, and the shot tier annotated the camera movement, i.e. medium close shots, close-ups, etc. (Matamala 2019: 527). The sound tier addressed sounds categorised by speech, paralinguistic elements, music and non-diegetic sound effects, among others. The character tier identified the characters on screen and, finally, the text tier encompassed the visual-verbal elements, i.e. on-screen titles and subtitles.

Against this background, this paper presents a pilot experience for a corpus study on opera AD scripts that aims to *1)* define the lexico-grammatical features of the scripts and *2)* analyse the distribution of the operatic semiotic meanings conveyed by the AD. The article is structured as follows: In Section 2, the conceptual framework examines opera translation and the hierarchisation of the operatic signs. In Section 3, the corpus is presented, and our two-fold methodology is introduced. The most salient lexico-grammatical features of the corpus are analysed following Quantitative Corpus Linguistics. Section 4 delves into the qualitative subsection of the corpus, where the aim is to question opera AD through the lens of semiotics. In

Section 5, the tentative results of this new area are discussed and future possibilities are examined.

2. A conceptual framework: translation, accessibility and the semiotics of opera

In order to establish the conceptual framework and the purpose of this article, two major topics are introduced: opera intelligibility through translation and the hierarchisation of operatic signs. As Mateo (2012: 115) puts it:

> [M]usic, performance and verbal text all collaborate in the creation of meaning in an opera piece. Nevertheless, the powerful presence of music has traditionally framed the conception of opera as an essentially musical genre, rather than as a dramatic art and this can be observed in sung translation: the music is normally considered untouchable and becomes the code that functions as the axis for translation decisions; the verbal text is thus subordinated to the music so that the translator must adhere to the notes and the tempo of the original score.

In this regard, it is not surprising that opera translation has generated controversy and that widespread translation strategies have been heterogenous regarding the historical period, tradition by country and even audience profile (cf. Desblache 2007). The controversy surrounding opera translation can be traced back to two general positions. While logocentrism prioritises the word and "may be characterized by the aphorism, *prima le parole e poi la musica*, musicocentrism is expressed in its opposite, *prima la musica e poi le parole*" (Gorlée 1997: 237). In the context of this study – Spanish opera houses – surtitles are the current widespread solution to maintain the integrity of the original vocal music while fostering linguistic accessibility. Accessibility services are also increasingly on offer. For illustrative purposes, Table 1 presents an overview of the most common translation strategies for opera (cf. Orero & Matamala 2007).

	Translation strategy	Fundamental characteristics
Less intrusive	No translation	The strategy presumes that the audience is familiar with the work.
	Libretto translation	The strategy does not imply synchrony. According to Orero and Matamala (2007: 262-263) libretto translations can be: 1. Reader-oriented (verse or prose) translations. 2. Translations to be sung or "composed to fit the music, which is untouchable" (Gorlée 1997: 245).
	Translated chant	Currently, not a widespread strategy in Europe (with some notable exceptions, i.e. adapted versions for children and the overall programme of the English National Opera). This is, however, the preferred approach for musical theatre.
	Surtitles	A condensed translation of the libretto, often projected above the proscenium or on small purpose-built screens on the auditorium seats. Virkkunen (2004) distinguishes two approaches to opera surtitles: 1. Libretto-like surtitles, with repetitions, imitating its verbal-musical structure. 2. Surtitles with the stage interpretation as the source text, condensing the core content.
		Creative surtitles. Also referred to as "meta-theatrical" surtitles. They can be used for added meaning and a visual or provocative effect (De Frutos 2013: 127).
More accessible	Opera accessibility	1. Relaxed performances for patrons within the autistic spectrum or with other sensory disabilities. 2. Adapted surtitles, Sign Language Interpreting and haptic technology for deaf and hard-of-hearing patrons. 3. Audio introduction, AD, touch tours, audio subtitling, braille and large-print libretto for patrons with visual impairment.

Table 1. Opera translation strategies

The convergence of the semiotic signs in opera can be rendered accessible for patrons with visual impairment through (standalone) audio introduction (AI), AD, touch tours, audio subtitling or a combination of the above. Standalone AI refers to a non-intrusive approach – first introduced by York (2007) at the London opera houses during the early 2000s – consisting of a 15-minute introduction of the cast, plot, and a vivid description of the visuals in the production. The AI would be followed by shorter, subsequent introductions before every act. Currently, some opera houses such as the Royal Opera House in London make AIs available via the audio platform Soundcloud. This is the closest approach to musicocentrism and has clear advantages in its cost-effectiveness and reutilisation potential. Nevertheless, standalone AI is not common in Spanish opera houses.

As for touch tours, Eardley-Weaver (2010) defines them as visits to the opera stage and backstage where patrons with visual impairment have the opportunity to touch items of the set, props, stage design models and sometimes cast members as an introduction to the visual elements of the production. This service is not offered at the Teatro Real or the Gran Teatre del Liceu (henceforth Liceu) opera house, but it is part of the accessible programme at the Teatro de la Zarzuela in Madrid.

Lastly, audio subtitling is an accessibility modality that consists of reading the subtitles or surtitles aloud. Orero (2007) reports on an experience with audio subtitling at the Liceu opera house where audio subtitles were tested with end-users in a concert version of *Roberto Devereux*. Even though the attendants were initially reticent to the continued overlapping with the music, they showed a high degree of acceptance in the questionnaire (Orero 2007: 146-147). As opposed to standalone AI, this service could be defined as logocentric and its main asset is that it could be reused unrestrictedly, provided that the audio subtitles are synchronised with the surtitles. Despite the positive outlook on this modality, it has not been established in either of the Spanish opera houses.

Shifting the focus to opera AD specifically, Corral and Lladó (2011: 163) define this modality as a commentary "designed to facilitate a comprehensive appreciation of artistic performances which incorporates both sound and visual dimensions". Its main divergence with other AD modalities lies in the fact that three specific elements coincide in the operatic medium:

mise-en-scène, surtitles and libretto (Matamala 2005). In other words, not only do visuals need to be translated semiotically, but also (sung) dialogues. Precisely, in contrast to other AD modalities such as film, television or theatre where the overlapping of AD with dialogue or other relevant sounds is not allowed (Maszerowska *et al*. 2014), the operatic chant may be interrupted by the AD in the context of both the Liceu opera house and the Teatro Real. Matamala (2007) further outlines the four main characteristics of opera AD: the multiplicity of linguistic codes, the overlapping of the AD with the chant, a reaction capability from the audio describer to adapt to unforeseen events, and an often necessary interpretation of the events in order to achieve a concise description.

For the purposes of this study, AD is understood as the combination of AI and AD throughout the performance, "intertwined in the gaps where lyrics are not generally heard and music is considered to be less important" (Orero & Matamala 2007: 270). The audio "through" description segment not only includes the translation of the visual elements of the performance, but also a synthetic translation of the surtitles. Currently, this is the most widespread approach among European opera houses and opera AD service providers: Accès Culture in France, VocalEyes in the UK, Aristia, Aptent and the TransMedia Catalonia research group in Spain (Cabeza-Cáceres & Matamala 2008; Orero *et al*. 2019) and the Macerata Opera Festival in Italy (Di Giovanni 2018a). The combination of AI and AD would fit somewhere between logocentrism and musicocentrism.

The second topic to be addressed in our conceptual framework is the hierarchisation of the operatic semiotic signs. Generally, semiotics in the field of opera has taken an interest in the classic dispute between music and *parole*, as well the hierarchisation of operatic signs. A first example is the question of whether music alone is able to convey meaning, which has been disputed extensively in diverse disciplines such as Musicology, Narratology and Translation Studies. In tackling the meaning-making potential of instrumental vs. vocal music, Tråvén (2005: 103-104) argues that instrumental music will always be more open to interpretation, as opposed to vocal music, which clearly conveys a message or an emotion:

> Music rhetoric lacked the precise quality or communicative skill of the spoken language, and since vocal music was considered to affect both the

> mind and the heart, as opposed to instrumental music that "spoke" more to the heart, it consequently had a higher standing than purely instrumental music.

As Ryan puts it (2004) music is an art made of signifiers without signifieds. The author outlines three basic conditions of narrativity (2004: 8-9):

1. A narrative text must create a world and populate it with characters and objects.
2. The narrative world must undergo changes of state that are caused by physical events: either accidents or deliberate human action.
3. The text must allow for the reconstruction of an interpretive network of goals, plans, causal relations and psychological motivations around the narrated events.

These three conditions apply to opera provided that some translation strategy is used and, in our case, sensory accessibility services are offered.

All in all, instrumental music conveys emotions, while (intelligible) vocal music is able to communicate an articulate message. Even though music is seen as the defining element of opera, current public expectations, undoubtedly in the Spanish context, require accessing both the signifiers and the signifieds. This is no different for blind and visually impaired patrons.

Shifting the focus to the hierarchisation of signs, multimodal studies have seen an interest in opera – the *Gesamtkunstwerk* – as a paradigm of multimodality itself, as it welds together literary text, dramatic (staged) action and music. These three semiotic resources are generally the basis for the semiotic taxonomies put forward by multimodality scholars. Hutcheon and Hutcheon (2010: 65), for instance, draw from Kress and van Leeuwen's distinction between *production* media "(such as the singers' voices, gestures, motions; the orchestra's musical sounds; the stage action and sets, etc.)" and *design* modes "(the musical score and libretto, the director's interpretive plan, the various designers' and performers' visions, etc.)" as a definition of social semiosis in opera. Furthermore, Rossi and Sindoni (2017) work from a Systemic Functional Linguistics approach and propose three semantic levels within an operatic play, in descending order: semiotic systems (language, music and mise-en-scène), semiotic resources (libretto, score, performance

and staging) and semiotic components (stage direction, kinesics, props, etc.). We will refer back to the semiotics of opera and opera AD in Section 4.

3. The pilot study: a comparable corpus of opera audio description

The following analysis aims to explore corpus combination in a small-sized corpus. For methodological clarity, it is important to make a distinction between corpus triangulation and corpus combination. The defining characteristic of corpus triangulation is the explicit integration of methods, subjects or materials from the outset of the study (Hansen 2010; Malamatidou 2018: 8). Corpus combination, on the other hand, "mixes" research methods, usually taking the form of a quantitative analysis followed by a qualitative analysis. In this paper, we cannot claim corpus triangulation, as different aims are established for the linguistic section, i.e. researching patterns in the language of operatic ADM; and the semiotic section, i.e. analysing how operatic signs are hierarchised in the AD scripts. Even though synergies are presented, they are not sufficient to claim a methodology based on triangulation, which is often the case in Translation Studies corpora.

Besides, the rationale for limiting the size of the corpus is precisely its qualitative analysis and manual semiotic tagging process, which is described in more detail in Section 4. In this regard, authors such as Koester (2010) and Evison (2010: 123) have supported the adequacy of smaller corpora in highly specialised registers. As Malamatidou (2018: 53) highlights, the possible limitations of smaller corpora are counterbalanced by the "depth and breadth of [the] understanding" they allow for. Koester (2010: 67) goes on to argue that researchers can gain greater familiarity with smaller corpora, as they are able to reflect on contextual features and link them to the linguistic patterning. Given that this paper deals with a niche modality, it was considered adequate to compile a 33,999-word pilot corpus for both quantitative and qualitative purposes.

The innovation of this corpus is twofold. Firstly, for the very modality of AD it deals with. To the best of our knowledge, there have been no previous corpus studies on opera AD. This may be linked to the more limited availability of the scripts – as opposed to film productions (Salway 2007; Reviers 2018) or museum AD (Soler Gallego 2018; Perego 2019) – and an

inherent issue with the performative, live nature of AD specifically in the case of the Liceu opera house (Cabeza-Cáceres 2010). Indeed, while some audio describers may read through the scripts with minimal modifications to adapt it to the pace, others will be more inclined to improvise. Therefore, the AD script may also need to be adapted to substantial changes in the stage performance (Orero & Matamala 2007: 271). In our case, given the fact that the recordings of the live AD have not been kept, we can draw an analogy with pre-production scripts. The semi-live approach at the Teatro Real, on the other hand, implies that the scripts compiled for this corpus accurately correspond to the broadcasted AD. In Section 3.1 we further illustrate the formal divergences between the approaches of the Liceu and the Teatro Real.

The second innovation of this study lies in its dual aim: the corpus is used to extract linguistic information from the scripts, as well as to analyse (qualitatively) operatic semiotic information. In the following subsection, the corpus description, design and operating software are presented (Section 3.1). Section 3.2 offers a quantitative lexico-grammatical analysis of the language of opera AD. The semiotic hierarchisation in operatic AD is analysed separately in Section 4, which is followed by a discussion of the results in Section 5.

3.1. Selection of the sample

As outlined in Section 2, combining AI and AD is common practice in both opera houses. Yet, from a technical and formal perspective, different approaches are applied at the Teatro Real in Madrid and at the Liceu in Barcelona. Aristia, the company behind the opera AD in the Teatro Real, follows a pre-recorded strategy. In other words, the AD is voiced and divided into audio fragments that are manually launched by a technician the day of the show. At the Liceu opera house, on the other hand, AD is delivered live. In addition, the AD in the Teatro Real can be accessed via the application Teatro Real Accesible. Once the blind and visually impaired patrons scan the QR (Quick Response) code at the entrance of the opera house, the application is ready to launch the pre-recorded AI and AD, which can be accessed from any of the theatre seats. Meanwhile, at the Liceu, a number of seats

and screens are equipped with headphone plugs where patrons can connect their own headsets and listen to the live AD.

Formally speaking, there is another difference between both opera houses: the pre-recorded AD from the Teatro Real distinguishes two different voices according to their function: one of them provides a summarised account of the surtitles and the other one describes the visual content of the play. Conversely, the live AD at Liceu opera house integrates both the surtitles and the descriptions of the visual content: scenography, the characters and their actions (Cabeza-Cáceres & Matamala 2008). In the written scripts that comprise this corpus, the Catalan AD thus interweaves descriptions with synthesised surtitles, while the Spanish AD differentiates the two voices by highlighting the content of the surtitles in bold.

The corpus consists of three AD scripts in Catalan that were delivered at the Liceu by three different members of the TransMedia Catalonia research group and three AD scripts in Spanish, delivered at the Teatro Real and produced by the same audio describer from the AD service provider Aristia. Three operas were chosen under the premise that they were described both in Catalan and Spanish: *Aida*, *The Magic Flute* and *Carmen*. While the ADs for *The Magic Flute* and *Carmen* correspond to the same creative productions, *Aida* was described for two different productions. Having obtained the text documents, they were arranged in four subcorpora, on the basis of language and distinguishing AI from AD, as illustrated in Table 2.

Opera	**Word count**		**Year of production**	
	Liceu	Teatro Real	Liceu	Teatro Real
Aida	AI: 1,637	AI: 768	2007	2018
	AD: 5,039	AD: 2,929		
The Magic Flute	AI: 859	AI: 1,262	2016	2016
	AD: 5,682	AD: 5,780		
Carmen	AI: 1,415	AI: 501	2011	2017
	AD: 4,099	AD: 4,028		
	Total words: AI: 3,911 (CAT), 2,531 (ES)			
	AD: 14,820 (CAT), 12,737 (ES)			

Table 2. Corpus description

As an introduction to the descriptive quantitative analysis that is central to this section, the number of total words accounts for 3,911 in Catalan and 2,531 in Spanish for the AI portion; and 14,820 words in Catalan and 12,737 words in Spanish for the description throughout the performance. The mean number of words per subcorpora (*AI_ES*=844; *AI_CAT*=1,304; *AD_ES*=4,246; *AD_CAT*=4,940) indicates lengthier descriptions in the context of the Liceu opera house. In this line, the standard deviation (*sd*) – a useful tool for calculating the variability of the length of text collections – accounts for 1,106.88 in the AD subcorpora, while the standard deviation for AI is considerably smaller (*sd*=432.85). In all, this preliminary set of descriptive statistics suggest that the Liceu ADs and AIs favour longer, more detailed descriptions.

3.2. Lexico-grammatical analysis of the corpus

As of the linguistic analysis of the study, a descriptive explanatory approach was adopted (Saldanha & O'Brien 2013: 50). In order to define the salient features of opera AD scripts, a lexico-grammatical analysis of the corpus was performed on the basis of four formal parameters: mean sentence length, open-class word frequencies, PoS distribution and type-token ratio (TTR).

The primary software employed to perform the linguistic analysis was SketchEngine, which allows for frequency list generation as well as automatic lemma annotation. As SketchEngine gathers general language corpora retrieved from the web, it also allows the researcher to compare their own sets of corpora to general language samples. The online tool performs concordance analysis too, which proved useful to disambiguate homographs.

3.2.1. Mean sentence length

The first parameter in our descriptive analysis was mean sentence length in all four subcorpora. For the Teatro Real AI subcorpus, the mean sentence length was 19.32 words, while the Liceu AI mean sentence length accounted for 21.85 words. As for the mean sentence length in the Teatro Real AD, the subcorpus displayed a mean sentence length of 6.80 words, while the same AD subcorpus for the Liceu accounted for 13.71 words. As such, we suggest that the higher percentage of words per sentence in the Catalan subcorpora is partly linked to the intermingling of surtitles and visual

descriptions. Conversely, in the Spanish AD subcorpus there was a clear separation between surtitles and descriptions of visual content: surtitles were written in bold in the text and distinguished by two different voices in the recording, and therefore surtitles were not always explicitly introduced. Moreover, the Spanish AD subcorpus was rich in short sentences such as: *Se tambalea* ("He wobbles"), *Saca su espada* ("He unsheathes his sword") or *Repite* ("She repeats").

For comparative purposes with general language corpora, the Catalan Web 2014 corpus, that is, the Catalan reference corpus integrated in SketchEngine, scores 20.99 words per sentence, in line with the European Spanish Web 2011, which shows 21.86 words per sentence. These scores fall closer to the AI subcorpora (*CAT*=21.85; *ES*=19.32) and, to a lesser extent, to the Catalan AD subcorpus, which features a mean of 13.71 words. Interestingly, the Spanish AD subcorpus (mean sentence length=6.80) seems to resemble the results from the VIW filmic corpus – in English, Spanish and Catalan – where the mean sentence length for professional and student AD was of 8.4 (Matamala 2018: 191). Overall, the Teatro Real AD subcorpus was richer in phrases, while the Liceu ADs included more compound and complex sentences.

3.2.2. Open-class word frequencies

Moving on to the lexical analysis of the corpus, Table 3 introduces the most frequent open class-words both in the AI and AD sections. For all four subcorpora, proper nouns (Aida, Carmen, Pamina, Sarastro, Papageno, Tamino) and common nouns linked to character identification (choir, queen, lady, soldier, woman and man), auxiliary verbs and copulae, as well as verbs of movement (to arrive, to go, to disappear, to enter, to stay) and verbs of communication (to say, to ask) comprised the most salient open-class words overall. These results are generally in line with the findings from previous corpus AD studies in the filmic modality (Salway 2007: 156; Matamala 2018: 196; Reviers 2018: 192). Yet, some elements are missing, namely the fact that there is no reference to body parts, objects or verbs linked to character description among the twenty most frequent open-class words in the four subcorpora.

Teatro Real AI_ES	Liceu AI_CAT	Teatro Real AD_ES	Liceu AD_ES
Tamino 28	Carmen 30	*ser* (to be) 142	*ser* (to be) 148
Papageno 19	*haver* (auxiliary verb) 27	*decir* (to say) 109	*haver* (auxiliary verb) 129
Pamina 17	*anar* (to go) 26	*ir* (to go) 106	*fer* (to do) 105
teatro (theatre) 16	José 18	*estar* (to be) 88	Radamès 97
haber (auxiliary verb) 14	*gran* (big, great) 18	*mujer* (woman) 88	Papageno 87
Real (real) 14	*acte* (act) 16	Tamino 83	Carmen 84
dirección (direction) 12	Egipci (Egyptian) 15	*haber* (auxiliary verb) 74	Tamino 84
Sarastro 11	Aida 15	*hombre* (man) 70	Aida 75
minuto (minute) 11	Tamino 14	Papageno 69	Pamina 71
amor (love) 11	*estar* (to be) 14	*aparecer* (to appear) 61	*aparèixer* (to appear) 68
mágico (magical) 10	Radamès 14	*desaparecer* (to disappear) 59	Jose 67
coro (choir) 10	*color* 13	*pedir* (to ask) 57	*entrar* (to enter) 65
ópera (opera) 9	*fer* (to do) 13	*entrar* (to enter) 53	*escenari* (stage) 62
escena (scene) 9	*soldat* (soldier) 12	Pamina 51	Amneris 60
prueba (trial) 8	Papageno 12	*suelo* (floor) 49	*dir* (to say) 59
reina (queen) 8	*arribar* (to arrive) 12	*centro* (centre) 48	*llum* (light) 54
noche (night) 8	*veure* (to see) 12	*llevar* (to take) 47	*blanc* (white) 53
parte (part) 8	Sarastro 11	*amor* (love) 46	*escena* (scene) 52
producción (production) 8	*òpera* (opera) 11	*quedar* (to stay) 44	*anar* (to go) 52
dama (lady) 7	Pamina 10	Carmen 42	*veure* (to see) 51

Table 3. Twenty most frequent open-class words in AI and AD

As introduced in Section 2, the AI segment offers a synthesised historical account of the composer and the play, a summary of the plot and the main aspects of the scenography and costumes (Cabeza-Cáceres & Matamala 2008: 99). Here, the prominent use of nouns referring to the internal structure of

the piece, i.e. opera, scene, part or act, further highlights a difference in purpose from the AD segment. Hereafter, the focus will be on the two AD subcorpora, as our aim is to determine the hierarchisation of the operatic signs when time constraints are imposed as the play moves forward.

Still considering open-class word frequencies, we now move onto the frequencies per open-class word. As illustrated in Table 4, shared units make up 60% of the twenty most frequent nouns in both the Catalan and the Spanish subcorpora. Character identification – both with proper and common nouns – proved to be particularly salient both in the Spanish and Catalan corpus. Other highly frequent words belonged to the location and body part semantic classes. This is in line with the findings of previous AD corpora on film (Salway 2007; Matamala 2018; Reviers 2018), although they also highlight objects, which, again, seem to be less frequent in the present corpus.

Nouns		**Adjectives**	
AD_ES	AD_CAT	AD_ES	AD_CAT
mujer (woman)	Radamès (proper name)	*oficial* (official)	*seu* (possessive)
Tamino (proper name)	Papageno (proper name)	*negro* (black)	*blanc* (white)
Papageno (proper name)	Carmen (proper name)	*oscuro* (dark)	*gran* (great, big)
hombre (man)	Tamino (proper name)	*rojo* (red)	*fosc* (dark)
Pamina (proper name)	Pamina (proper name)	*blanco* (white)	*negre* (black)
suelo (floor)	José (proper name)	*lleno* (full)	*meu* (possessive)
centro (centre)	*scenario* (stage)	*mecánico* (mechanical)	*altre* (other)
amor (love)	Amneris (proper name)	*nuevo* (new)	*summe* (high)
Carmen (proper name)	*llum* (light)	*grande* (big)	*primer* (first)
derecha (right)	*escena* (scene)	*derecho* (right)	*mateix* (same)

Amneris (proper name)	*mà* (hand)	*eterno* (eternal)	*teu* (possessive)
Radamés (proper name)	Sarastro (proper name)	*divino* (divine)	*segon* (second)
izquierda (left)	*sacerdot* (priest)	*egipcio* (Egyptian)	*vermell* (red)
círculo (circle)	*amor* (love)	*pequeño* (small)	*nostre* (possessive)
José (proper name)	*home* (man)	*feliz* (happy)	*nou* (new)
pared (wall)	*noia* (girl)	*culpable* (guilty)	*superior* (superior)
soldado (soldier)	*terra* (land, floor)	*izquierdo* (left)	*blau* (blue)
niño (child)	*cor* (choir, heart)	*dulce* (sweet)	*egipci* (Egyptian)
Sarastro (proper name)	*esquerra* (left)	*hermoso* (beautiful)	*sol* (alone)
mano (hand)	*dona* (woman)	*alto* (tall)	*ple* (full)

Table 4. Twenty most frequent nouns and adjectives in the AD subcorpora

Regarding the adjective PoS, 40% of the twenty most frequent adjectives were shared by both the Teatro Real and the Liceu ADs. Overlapping adjectives corresponded to the following semantic sub-groups: colour (black, red, white), light (dark), size (big), origin (Egyptian), attribute (new), and value (full).

Verbs		**Adverbs**	
AD_ES	AD_CAT	AD_ES	AD_CAT
ser (to be)	*ser* (to be)	*no* (no)	*no* (no)
decir (to say)	*haver* (auxiliary)	*arriba* (above)	*más* (more)
ir (to go)	*fer* (to do)	*abajo* (below)	*davant* (in front)

estar (to be)	*aidar* (to help)	*luego* (then)	*tot* (all)
haber (auxiliary verb)	*aparèixer* (to appear)	*más* (more)	*després* (after)
aparecer (to appear)	*entrar* (to enter)	*pronto* (soon)	*darrere* (behind)
desaparecer (to desappear)	*dir* (to say)	*ahora* (now)	*molt* (very)
pedir (to ask)	*anar* (to go)	*siempre* (always)	*ara* (now)
entrar (to enter)	*veure* (to see)	*dentro* (inside)	*només* (just)
llevar (to take/to wear)	*sortir* (to exit)	*cerca* (near)	*abans* (before)
quedar (to stay)	*estar* (to be)	*sólo* (just)	*ja* (already)
querer (to want/to love)	*quedar* (to stay)	*ya* (already)	*mai* (never)
poder (can)	*tornar* (to turn, to go back)	*delante* (in front)	*encara* (yet)
llegar (to arrive)	*acabar* (to finish)	*tan* (such)	*així* (so, thus)
dar (to give)	*tenir* (to have)	*menos* (less)	*sempre* (always)
acercar (to approach)	*estimar* (to love)	*también* (too)	*aviat* (soon)
volver (to come back)	*trobar* (to find)	*aun* (even)	*dalt* (above)
tener (to have)	*poder* (can)	*jamás* (never)	*mentrestant* (meanwhile)
ver (to see)	*donar* (to give)	*despacio* (slowly)	*finalment* (finally)
volar (to fly)	*sentir* (to feel)	-	*lentalment* (slowly)

Table 5. Twenty most frequent verbs and adverbs in the AD subcorpora

As for verbs, shared units made up 60% of the twenty most frequent word list (see Table 5). Following Faber and Mairal Usón's classification (1999) for lexical domains, overlapping units belonged to the following categories: existence (to be, to appear), action (to give, can), perception (to see), sentiment (to love), position (to stay), possession (to have), speech (to say) and movement (to go, to enter).

Lastly, the variability in adverbs proved to be limited in our study: the twenty most frequent adverbs practically made up the entire range of this PoS. Furthermore, the top four adverbs in the Spanish subcorpus (no, above, below and then) made up for half of the occurrences, as the top five adverbs do in the Catalan subcorpus (no, more, in front, all, after).

3.2.3. PoS distribution

Having established the most frequent open-class words, we now move on to the overall PoS distribution of the AD subcorpora. For the purpose of this study, the Catalan and the Spanish FreeLing PoS tagsets were applied. These tagsets are integrated in SketchEngine and are used to lemmatise and tag the PoS. Table 6 indicates the relative frequency results for each PoS as per 1000 words. Words tagged as numerals or "unknown", as well as punctuation, were excluded.

Overall lexicon size	AD_ES	AD_CAT
Adjective	2.67	5.48
Adverb	3.45	3.80
Conjunction	8.36	9.13
Noun	26.00	27.34
Preposition	19.72	21,24
Pronoun	14.96	11.62
Verb	24.84	21.38

Table 6. Relative frequency of words

When analysing the most frequent PoS, nouns (*ES*=26.00; *CAT*=27.34) made up the most frequent category, followed by verbs (*ES*=24.84; *CAT*=21.38) and prepositions (*ES*=19.72; *CAT*=21.24). Interestingly, the greatest disparity

between our two subcorpora was a deviation of 2.81 for adjectives. In fact, adjectives made up the less salient word class in the Teatro Real subcorpus, while adverbs ranked last for the Liceu AD subcorpus.

Furthermore, when compared to other AD corpora, our results resemble those from the VIW corpus (Matamala 2018: 193) and from the filmic AD corpus in Dutch (Reviers 2018), with nouns and verbs ranking first and adjectives and adverbs scoring the lowest frequencies.

3.3.4. Type-token ratio

The fourth and last parameter in the linguistic analysis is devoted to TTR, which is an established indicator of lexical variation in corpus analysis (Baker 1995). TTR is calculated by dividing the number of unique words (types) by the total running words (tokens) in a corpus. A low TTR thus implies greater repetition, that is, less lexical variation. Typically, spoken genres and L2 learner corpora (Durán *et al*. 2004) present a lower TTR. The reliability of this measurement has nonetheless been challenged, as the ratio will be largely influenced by the text length – the shorter the text, the bigger the ratio (McCarthy & Jarvis 2010: 382), and, in the case of translation corpora, it may be an insufficient measurement due to the differences in morphology between languages (Corpas-Pastor 2008). Standardised TTR (calculated on the basis of 1000 words) is a measurement that can mitigate the influence of the text length.

As for the results in our corpus, the Liceu AD subcorpus scored 20.53% TTR, with an overall lexicon size of 3,043 words. Regarding the Teatro Real AD subcorpus, the TTR amounted to 22.31%, with a lexicon size of 2,841 words. Both subcorpora therefore showed a low degree of lexical variation.

If we look at the standardised TTR for each individual AD script (Table 7), however, we find higher lexical variation. This suggests that a large portion of the lexicon is shared by the scripts; that is, the same words are repeated not so much within the same script, but throughout the ADs from the same genre.

	Teatro Real AD_ES			Liceu AD_CAT		
	Aida_ES	Carmen_ ES	Magic flute_ES	Aida_ CAT	Carmen_ CAT	Magic flute_ CAT
Standardised TTR	38.40%	37.05%	36.62%	39.94%	39.20%	38.36%
Overall standardised TTR	37.10%			39.16%		

Table 7. Standardised TTR

On a final note, for our opera TTR and standardised TTR to be compared with the results from other AD corpora, we first need to distinguish between the studies that apply each of the two measurements. Arma's (2011) study on filmic AD reports 26.0% TTR for English and 31.5% for Italian AD. On the other hand, Perego's (2019) study comprising 18 standalone ADs from the British Museum scores 51.07% TTR, a much higher ratio. As for standardised TTR, Revier's (2018) study on Dutch filmic AD indicates 38%, and Soler Gallego's (2018) corpus on museum AD in English scores a median standardised TTR of 42.5%. We will expand on the comparability of the results in the discussion section.

4. A semiotic analysis of the corpus

The second aim of the study was to analyse and hierarchise the semiotic information in our AD corpus. Specifically, we raise the following questions: How are the aesthetics of opera reflected in opera AD scripts? Can we extract any information at all about the hierarchisation of the signs in contemporary opera productions? In an attempt to answer these questions, a semiotic tagging system was adopted and applied to the corpus.

4.1. A conceptual framework for the semiotic tagging of opera AD

The polysemiotic nature of audiovisual texts has been central in AVT studies. For instance, Delabastita (1989) synthesized the simultaneous signifying codes in audiovisual products as follows: acoustic verbal signs (dialogue,

dubbing, lyrics), visual verbal signs (subtitles, text on screen), acoustic non-verbal signs (music, diegetic and non-diegetic sounds) and visual non-verbal signs (images, camera language). Another relevant contribution is Gottlieb's (2005: 54-55) semiotic taxonomy of translation, where audio description falls within the umbrella of adaptational, intersemiotic verbalised translation.

In this study, we suggest a semiotic tagset and apply it to a textual corpus in an attempt to find out how operatic semiosis translates into the Catalan and Spanish operatic AD. In order to set a comprehensive semiotic tagset, we depart from the three pillars of opera, namely "*le parole*", "*la musica*" and "*le immagini*", from which Rędzioch-Korkuz (2016: 42-43) proposes a semiotic division of elements that make up the operatic macrocode:

(1) LANGUAGE
- verbal signs
- paralinguistic and prosodic codes

(2) MUSIC and SOUNDS
- instrumental music
- vocal music
- sound effects
- noises
- non-verbal reactions of the audience

(3) THEATRICAL FORMS
- kinesics
- proxemics
- acting
- scenography
- dancing
- theatre architecture

Departing from Rędzioch-Korkuz's proposal (2016) and drawing from the TRACCE narratology tagset (Jiménez Hurtado 2010: 71), a character identification tag was added, as well as two semiotic components within the scenography category: lighting and props. Furthermore, an additional tag was created to encompass AD inserts.

Semiotic Tagset			
	Language	(Synthesised) surtitles	
	Music and sounds	Instrumental music Vocal music Noise	
	Theatrical forms	Character Kinesics Proxemics Acting Scenography Wardrobe Theatre architecture Dancing	Lighting Props
	AD insert	(notes from the describer, references to the AD structure, mentioning of elements outside the stage space)	

Table 8. Conceptual map for the semiotic tagset. Adapted from Rędzioch-Korkuz (2016) and Jiménez Hurtado (2010)

Table 8 shows the semiotic tagset applied in the analysis. The Language category corresponds to the translated chant. Describers for the Teatro Real and for the Liceu are provided with a list of the surtitles as part of their working material. Surtitles are therefore included in the AD scripts and they are either reformulated and synthesised or directly read aloud. At the Liceu, surtitles are generally synthesised and grouped into longer descriptive units, while the Teatro Real approach favours a word-by-word reading of the surtitles, which are clearly separated from the descriptive content. The second category, music and sounds, shows allusions to instrumental music (*a ritme de flauta i oboe*: to the beat of the flute and the oboe; *acompanyats del so de tompetes*: accompanied by the sound of trumpets), vocal music (*cantan loas al vencedor*: they sing the praises of the victor; *canten el cèlebre «trio de les cartes»*: they sing the notorious Card Trio), and noises (*en sentir sorolls*: hearing noises; *sent un gemec*: [he] hears a moan). The theatrical forms, our third category, includes the majority of tags. Among them, kinesics, proxemics and acting are perhaps the closest signs, sometimes displayed

interchangeably or simultaneously (Bobes Naves 2004). As such, it may be difficult to tell them apart. For the sake of clarity, kinesics refers to gestures, mimicry and facial expressions[2]; proxemics refers to the stage position of the performers, the distance among them and the distance between performers and the audience. As for acting, the tag refers to the performers' movements, actions and behaviour (*cae al suelo agotado*: exhausted, he falls to the ground; *torna a sanglotar*: he starts to sob again).

Once the tags were established and tested for ambiguity in a separate AD script, the manual tagging process was performed using the qualitative analysis software Atlas.ti. Among other uses, this software allows for the creation of tags or "codes" that can be arranged in interlinked networks. The texts were thus annotated, with words acting as the minimal sense unit and sentences as the maximal sense unit, provided that the sentence concerned only one of the semiotic tags. Here, an example is presented for both cases:

(1) [character] Micaela i José [end/character] [vocal music] acaben la cançó [end/vocal music] [proxemics] ben abraçats [end/proxemics] [scenography] al costat de la cabina de telèfon [end/scenography].

[character] Micaela and José [end/character] [vocal music] finish their song [end/vocal music] [proxemics] hugging tightly [end/proxemics] [scenography] next to the phone booth [end/scenography].

(2) [scenography] A la izquierda hay un árbol con una horca y a la derecha una bomba [end/scenography].

[scenography] To the left there is a gallows tree and to the right there is a bomb [end/scenography].

Quantifying the frequencies for each tagset allows for the identification of the most salient semiotic resources in the AD scripts. As shown in Figure 1, the most salient semiotic tagset was undoubtedly character identification (36.16%). This is no surprise if we refer back to the most frequent nouns in Table 4. Even more so than in other modalities, such as filmic AD, operatic characters need to be identified not only when they are firstly introduced or

2. Rędzioch-Korkuz (2016: 46-47) argues that kinesics may be peripheral in the operatic context, as the singing interferes with the performers' ability to fully control their face muscles.

when they enter, leave and move around the stage, but also when surtitles are introduced, as the overlapping in singing makes character identification more confusing and the entire cast can be on stage simultaneously. Yet another factor may be that recognising an actors' voice in one's own language is easier than recognising a singing voice in another language. For comparative purposes, the character introduction tag within the narratology tagset in TRACCE's filmic corpus presented a very low frequency (120 occurrences, as opposed to the most frequent tags: Action=7,985, and Indoor location=571 (Jiménez Hurtado & Soler Gallego 2013: 584).

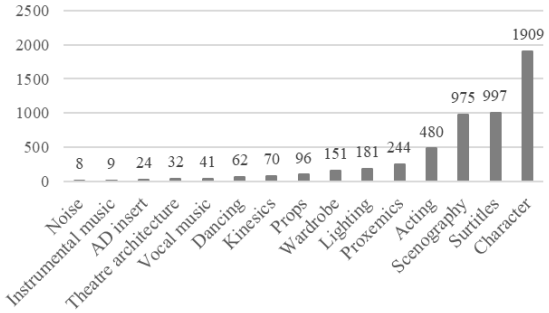

Figure 1. Frequency analysis of the semiotic tagset

Surtitles accounted for 18.89% of the total semiotic occurrences (*N*=5279). In third place, the scenography tagset was identified (18.47%), encompassing a large number of visual meaning-making signs, namely the scene – for instance, a gyratory stage, and the characters' interaction with the stage and its inanimate elements. When it comes to the performers' movement and physical expression of emotions, acting (9.09%) encompasses most of the meaning-making occurrences while proxemics (4.62%) and kinesics (1.33%) appear to be less prominent.

Finally, we present the semiotic similarities and differences between the two AD subcorpora (Figures 2 and 3), which are expanded on in the discussion section.

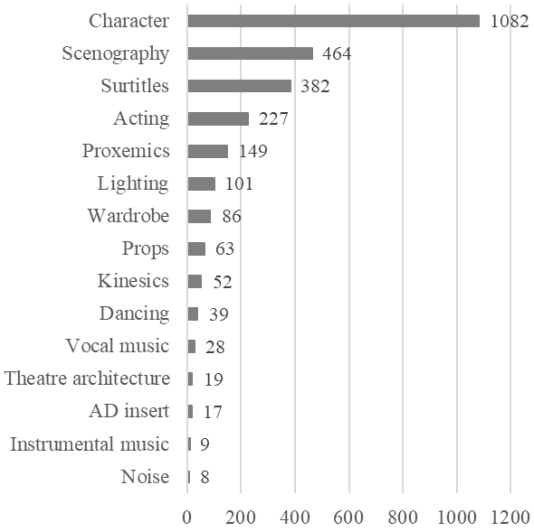

Figure 2. Semiotic distribution of the Liceu AD

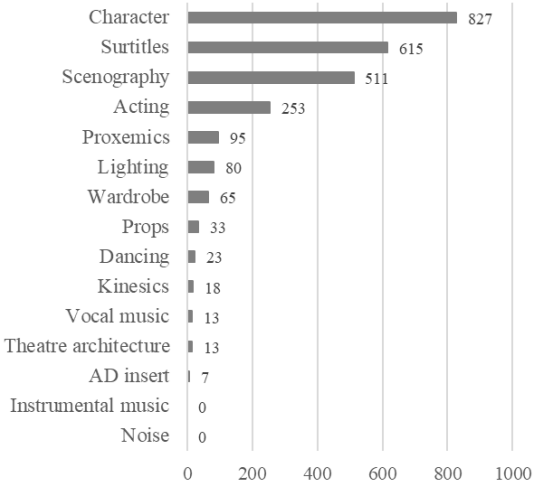

Figure 3. Semiotic distribution of the Teatro Real AD

5. Discussion

In this section, the results from the pilot study are discussed with the aim of putting forward a tentative set of lexico-grammatical and semiotic characteristics of opera AD in the Spanish context. Limitations from the study are addressed later on. Overall, in spite of the formal differences between the Teatro Real and the Liceu AD, both subcorpora share a great number of lexico-grammatical features, i.e. overlapping of the most frequent nouns and verbs, similar PoS distribution and resembling TTR and standardised TTR. Yet, there are some significant discrepancies, i.e. mean sentence length: (*AD_CAT*=13.71; *AD_ES*=6.80).

When compared to other AD modalities, namely film, our opera subcorpora also reveal certain discrepancies. In terms of lexico-grammatical patterns, the mean sentence length for the AI (*CAT*=21.85; *ES*=19.32) resembles that of the general language corpora for Catalan and Spanish. Soler Gallego's (2018: 236) corpus on museum AD in English shows similar results, with 19 words per sentence. The Teatro Real AD mean sentence length (*ES*=6.80), however, is closer to the results from the VIW corpus, with 8.4 words per sentence (Matamala 2018: 191); while the mean sentence length from the Liceu AD subcorpus (13.71 words per sentence) falls closer to Reviers' (2017: 84) corpus on filmic AD in Dutch (14.27). Nonetheless, these resemblances are not sufficient to establish any tangible comparisons, as they leave out language-specific morphology and syntactic elements (Saldanha & O'Brien 2013: 88). A certain trend has been spotted in the Liceu subcorpora consisting of complex and compound sentences, as opposed to the higher prevalence of phrases and independent clauses in the Teatro Real AD subcorpus.

Regarding open-class word frequencies: a high percentage of the most frequent nouns, adjectives and verbs are shared between our subcorpora. Conversely, some idiosyncratic lexico-grammatical features shared in filmic AD corpora (Salway 2007; Matamala 2018; Reviers 2018) are not particularly salient in our operatic corpus, e.g. verbs linked to character description, and nouns indicating objects and indoor and outdoor locations. These locations are replaced by more scenic and less cinematographic positions, such as *left*, *right*, *stage*, *wall*, *scene* and *floor*.

As for the PoS distribution, even though adjectives represent less than 6% of the overall word count in both AD subcorpora, they are more prevalent in the Liceu scripts. In this sense, adjectival phrases are thought to contribute to the lengthier sentences in the Catalan corpus.

When discussing lexical variation, TTR results from the opera pilot corpus (*ES*=22.31%; *CAT*=20.53%) differ greatly from the standardised TTR (*ES*=37.10%; *CAT*=39.16%). As we seek to compare our results with those in previous AD corpora, we face the challenge of non-uniformity: some studies apply the TTR measurement (Arma 2011; Perego 2019), while others report their findings with standardised TTR (Reviers 2018; Soler Gallego 2018). In any case, excessive lexical variation is generally avoided in AD, as the aim is to foster access, irrespective of the patrons' "language competence, background preparation and individual attention patterns" (Hutchinson & Eardley 2018: 8–9, in Perego 2019: 338). What we can gather is that both TTR and standardised TTR results for AD rarely surpass a 50% ratio.

Regarding the semiotic dimension of the analysis, both subcorpora are surprisingly similar in terms of sign saliency. Given that the sign frequencies were not normally distributed, we conducted a Mann-Whitney test and found no statistical differences between the Teatro Real and the Liceu corpus (U=215, p=0.48). Nevertheless, as this is a pilot study and our aim is to incorporate more data to the sample, for now we stick to general frequency results.

Overall, character identification is the most relevant tag at both the Liceu and the Teatro Real. This is closely linked to the added difficulty in recognising singing voices, particularly in ensembles where the singers intrude on each other's chant. Read-aloud surtitles also play a salient role in the Teatro Real subcorpus, more so than in the Liceu subcorpus, where surtitles are synthesised and intermingled with the visual content. In both cases, surtitles convey a great deal of the emotional charge and they are responsible for moving the plot forward. Following the terminology of Social Semiotics applied to opera by Hutcheon and Hutcheon (2010), these two first tagsets coincide with the *design* of the opera, that is, the invariable elements, regardless of the specific aesthetics of a particular spectacle. Theatrical forms, on the other hand, represent the *production* of the opera, the creative output that depends on each specific staging. The visual aesthetics of opera are therefore expressed in the AD through scenography (18.47%), acting (9.09%),

proxemics (4.72%), lighting (3.43%), wardrobe (2.86%), props (1.82%), kinesics (1.33%) and dancing (1.17%).

Divergences between the Liceu and the Teatro Real lie mostly in the preponderance of scenography (*AD_CAT*=17.02%; *AD_ES*=20.03%) vs. surtitles (*AD_CAT*=14.01%; *AD_ES*=24.12%). The scenography tag ranks second in frequency for the Liceu AD, while surtitles occupy this position for the Teatro Real. Aside from this difference, the rest of the semiotic signs are distributed in the same order for both subcorpora, though there is no mention at all of instrumental music or noise in the Spanish subcorpus.

In short, after recognising the need to constantly name the characters and given the fact that the musical element does not need to be described, the aesthetic value in the AD is mostly placed on the scenography and the surtitles (or the operatic *parole*), and, to a lesser extent, on the performer (through acting, proxemics and kinesics).

As highlighted throughout this article, multimedia-annotated corpora encompass several technical and methodological challenges, such as representativeness, arduous alignment and annotation procedures (Soffritti 2018), and even more so in an art form which is performative by nature. Although this paper examines the textual representation of multimodality in AD, the study itself cannot be deemed multimodal, as its source of analysis is solely based on the AD scripts, and not their interaction with other modes and media, i.e. images and sounds (Jiménez Hurtado & Soler Gallego 2013: 577).

On a final note, this is an initial pilot study for a larger corpus, which will allow us to apply more in-depth statistical measurements and explore further linguistic and semiotic patterns.

6. Final remarks

This pilot study is a contribution to what is still an emerging subfield of corpus AD analysis. Going back to our initial research questions: Did we find any idiosyncratic or specialised language, and if so, was it significantly different to previous (filmic) AD corpora? Is there a clear hierarchisation of the operatic signs in AD scripts?

First of all, the present study has found a great number of lexico-grammatical commonalities between the Liceu and the Teatro Real subcorpora,

with some disparities in sentence length and adjective frequency. Conversely, some of the idiosyncratic lexico-grammatical features that are common in filmic AD corpora (Salway 2007; Reviers 2018; Matamala 2018), such as the semantic category of objects and indoor and outdoor locations, were not salient in the operatic medium. We can therefore corroborate that there are lexico-grammatical characteristics that are idiosyncratic to AD (regardless of the modality), but differences in genre exist and deserve further reflection.

Second, the article aimed to test the saliency of semiotic resources in opera AD. More specifically, we wondered whether the prioritisation of the signs in AD could tell us something about contemporary opera *design* and *production* (Hutcheon & Hutcheon 2010). In this light, character identification and surtitles, ultimately the two most common tags, belong to the *design* of the opera, just as much as the musical score. Here, we emphasise that surtitles (either read-aloud or synthesised) could potentially be reused for every performance in the same production (Orero 2007).[3]

Within the *production* media, scenography is the clear priority in operatic AD. Yet, is scenography as important as, for instance, the music for the audience? We raise the possibility of conducting a reception study to test our findings with viewing and blind audiences alike: Why do they go to the opera? What is the current perception of the music vs. *parole* debate, if still relevant at all? Has the aesthetic value of the stage performance ultimately overshadowed other design resources?

3. Accessibility services for live performances in the Spanish context are usually offered on one or two different days per run of the play, thus contravening the flexibility-in-use principle of universal design.

References

Arma, Saveria. (2011) *The language of filmic audio description: a corpus-based analysis of adjectives*. Naples: Università degli Studi di Napoli Federico II. Unpublished PhD thesis.

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

Baños, Rocío; Silvia Bruti & Serenella Zanotti. (2013) "Corpus linguistics and Audiovisual Translation: in search of an integrated approach." *Perspectives. Studies in Translatology* 21:4, pp. 483-490.

Bobes Naves, María del Carmen. (2004) "Teatro y semiología." *Arbor. Ciencia, Pensamiento y Cultura* CLXXVII, 699/700, pp. 497-508.

Cabeza-Cáceres, Cristobal & Anna Matamala. (2008) "La audiodescripción de ópera: una nueva propuesta." In: Pérez-Ugena, Álvaro & Ricardo Vizcaíno-Laorga (eds.) 2008. *ULISES. Hacia el desarrollo de tecnologías comunicativas para la igualdad de oportunidades*. Madrid: Observatorio de las Realidades Sociales y de la Comunicación, pp. 95-106.

Cabeza-Cáceres, Cristobal. (2010) "Opera audio description at Barcelona's Liceu Theatre." In: Díaz-Cintas, Jorge; Anna Matamala & Josélia Neves (eds.) 2010. *New insights into audiovisual translation and media accessibility*. Amsterdam: Rodopi, pp. 227-237.

Corpas-Pastor, Gloria. (2008) *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt am Main: Peter Lang.

Corral, Anna & Ramón Lladó. (2011) "Opera multimodal translation: audio describing Karol Szymanowski's Król Roger for the Liceu Theatre, Barcelona." *JosTrans. Journal of Specialised Translation* 15, pp. 163-179.

De Frutos, Rocío. (2013) *El debate en torno al canto traducido. Análisis de criterios interpretativos y su aplicación práctica: Adaptación al castellano de la ópera "Il barbiere di Siviglia" de G. Paisiello*. Sevilla: Universidad de Sevilla. Unpublished PhD thesis.

Delabastita, Dirk. (1989) "Translation and mass-communication: Film and T.V. translation as evidence of cultural dynamics." *Babel* 35:4, pp. 193-218.

Desblache, Lucile. (2007) "Music to my ears, but words to my eyes? Text, opera and their audiences." *Linguistica Antverpiensia* 6, pp. 155-170.

Di Giovanni, Elena. (2018a) "Audio description for live performances and audience participation." *JosTrans. Journal of Specialised Translation* 29, pp. 189-211.

DI GIOVANNI, Elena. (2018b) "Participatory accessibility: creating audio description with blind and non-blind children." *Journal of Audiovisual Translation* 1:1, pp. 155-169.

DURÁN, Pilar; David Malvern; Brian Richards & Ngoni Chipere. (2004) "Developmental trends in lexical diversity." *Applied Linguistics* 25:2, pp. 220-242.

EARDLEY-WEAVER, Sarah. (2010) "Opening doors to opera. The strategies, challenges and general role of the translator." *InTralinea* 12.

EVISON, Jane. (2010) "What are the basics of analysing a corpus." In O'Keeffe, Anne & Michael McCarthy (eds.) 2010. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp. 122-145.

FABER, Pamela & Ricardo Mairal Usón. (1999) *Constructing a lexicon of English verbs*. Berlin and New York: Mouton de Gruyter.

FRYER, Louise. (2016) *An introduction to audio description: a practical guide*. Oxford: Routledge.

GORLÉE, Dinda L. (1997) "Intercode translation: words and music in opera." *Target* 9:2, pp. 235-270.

GOTTLIEB, Henrik. (2005) "Multidimensional Translation: Semantics turned Semiotics." In Gerzymisch-Arbogast, Heidrun & Sandra Nauert (eds.) 2005. *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation*. Saarbrücken, pp. 33-61. <https://www.euroconferences.info/proceedings/2005_Proceedings/2005_Gottlieb_Henrik.pdf>

HANSEN, Gyde. (2010) "Integrative description of translation processes." In: Shreve, Gregory M. & Erik Angelone (eds.) 2010. *Translation and Cognition*. Amsterdam and Philadelphia: John Benjamins, pp. 189-211.

HUTCHEON, Linda & Michael Hutcheon. (2010) "Opera: Forever and always multimodal." In Page, Ruth (ed.) 2009. *New perspectives on narrative and multimodality*. London: Routledge, pp. 65-77.

JIMÉNEZ HURTADO, Catalina. (2010) "Fundamentos metodológicos del análisis de la AD." In Jiménez Hurtado, Catalina; Ana Rodriguez & Claudia Seibel (Coord.) 2010. *Un corpus de cine. Teoría y práctica de la audiodescripción*. Granada: Tragacanto, pp. 57-110.

JIMÉNEZ HURTADO, Catalina & Claudia Seibel. (2012) "Multisemiotic and multimodal corpus analysis in audio description: TRACCE." In Remael, Aline;

Pilar Orero & Mary Carroll (eds.) 2012. *Audiovisual Translation and Media Accessibility at the Crossroads*. Amsterdam: Rodopi, pp. 409-425.

Jiménez Hurtado, Catalina & Silvia Soler Gallego. (2013) "Multimodality, translation and accessibility: a corpus-based study of audio description." *Perspectives. Studies in Translatology* 21:4, pp. 577-594.

Koester, Almut. (2010). "Building small specialised corpora." In: O'Keeffe, Anne & Michael McCarthy (eds.) 2010. *The Routledge handbook of corpus linguistics*. London: Routledge, pp. 66-79.

Malamatidou, Sofia. (2018) *Corpus triangulation: Combining data and methods in corpus-based translation studies*. London: Routledge

Maszerowska, Anna; Anna Matamala & Pilar Orero. (2014). *Audio description: New perspectives illustrated*. Amsterdam: Rodopi.

Matamala, Anna. (2005) "Live Audio Description in Catalonia." *Translating Today* 4, pp. 9-11.

Matamala, Anna. (2007) "La audiodescripción en directo." In Jiménez Hurtado, Catalina (ed.) 2007. *Traducción y accesibilidad. Subtitulación para Sordos y audiodescripción para ciegos: nuevas modalidades de Traducción Audiovisual*. Frankfurt: Peter Lang, pp. 121-132.

Matamala, Anna. (2018) "One short film, different audio descriptions. Analysing the language of audio descriptions created by students and professionals." *Onomázein* 41, pp. 185-207.

Matamala, Anna. (2019) "The VIW project. Multimodal corpus linguistics for audio description analysis." *Revista Española de Lingüística Aplicada* 32:2, pp. 515-542.

Mateo, Marta. (2012) "Music and translation." In: Gambier, Yves & Luc van Doorslaer (eds.) 2012. *Handbook of translation studies*. Amsterdam: John Benjamins, vol. 3, pp. 115-121.

McCarthy, Philip M., & Scott Jarvis. (2010) "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment." *Behavior Research Methods* 42, 381-392.

Orero, Pilar. (2007) "Audiosubtitling: a possible solution for opera accessibility in Catalonia." *TradTerm* 13. São Paulo: Humanitas, pp. 135-149.

Orero, Pilar & Anna Matamala. (2007) "Accessible opera: Overcoming linguistic and sensorial barriers." *Perspectives. Studies in Translatology* 15:4, pp. 262-277.

ORERO, Pilar; Joan Bestard; Miquel Edo; Gonzalo Iturregui-Gallardo; Anna Matamala & Iris C. P. H. de Solás. (2019) "Opera accessibility in the 21st century: new services, new possibilities." *TRANS. Revista de Traductología* 23, pp. 245-256.

PEREGO, Elisa. (2019) "Into the language of museum audio descriptions: a corpus-based study." *Perspectives. Studies in Translatology* 27:3, pp. 333-349.

PFANSTIEHL, Margaret & Cody Pfanstiehl. (1985) "The play's the thing. Audio description in the theatre: Margaret and Cody Pfanstiehl." *British Journal of Visual Impairment* 3:3, pp. 91-92.

REVIERS, Nina. (2017) *Audio-description in Dutch: A corpus-based study into the linguistic features of a new, multimodal text type*. Antwerp: University of Antwerp. Unpublished PhD thesis.

REVIERS, Nina. (2018) "Studying the language of Dutch audio description." *Translation and Translanguaging in Multilingual Contexts* 4:1, pp. 178-202.

RĘDZIOCH-KORKUZ, Anna. (2016) *Opera surtitling as a special case of audiovisual translation. Towards a semiotic and translation based framework for opera surtitling*. Frankfurt am Main: Peter Lang.

ROSSI, Fabio & Maria Grazia Sindoni. (2017) "The phantoms of the opera: Toward a multidimensional interpretative framework of analysis." In: Sindoni, Maria Grazia; Janina Wildfeuer; Kay L. O'Halloran (eds.) 2017. *Mapping multimodal performance studies*. London/New York: Routledge, pp. 61-84.

RYAN, Marie-Laure. (2004) *Narrative across media: The languages of storytelling*. Lincoln, NB: University of Nebraska Press.

SALDANHA, Gabriela & Sharon O'Brien. (2013) *Research methodologies in Translation Studies*. Manchester: St. Jerome Publishing.

SALWAY, Andrew. (2007) "A corpus-based analysis of audio description." In Díaz-Cintas, Jorge; Pilar Orero & Aline Remael (eds.) 2007. *Media for all: Subtitling for the deaf, audio description, and sign language*. Amsterdam: Rodopi, pp. 151-174.

SEVERAL AUTHORS (ITC). (2000) *ITC guidance on standards for audio description*. <http://audiodescription.co.uk/uploads/general/itcguide_sds_audio_desc_word3.pdf >.

SOFFRITTI, Marcello. (2018) "Multimodal corpora and audiovisual translation studies." In Pérez-González, Luis (ed.) 2018. *The Routledge Handbook of Audiovisual Translation*. London & New York: Routledge, pp. 334-349.

Soler Gallego, Silvia. (2018) "Audio descriptive guides in art museums: A corpus-based semantic analysis." *Translation and Interpreting Studies* 13:2, pp. 230-249.

Tråvén, Marianne. (2005) "Musical rhetoric – the translator's dilemma: A case for Don Giovanni." In: Gorlée, Dinda L. (ed.) 2005. *Song and significance: Virtues and vices of vocal translation*. Amsterdam: Rodopi, pp. 103-120.

Tuominen, Tiina; Catalina Jiménez Hurtado & Anne Ketola. (2018) "Why methods matter: Approaching multimodality in translation research." *Linguistica Antverpiensia* 17, pp. 1-21.

Virkkunen, Riitta. (2004) "The Source Text of Opera Surtitles." *Meta*, 49:1, pp. 89-97.

York, Greg. (2007) "Verdi made visible: audio introduction for opera and ballet." In Díaz-Cintas, Jorge; Pilar Orero & Aline Remael (eds.) 2007. *Media for all: Subtitling for the deaf, audio description, and sign language*. Amsterdam: Rodopi, pp. 215-231.

BIONOTE / BIONOTA

Irene Hermosa-Ramírez is a PhD student in Translation and Intercultural Studies at Universitat Autònoma de Barcelona (UAB). Her research interests include Multimodality, Media Accessibility and accessibility for the scenic arts. Her doctoral thesis focuses on opera audio description. She holds a B.A. in Translation and Interpreting from the University of the Basque Country and an M.A. in Audiovisual translation from UAB. She is a member of the TransMedia Catalonia research group (2017SGR113), where she collaborates in the RAD project. She is the secretary of the Catalan Association for the Promotion of Accessibility.

Irene Hermosa-Ramírez es doctoranda en el programa de Traducción y Estudios Interculturales de la Universitat Autònoma de Barcelona (UAB). Sus intereses de investigación incluyen los Estudios Multimodales y la accesibilidad a los medios y a las artes escénicas. Su tesis doctoral se centra en la audiodescripción operística. Previamente estudió el Grado en Traducción e Interpretación en la Universidad del País Vasco/Euskal Herriko Unibertsitatea y el Máster en Traducción Audiovisual en la UAB. Forma parte del grupo

de investigación TransMedia Catalonia (2017SGR113) y colabora en el proyecto RAD. Es la secretaria de la Asociación Catalana para la Promoción de la Accesibilidad.

Recibido / Received: 13/04/2020
Aceptado / Accepted: 01/08/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.07

Para citar este artículo / To cite this article:
MEJÍAS-CLIMENT, Laura. (2021) "Los estudios de corpus y la localización: Una propuesta de análisis para material interactivo." En: CALZADA, María & Sara LAVIOSA (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection*. *MonTI* 13, pp. 220-250.

LOS ESTUDIOS DE CORPUS Y LA LOCALIZACIÓN: UNA PROPUESTA DE ANÁLISIS PARA MATERIAL INTERACTIVO

LAURA MEJÍAS-CLIMENT
lmejias@uji.es
Universitat Jaume I

Resumen

Este artículo pretende dar cuenta de las sincronías empleadas en el doblaje al español peninsular de un corpus multimodal compuesto por tres videojuegos del género interactivo de la acción-aventura. La metodología, de enfoque descriptivo, triangula datos cualitativos y cuantitativos obtenidos, por una parte, del análisis empírico del corpus multimodal y, por otra, del contacto directo con profesionales de la industria mediante entrevistas semiestructuradas. Asimismo, se revisarán algunos planteamientos previos de los estudios de corpus, estrechamente relacionados con los Estudios Descriptivos en Traducción, y de la práctica profesional de la localización, abordándola desde el enfoque de la Traducción Audiovisual (TAV). Se busca así exponer cómo combinar distintos métodos y perspectivas para analizar la modalidad de TAV del doblaje en un producto multimodal interactivo, aspecto escasamente investigado empíricamente en la esfera académica por el momento, a pesar de la utilidad que ya han demostrado los estudios de corpus en el panorama traductológico.

Palabras clave: Corpus multimodal; Interacción; Localización; Doblaje; Videojuegos.

Abstract

This article aims to analyze the dubbing synchronies used in a multimodal corpus composed of three video games, dubbed into Castilian Spanish, belonging to the interactive genre of action-adventure. The methodology, adopting a descriptive approach,

triangulates qualitative and quantitative data obtained, on the one hand, from the empirical analysis of the multimodal corpus and, on the other hand, from direct contact with professionals in the industry through semi-structured interviews. Additionally, some previous approaches within Corpus-Based Translation Studies—closely linked to Descriptive Translation Studies—will be reviewed, as well as the professional practice of localization, from the perspective of audiovisual translation (AVT). The goal is thus to present how different methods and perspectives can be combined to analyze the AVT mode of dubbing in a multimodal and interactive product, which remains largely unexplored in academia so far, despite the efficacy that corpus studies have demonstrated in translation studies.

Keywords: Multimodal corpus; Interaction; Localization; Dubbing; Video games.

1. El poliédrico marco de trabajo de los Estudios de Traducción

Uno de los atractivos del ámbito de la traducción es la versatilidad de enfoques, métodos y áreas que este abarca, tanto profesionales como investigadores. Esto nos permite diseñar modelos de investigación como el que se presenta en estas páginas, cuyo principal objetivo es determinar los tipos de ajuste reflejados en las situaciones de juego de un corpus compuesto por tres videojuegos del género interactivo de la acción-aventura: *Batman: Arkham Knight* (Rocksteady Studios, 2015), *Assassin's Creed Sydicate* (Ubisoft, 2015) y *Rise of the Tomb Raider* (Crystal Dynamics, 2016). Para ello, describiremos a continuación el planteamiento teórico y los estudios previos en los que se fundamenta nuestro diseño, así como la metodología empleada y los resultados obtenidos, de carácter tanto cuantitativo como cualitativo.

En las últimas décadas hemos presenciado el arraigo de los Estudios de Traducción basados en Corpus desde el artículo de referencia de Baker en 1993, que ponía de manifiesto la utilidad de la lingüística de corpus para la traducción (Zanettin 2014). La metodología basada en corpus se ha asociado estrechamente con los Estudios Descriptivos en Traducción, que gozan de más recorrido en el mundo académico desde las primeras publicaciones de los años 70 (Hermans 2020). Hasta nuestros días, han sido numerosísimas las propuestas metodológicas tanto en el uso de corpus como en los planteamientos descriptivos. Nos remitimos a Bernardini & Kenny (2020)

y Hermans (2020), respectivamente, para una sucinta revisión de las principales publicaciones asociadas a ambos campos.

Por otra parte, desde el prisma profesional, el desarrollo de las tecnologías ha traído consigo el establecimiento de la localización como una potente industria para la adaptación, tanto cultural como lingüística, legal y, en conjunto, funcional y cultural de todo producto multimedia (Jiménez-Crespo 2020). Sus orígenes se remontan a finales de la década de los setenta. Desde su consolidación entre los 80 y los 90 hasta el presente, la localización se ha desarrollado a un ritmo vertiginoso y se asocia con la adaptación de tres grandes grupos de productos: contenidos web, *software* y videojuegos (Mata Pastor 2005), además de los dispositivos móviles o *small devices* que añade Jiménez-Crespo (2013b).

Dichos productos, entre otros muchos, pueden considerarse ejemplos de textos multimodales (Kress y Van Leeuwen 2001), pues combinan múltiples modos semióticos integrados para crear su sentido conjunto. Aunque se trata de un concepto sobre el cual aún quedan cuestiones clave en las que alcanzar un consenso (Pérez-González 2020), la multimodalidad, desde su incursión en el ámbito académico en la década de los 90 (Jewitt, Bezemer y O'Halloran 2016), ha permitido el análisis holístico de la comunicación, teniendo en cuenta no únicamente el código lingüístico, sino también todo modo semiótico que lo acompaña y completa, ya sea visual, acústico, kinésico, etc. (Caple 2018).

De especial utilidad en el actual panorama multimedia ha demostrado ser la convergencia entre multimodalidad y estudios de corpus, mediante el análisis de corpus multimodales o MMC (Soffritti 2018), para dar cuenta de la configuración semiótica completa de los productos multimedia. En los MMC nos centraremos en estas páginas, con la elección de los videojuegos como producto multimodal y, además, interactivo, sobre el que aún quedan numerosísimos rincones que explorar (O'Hagan y Mangiron 2013).

Por último, un entorno más al que debemos referirnos, y en estrecha relación con la localización, con el uso de corpus y con la multimodalidad, es la Traducción Audiovisual (TAV), algo más antigua que la práctica localizadora, ya que surgió de la mano de la industria cinematográfica (Díaz Cintas y Anderman 2009) y, desde entonces, su flexibilidad ha sido evidente (Orero 2004) a la hora de acomodarse al amplio y cambiante mercado de productos

audiovisuales, no solo desde la práctica profesional, sino también con una gran variedad de planteamientos investigadores (Chaume 2018a).

Aprovechando, como se decía, la versatilidad de la traducción, todas estas perspectivas confeccionan el marco de interés del presente artículo, con el que pretendemos ofrecer un planteamiento original, empleando perspectivas mixtas, en la combinación de los Estudios Descriptivos y basados en Corpus, con la práctica de la localización y el ámbito de la TAV.

Dado el amplio alcance que tanto la localización como la TAV ofrecen, nos centraremos en una modalidad de TAV, el doblaje, escasamente explorada hasta ahora en la práctica localizadora profesional (O'Hagan y Mangiron 2013). Del mismo modo, también son "escasos [los] estudios sobre el uso de corpus en localización" (Jiménez-Crespo 2009: en línea), motivo por el cual hemos decidido basar la investigación en un corpus multimodal compuesto por los tres citados videojuegos.

Para delimitar el estudio de dicho corpus, nos centraremos en analizar las sincronías, una de las características principales del doblaje. Así, partiremos del polémico nexo entre TAV y localización para exponer la metodología particular que se ha diseñado para trabajar empíricamente con un corpus multimodal de videojuegos, cuya originalidad con respecto a estudios previos radica en la contemplación de una dimensión interactiva y en el uso de las *situaciones de juego* como unidad de análisis.

Además de la complejidad que ya presentan los estudios basados en corpus multimodales, en los que deben tenerse en cuenta diversos modos de transmisión del sentido (Soffritti 2018), en este caso sumamos un canal táctil no explorado en estudios empíricos en traducción hasta ahora, para lo cual se propone adaptar el planteamiento analítico a la medida de los intereses de la investigación y de la idiosincrasia del producto estudiado.

La complejidad aquí va más allá al contemplar una dimensión interactiva. Reconocemos que el diseño de esta investigación, tal y como señalan Taylor y Marchi (2018: 11), podría influir en la dirección que tome el análisis y en los resultados obtenidos. Por ello, hemos querido recurrir a la triangulación de datos para completar la información empírica con información procedente de los testimonios de profesionales del sector. No obstante, se trata de un diseño preliminar en el que se identificarán algunos puntos fuertes y débiles

a lo largo de las siguientes páginas y que puede tomarse como estudio de partida para refinar el procedimiento en el futuro.

Dado que este análisis contempla diversos campos y enfoques dentro de los Estudios de Traducción, los siguientes apartados reflejan el fundamento poliédrico de nuestro trabajo: la TAV y la localización como ámbitos profesionales analizados, los estudios de corpus como enfoque metodológico y, por último, los textos interactivos como producto multimodal que ofrece aún numerosas aristas de análisis.

2. Traducción Audiovisual y localización: encuentros y desencuentros

Ya desde comienzos de siglo (Orero 2004) y más aún en nuestros días (Chaume 2018b), la TAV parece funcionar como un concepto de gran amplitud bajo el que situar muy diversas modalidades de traducción, entendidas según los métodos técnicos empleados para trasvasar el código lingüístico de un texto audiovisual de partida a uno meta (Chaume 2004a: 31) o el modo del discurso del texto original y del meta (Hurtado Albir 2011: 69-70).

Tradicionalmente, la TAV engloba a aquellas modalidades que añaden una banda sonora al producto original (*revoicing*) y aquellas basadas en insertar texto en la pantalla en la que se proyecta el texto audiovisual de partida (*captioning*) (Chaume 2018b). En este caso, nos centraremos en la modalidad de *revoicing* de mayor tradición histórica en nuestro país: el doblaje.

Orero (2004: VIII) ya llamaba la atención sobre la flexibilidad del concepto de TAV para abarcar tantas diversas formas de transferencia multisemiótica como las nuevas tecnologías trajeran consigo:

> Audiovisual Translation will encompass all translations — or multisemiotic transfer — for production or postproduction in any media or format, and also the new areas of media accessibility […]. Technological developments which have changed paper oriented society towards media oriented society have also made Audiovisual Translation the most dynamic field of Translation Studies.

En el presente, además de transferencias inter e intralingüísticas, contamos también con la intersemiótica (Jakobson 2000), con modalidades de TAV como la audiodescripción o las audioguías para museos; también puede ubicarse bajo el paraguas de la TAV la transadaptación (Gambier 2003,

Neves 2005, Pruys 2009), la transcreación (O'Hagan y Mangiron 2013, Bernal Merino 2015), las narrativas transmedia (Bernal Merino 2015, Pujol Tubau 2015) y las adaptaciones o *remakes*, así como la localización. Es aquí donde nos interesa hacer referencia al actual debate sobre TAV y localización como prácticas estrechamente relacionadas entre las cuales los límites no están claros, si es que alguna vez lo han estado (O'Hagan y Mangiron 2013), y menos aún lo están a medida que avanzan los medios tecnológicos con nuevos productos y géneros.

Por una parte, un sector de los profesionales dedicados a la localización defiende su clara diferenciación de la TAV al concebir la traducción desde una posición reduccionista y meramente lingüística (Cadieux y Esselink 2004, Maxwell-Chandler y Deming 2012). También otros profesionales y algunos investigadores defienden la visible separación de los ámbitos de la localización y la TAV, por tratarse la primera de un proceso industrial particular que abarca muchas más modificaciones y etapas que la traducción (Jiménez-Crespo 2013b, Pym 2014, Muñoz Sánchez 2017, Méndez González y Calvo-Ferrer 2017).

Por otro lado, autores como Bernal Merino (2006) y O'Hagan y Mangiron (2013) adoptan el uso del término *localización*, dado su extendido empleo en la industria, aunque reconocen que no contempla nada que no abarque ya un sentido amplio de la traducción, por lo que no sería necesaria una estricta separación en dos ámbitos diferenciados.

Ante ambas posturas, la clave parece estar entonces en la perspectiva que se adopte: bien la posición de la industria, que aboga por la diferenciación de la localización, o bien, la académica, donde no parece ser estrictamente necesario crear un nuevo paradigma. En estas páginas, emplearemos el término *localización* entendiendo tal práctica como la plena adaptación de un producto multimedia (contenidos web, *software* o videojuegos) a un mercado local meta, en línea con los enfoques de O'Hagan y Mangiron (2013) y Bernal-Merino (2015). Se diferencia así de la práctica profesional de la TAV, entre otros aspectos, en los productos de los que se ocupa y la diversidad de procedimientos que abarca.

En lo que respecta al planteamiento metodológico de este estudio, como se verá en la sección 4, tal concepción de la localización no nos impide situarnos en la misma posición adoptada por Vázquez Rodríguez (2018),

quien, ante el debate sobre la distinción entre TAV y localización, propone adaptar la metodología descriptivista ya empleada en la TAV para contemplar particularidades que pueda presentar el producto audiovisual y multimedia analizado, sea cual sea —en este caso, el principal rasgo particular es la presencia de la interactividad en un corpus de videojuegos—. Por lo tanto, entendemos la localización y la TAV como prácticas profesionales que se ocupan tradicionalmente de distintos productos, pero cuya convergencia en numerosos aspectos favorece la adopción de metodologías de investigación comunes.

3. Los Estudios de Traducción basados en Corpus y los productos multimodales

Tal y como Baños, Bruti y Zanotti (2013) ponen de manifiesto, la lingüística de corpus ha demostrado ampliamente su utilidad en los Estudios de Traducción en general (Baker 1996, Olohan 2004, Laviosa 2012), así como en la TAV en particular (Mangiron 2017), según muestran las contribuciones de Forchini, Baños, Valentini, Valirano, Jiménez Hurtado y Soler Gallego, y Bywood *et al*. en la citada edición de Baños, Brutti y Zanotti (2013) sobre la aplicación de la lingüística de corpus a la TAV, entre muchísimas otras obras.

A pesar de los grandes avances que los corpus han aportado a la investigación en TAV en términos de generalizaciones descriptivas y mejora de la calidad, aún quedan aspectos en los que trabajar, como la representatividad, el tamaño, la comparabilidad, la transcripción, la anotación y la alineación de estos corpus audiovisuales (Pavesi 2018). Aunque nuestro estudio es aún un punto de partida en el análisis de un pequeño corpus interactivo y adolece de varias limitaciones en las citadas categorías, consideramos que representa una propuesta preliminar original que puede apuntar a nuevas perspectivas en el estudio de los corpus multimodales.

Situar el análisis de videojuegos en el paradigma de la TAV ofrece la ventaja de que, casi desde sus inicios, el objeto de estudio se ha abordado desde un enfoque multimodal (Kaindl 2013: 263), teniendo en cuenta explícitamente tanto elementos verbales como no verbales, aunque, ciertamente, los videojuegos y la interacción que traen con ellos son todavía más recientes, pues su origen puede situarse alrededor de los años 60 (Kent 2001, López

Redondo 2014) y la metodología de estudio ha de ajustarse a dicha interacción adecuadamente, en el conjunto multimodal.

Mangiron (2017) llama la atención sobre la necesidad de emprender estudios de corpus en el ámbito de la localización. Así, podría recuperarse una notable cantidad de datos sobre tendencias y regularidades en la localización de videojuegos. En este campo se han estudiado ampliamente el proceso (Maxwell-Chandler y Deming 2012, O'Hagan y Mangiron 2013; Bernal Merino 2015), las restricciones (Dietz 2007, Loureiro Pernas 2007, Muñoz Sánchez 2017) y las estrategias de traducción (Fernández Costales 2012), pero no se han desarrollados tantos estudios empíricos basados en corpus como en otros tipos de texto.

Hasta la fecha existen diversos estudios de caso que analizan videojuegos concretos (Crosigniani y Ravetto 2001, Fernández Torné 2007, Mangiron 2010, Ensslin 2012, Müller Galhardi 2014 y Van Oers 2014, entre otros), pero únicamente se han empleado corpus de videojuegos, hasta donde se ha podido comprobar, en la obra de Pujol Tubau (2015) —aunque no exclusivamente, pues su análisis reúne producción transmedia— y en la obra de Vázquez Rodríguez (2018). En ninguno de estos trabajos, sin embargo, se emplean las situaciones de juego como unidades de análisis (véase sección 4.1.), aspecto que suma originalidad a nuestra propuesta.

Tras la introducción de los corpus en los Estudios de Traducción a comienzos de los 90, fue haciéndose patente la necesidad de adoptar enfoques más allá de los puramente descriptivistas y lingüísticos para dar cuenta de la complejidad de la traducción ampliando el marco de estudio mediante su contextualización y la integración de herramientas de análisis de otras áreas (Olohan 2004). La integración de corpus y métodos mixtos de investigación del proceso de traducción ha resultado especialmente fructífera (Malamatidou 2018). Además de diseños de investigación basados en distintos métodos, algunos proyectos recientes han ido expandiéndose hacia fuentes previamente ignoradas, quizá por su complejidad, y producidas bajo condiciones relativamente recientes (Bernardini y Kenny 2020: 113), como es el caso de los trabajos de Jiménez-Crespo (2013a, 2015), centrados en el análisis de corpus compuestos por una modalidad (la localización web) y un género (los sitios de redes sociales) que ni siquiera existían cuando comenzó a sistematizarse el estudio de tendencias en traducción y el uso de corpus.

En estas nuevas investigaciones, a cuyo paradigma la nuestra pretende sumarse, uno de los aspectos cruciales es la contemplación de la dimensión no verbal y la configuración semiótica multimodal en el diseño del corpus. Combinando la teoría multimodal y los adecuados diseño y consulta del corpus, la investigación puede abrirse a la naturaleza multisemiótica del producto y su impacto en la traducción. Por lo tanto, un enfoque integrador que combine la lingüística de corpus con el análisis multimodal y recursos y fuentes extralingüísticos (entrevistas), triangulando los datos, podrá dar cuenta de la naturaleza híbrida y compleja del producto (Baños, Bruti y Zanotti 2013).

4. Metodología: triangulación de datos con un corpus interactivo

A continuación, describiremos el proceso de trabajo con un pequeño corpus paralelo (Laviosa 2002) en el que se han analizado los tipos de ajuste empleados en las versiones de partida, en inglés, y doblada al español peninsular, en los segmentos originales y sus correspondientes traducciones (Toury 1995). Se trata de un corpus multimodal o multimedia (Soffritti 2018), ya que recoge tanto vídeos (imagen y audio) como texto. El corpus abarca 2635 registros con cadenas de texto de entre 2 y 300 palabras (1326 registros en español y 1309 en inglés) que se extienden durante 76 horas de juego. Los datos extraídos del análisis del corpus se triangularán con información procedente de entrevistas semiestructuradas que también presentaremos.

Ante la escasez de estudios académicos que ahondaran en el nexo entre doblaje y localización de videojuegos, el punto de partida de este estudio es la siguiente pregunta de investigación: ¿qué tipos de ajuste pueden asociarse a cada situación de juego presente en videojuegos de acción-aventura? Será necesario ampliar el corpus en el futuro para confirmar los resultados obtenidos aquí, que pueden entenderse como una aproximación inicial a esta pregunta, dada la limitación del corpus a una selección de tres videojuegos de este género. Los métodos que se emplearán para responderla derivarán de las necesidades que la propia pregunta nos plantea (Taylor y Marchi 2018: 3), así como de la creatividad necesaria para enfocar un estudio de corpus tan particular (*ibid*., 6).

4.1. El análisis de los textos interactivos: semiótica y segmentación del corpus

Al abordar el análisis de un corpus compuesto por videojuegos bajo el amplio paraguas de la TAV es necesario, para comenzar, delimitar el concepto y la configuración semiótica particular del objeto de estudio con el fin de poder diseñar posteriormente el análisis de acuerdo con su idiosincrasia. Ciertamente, se trata del producto audiovisual (y multimodal) actual más complejo, dados los variados modos que coexisten en su entramado semiótico (Maietti 2004).

Como producto multimedia, los videojuegos encajan en la ya conocida concepción de todo texto audiovisual como aquél que se transmite a través de los canales acústico y visual y cuyos diversos códigos se entrelazan para tejer su sentido completo (Chaume 2012). Ahora bien, en el caso de los videojuegos, además de los códigos acústicos y visuales de dichos canales, debe tenerse en cuenta también la dimensión interactiva, que quizás sea la clave en el éxito mundial que han experimentado estos productos en tan solo unas décadas (López Redondo 2014), puesto que convierten al espectador pasivo en figura activa y protagonista de los hechos en pantalla. Así, a la ya conocida configuración semiótica de todo producto audiovisual, en videojuegos debemos sumar un canal táctil (Pujol Tubau 2015) que vehicula códigos hápticos bidireccionales (juego ↔ jugador) cuando se emplea un mando o algún otro periférico para jugar. Nos remitimos a Mejías-Climent (2019) para una revisión del funcionamiento del canal táctil y otras formas de juego sin mando.

En el caso de los tres videojuegos que componen nuestro corpus, el jugador recibirá información a través del canal táctil mediante códigos hápticos si el mando vibra. Pero, sobre todo, este usuario devolverá información empleando el tacto en movimiento (de nuevo, códigos hápticos en forma de pulsación de botones o palancas en el periférico), que el juego interpretará en función de sus reglas y ante los que reaccionará para completar así la creación del sentido conjunto del videojuego.

En este tipo de productos interactivos, el desarrollo de los hechos depende, por tanto, de la acción continua de quien juega, que va interactuando con el juego y causando la alternancia constante de distintas *situaciones de juego* (Pujol Tubau 2015, Mejías-Climent 2017), las cuales implican

diferentes niveles de interacción con el usuario. En concreto, se trata de tareas, diálogos, acción de juego y cinemáticas. Las dos primeras pueden darse durante interacción plena o parcial (el jugador puede recibir tareas o dialogar con otros personajes mientras actúa plenamente o, por el contrario, sus movimientos se limitarán parcialmente, aunque no por completo, para obligarlo a atender a dichas instrucciones o conversaciones); la acción representa la plena interacción del jugador con el producto, mientras que las cinemáticas detienen por completo toda interacción para trasladarnos por unos segundos a un vídeo tradicional que recurre a la configuración cinematográfica.

La estructura audiovisual de un videojuego, por lo tanto, no está cerrada de antemano, como sí sucede en películas o series, de forma que el análisis de un corpus compuesto por productos interactivos habrá de tener en cuenta esta particularidad a la hora de estructurarse. Aprovechando los distintos niveles de interacción y la clara diferenciación de las citadas cuatro situaciones de juego, estas se tomarán como la unidad de análisis que permitirá identificar y situar el fenómeno estudiado dentro del corpus (las sincronías del doblaje) en lugar de recurrir a los códigos de tiempo, como se haría en corpus compuestos por películas o series, de duración y estructura únicas y cerradas.

4.2. El doblaje como nexo común y delimitador del corpus

Además de constituir el fenómeno específico que se analiza en este corpus, la sincronía, por su parte, constituye uno de los estándares de calidad que propugna la industria del doblaje en nuestro país (Chaume 2007). Representa la coherencia entre la imagen que se aprecia en pantalla y el elemento sonoro, y se ha clasificado en tres tipos para cine y televisión: sincronía fonética o labial (reproducción de la articulación de los labios de los personajes, siempre que se les vea en primer plano o plano detalle), cinésica (correspondencia de la traducción con los movimientos y expresividad de los personajes) e isocronía (misma duración de los enunciados originales y traducidos) (Chaume 2004b).

En el caso del doblaje de videojuegos, varias son las diferencias reseñables en el proceso y los materiales disponibles en comparación con la práctica

del doblaje en productos no interactivos (Mejías-Climent 2019). De entre todas ellas, cabe señalar en especial la no disponibilidad de los vídeos que se doblarán, ni para los traductores —que a menudo trabajan con meras restricciones de espacio marcadas en las cadenas de texto descontextualizadas que traducen—, ni para los agentes en sala de doblaje (actores, directores y técnicos de sonido), cuya única referencia, en la mayoría de los casos, son las ondas de audio originales.

La sincronía en videojuegos, por tanto, ha de entenderse como una gradación de restricciones, que pueden aplicarse hasta en cinco niveles (es decir, se contabilizan cinco tipos de ajuste o sincronía) (Pujol Tubau 2015; Mejías-Climent 2018, 2019, 2020). Cada una de ellas implica lo siguiente:

(1) Libre: segmentos traducidos sin restricción (voces en *off*).
(2) Temporal: los segmentos traducidos pueden ser un 10 o 20 % más cortos o largos que los originales.
(3) Temporal exacto: los segmentos traducidos han de tener exactamente la misma duración que los originales, sin tener en cuenta pausas ni articulaciones internas.
(4) Sonoro: los segmentos traducidos han de tener exactamente la misma duración que los originales y, además, respetar pausas o entonaciones particulares.
(5) Labial: los segmentos traducidos han de ser idénticos a los originales en cuanto a duración y articulación, similar al ajuste labial cinematográfico.

Estos tipos de ajuste guiarán la pregunta principal que se consultará en el corpus, pues se buscará identificar una relación entre situaciones de juego, que implican distintos niveles de interacción, y tipos de ajuste, que pueden ser más o menos restrictivos, según se ha explicado. Los datos que se obtengan serán de carácter cuantitativo, en función de los porcentajes asignados a cada tipo de ajuste identificado para cada situación de juego, lo cual nos permitirá, a su vez, llegar a conclusiones cualitativas sobre los niveles de restricción de las situaciones de juego en relación con las sincronías del doblaje. Esta información se completará con entrevistas a los agentes del proceso (véase apartado 4.4.).

4.3. Diseño del estudio: ficha de análisis

Un estudio empírico para trazar tendencias en el uso de las sincronías del doblaje en un determinado contexto sociocultural requiere, en primer lugar, de un corpus de análisis claramente delimitado sobre el que puedan detectarse los patrones, en este caso, empleados en el doblaje al español peninsular de videojuegos originalmente desarrollados en inglés. Como vemos, se trata de un corpus multimodal paralelo, bilingüe y unidireccional (inglés>español) (Laviosa 2012). Posteriormente, una serie de entrevistas con los principales agentes del proceso de traducción y doblaje nos darán las claves principales sobre el contexto sociocultural en el que se sitúa el corpus.

Siguiendo las fases de la metodología que describe Laviosa (2012: 68), para empezar, se establecieron los criterios para seleccionar el corpus de entre una población de 106 videojuegos comercializados en España, entre 2015 y 2016, con los diálogos sonoros disponibles en español. Para reducir esta población a un catálogo de menor tamaño se añadió el requisito de que la lengua original de desarrollo (y, por tanto, de partida para el doblaje) fuera el inglés. Los 80 videojuegos resultantes, de nuevo, se acotaron aplicando los siguientes filtros: el *género interactivo* (Wolf 2005; Mejías-Climent 2019) ha de ser la acción-aventura; el *modo*, visión en 3.ª persona —de forma que la perspectiva no pueda alternar entre 1.ª y 3.ª, a elección del jugador, y la apreciación de las sincronías se mantenga así más homogénea— y para un solo jugador —la inclusión de más participantes ampliaría las posibilidades del juego y dificultaría aún más la replicación de la ruta jugada—; las *desarrolladoras y distribuidoras* serán distintas y, por último, también las *empresas de localización* encargadas de la traducción y el doblaje serán diferentes, en ambos casos, para evitar detectar patrones limitados a una misma empresa, tanto desarrolladora/distribuidora como localizadora, cuyas preferencias en el doblaje pudieran ofrecer resultados asociados a un enfoque particular corporativo, más que al tipo de videojuegos elegidos. De esta forma, los videojuegos que componen finalmente el corpus se vieron reducidos únicamente a tres: *Batman: Arkham Knight* (BAK), *Assassin's Creed Sydicate* (ACS) y *Rise of the Tomb Raider* (RTR).

La segunda fase la constituye la segmentación del corpus y la alineación entre origen y meta para determinar las relaciones que buscamos entre ambas

versiones (Laviosa 2012: 68). Como se ha explicado, este corpus multimodal se divide siguiendo, en última instancia, la alternancia continua de situaciones de juego a lo largo de cada uno de los tres juegos analizados. Con algunas salvedades, causadas por el azar presente en todo producto interactivo, la alineación de los segmentos origen y meta, realizada de forma manual recogiendo los datos en columnas contiguas de Excel, resultó exitosa, a pesar de la laboriosa tarea de jugar en ambas versiones, español e inglés, procurando repetir exactamente los mismos pasos en cada uno de los tres juegos. Esta, sin embargo, puede constituir una de las principales limitaciones en un análisis de material interactivo, pues el azar puede causar la no concordancia exacta de segmentos originales y meta en algunos casos. No obstante, en nuestro corpus los porcentajes de no coincidencia son mínimos y, por tanto, poco representativos (las versiones original y meta en BAK coinciden al 96,75 %; en ACS, al 100 % y en RTR, al 99,73 %).

Para estructurar esta segmentación de manera que pudiera accederse al corpus posteriormente de forma cómoda, se grabaron en vídeo las *gameplays* de cada videojuego en ambas versiones y cada fragmento de vídeo se etiquetó convenientemente, de forma manual, para anotar en el Excel de trabajo dónde se encontraba cada elemento que se iba analizando. Se obtuvieron un total de 300 GB de vídeo, algo más de 76 horas de juego, que recogen 2635 registros con cadenas de texto (1326 en español y 1309 en inglés entre los tres juegos). Este Excel (uno para cada videojuego) constituye la ficha básica de análisis, distribuido de la siguiente manera:

Momento argumental	Situación de juego	Tipo de ajuste (ES)	Tipo de ajuste (EN)	Tipo de cadena textual	Comentarios y código de vídeo (ES)	Comentarios y código de vídeo (EN)

Tabla 1. Ficha de análisis

La tercera fase de nuestra metodología, una vez llevada a cabo la extracción de datos, fue realizar las correspondientes generalizaciones de primer nivel sobre las tendencias observadas entre el texto origen y el texto meta. Estas

apreciaciones las recogemos en el apartado siguiente, referido a los resultados. Como señala Laviosa (2012: 68), será pertinente —y necesario— corroborar estos resultados con futuras ampliaciones del corpus para conseguir mayores niveles de generalización que apunten a tendencias más claras en el uso de las sincronías del doblaje en videojuegos.

Las principales limitaciones que puede presentar un corpus de estas características, además de la ya mencionada, se refieren a su representatividad, a la segmentación y a la transcripción y los métodos de anotación para dar cuenta de su complejidad semiótica (Baños, Bruti y Zanotti 2013; Pavesi 2018).

Aunque la representatividad pueda resultar algo limitada en comparación con otros grandes corpus de cientos de miles de palabras, en este caso debe reconocerse que alcanzar unos criterios de representatividad comunes a todo corpus (multimodal o basado únicamente en texto escrito) no es nada sencillo, dada la ingente cantidad de material y datos que un corpus multimedia genera (Soffritti 2018: 340). Por tanto, siguiendo las recomendaciones de este autor, se han aplicado los filtros ya enumerados al catálogo de videojuegos para reducirlo a un corpus de estudio abarcable que, si bien no da cuenta de una realidad de gran alcance, sí se ha seleccionado en función del fenómeno que se analiza y permite extraer generalizaciones preliminares sobre las que ahondar en futuros trabajos, tal y como se ha hecho anteriormente en otros muchos estudios descriptivos con corpus audiovisuales.

La segmentación, como se ha expuesto, se basa en las situaciones de juego, que representan la unidad de análisis de las sincronías del doblaje y son causa directa y diferenciadora de la presencia de la interacción. Cada situación de juego puede contener diversas cadenas de texto. La alineación presenta ciertas limitaciones inevitables en todo producto sujeto a un considerable grado de aleatoriedad, puesto que, a pesar de reproducir exactamente el mismo camino en las versiones original y traducida, el juego no devuelve siempre respuestas completamente idénticas. Según explica Wolf (2005: 7):

> Instead of fixed, linear sequences of text, image, or sound which remain unchanged when examined multiple times, a video game experience can vary widely from one playing to another.

A pesar de ello, en este trabajo nos centramos en las sincronías, un fenómeno que depende del código lingüístico, pero no se limita a este, lo que

nos permite analizar segmentos, en términos de sincronías, cuyo contenido lingüístico no sea exactamente el mismo.

La transcripción y anotación (véase figura 1) se ha hecho sobre la ficha de análisis (tabla 1) en formato de hoja de cálculo a medida que se ha ido jugando y grabando la *gameplay*, y también posteriormente. En ella se puede acudir a la herramienta *filtros* y hacer consultas rápidas, y las anotaciones abarcan todos los canales de comunicación, tanto aspectos visuales relevantes para la configuración de las sincronías como cuestiones acústicas o relativas al tipo de interacción que hayan podido influir en la forma de ajuste empleada. Por ahora, esta es la única opción para gestionar y analizar componentes de audio traducibles en un corpus compuesto por videojuegos, dado que no disponemos de las hojas de cálculo que reciben los traductores ni del guion *as rec*, según sale del estudio de doblaje. Tampoco podría contarse inicialmente con una transcripción completa del texto, ya que, al ser material interactivo y dinámico, este guion no existe ni siquiera antes de jugar, sino que se va construyendo según la interacción entre el jugador y el juego alternando distintas situaciones. Asimismo, este tipo de análisis con corpus multimodales presenta la limitación de que el texto no puede tratarse mediante ningún tipo de *software*, al menos por el momento, sino que debe analizarse manualmente, con la lentitud e imprecisión inherentes a este proceso. La figura 1, a continuación, muestra cuatro registros anotados para BAK, a modo de ejemplo.

1	Momento argumental	Situación de jueg	Ajuste (ES)	Ajuste (EN)	Cadena text	Comentarios (ES)	Comentarios (EN)
372	Objetivo: Devolver a los infectados por el Joker a su celda 14	Acción	Libre	Libre	Enunciados breves PNJ	Harley Quinn amenaza a Batman continuamente. Ejemplo de enunciado muy natural: «Por supuesto. ¿Te has enterado, pedazo de cabrón? ¡Vamos a por ti!». [BAK10_ES 00:36:38]	«You hear that, Bat-freak!? We're coming to get you!» [BAK7_EN 00:57:12]
373	Objetivo: Devolver a los infectados por el Joker a su celda 15	Tarea	Libre	Libre	Instrucciones del personaje para el jugador	[BAK10_ES 00:38:40 + 00:39:32]	[BAK7_EN 01:00:02]
374	Objetivo: Devolver a los infectados por el Joker a su celda 16	Cinemática	Sonoro	Sonoro	Enunciado breve PNJ	*Quick-time event* para contraatacar a la infectada que se lanza sobre el jugador. [BAK10_ES 00:40:14]	[BAK7_EN 01:00:35]
375	Objetivo: Devolver a los infectados por el Joker a su celda 17	Diálogo	Libre	Libre	Diálogo en acción de juego	Robin se queja del ataque. Incoherencia en la respuesta de Batman: «R: ...Sé, sincero, ¿qué pinta tengo? B: Tan lento como tú». [BAK10_ES 00:40:35]	Ni siquiera mueven los labios. En inglés: «R: …Be honest. How's it look? B: Like you're too slow». [BAK7_EN 01:00:55]

Figura 1. Muestra de la segmentación y anotación de cuatro situaciones de juego en BAK

4.4. La triangulación de datos: entrevistas a los agentes de la traducción

En la búsqueda de tendencias en el uso de las sincronías del doblaje también hemos recurrido al contexto del que emergen los textos multimodales que constituyen el corpus de estudio. La triangulación, según apuntan Taylor y Marchi (2018: 10), es un método valioso para completar el análisis, puesto que tiende a aportar información complementaria que amplía la perspectiva de los datos recogidos, además de ser la base de una investigación sólida que analice el fenómeno en su conjunto (Malamatidou 2018). En este caso, no emplearemos la triangulación basada en la combinación de distintos tipos de corpus, sino, más bien, en la combinación de métodos cuantitativos (el estudio empírico) y cualitativos (entrevistas), en términos de Malamatidou (2018).

Además de la habitual revisión y preparación del marco teórico, de carácter interdisciplinar, dada la polifacética naturaleza de los videojuegos, la consulta de las fuentes textuales se completa con entrevistas semiestructuradas (Kvale y Brinkmann 2009) que dan cuenta de los factores extratextuales referidos a la producción del material analizado.

Dada la estricta confidencialidad bajo la que se trabaja en el ámbito de los videojuegos (Mangiron 2017: 85-86), omitiremos los nombres de las personas entrevistadas, quienes representan la cadena completa de agentes que intervinieron en el proceso de traducción y doblaje para los videojuegos en cuestión. El planteamiento inicial fue entrevistar a todos los agentes desde que se realiza el encargo de traducción hasta que se entrega, es decir, el cliente y la administración (gestores de proyectos) en la preparación del proyecto; traductores y revisores durante la producción; y directores y actores de doblaje, además de los técnicos de sonido, durante la posproducción de la traducción. Sin embargo, los clientes tuvieron que descartarse por la imposibilidad de acceder a ellos. De las 18 figuras restantes, logramos entrevistar a 16, mediante llamada telefónica o videoconferencia, con la posterior transcripción manual de toda la conversación para analizar la información convenientemente.

El diseño de las entrevistas para todos los agentes respondió a tres grandes bloques de preguntas: el perfil de la persona, las indicaciones del encargo

y el funcionamiento del proceso de traducción y, por último, la fase de doblaje en sala.

La intención de esta herramienta de investigación fue integrar (Malamatidou 2018: 9) la información cualitativa con la extracción empírica para complementar los datos analizados empíricamente, de forma que pudiéramos comprobar si las tendencias observadas se correspondían, en realidad, con la manera en que todos estos profesionales aplicaron las sincronías intencionadamente o si, por el contrario, la aplicación de sincronías es responsabilidad únicamente de algunos de los agentes.

5. El análisis de las sincronías en nuestro corpus multimodal

Veamos, pues, cuáles han sido los resultados obtenidos tras delimitar el corpus, los fenómenos analizados y las herramientas de trabajo empleadas. La información cuantitativa resultante del análisis empírico se complementará con el enfoque cualitativo que ofrecen las entrevistas a los agentes del proceso de traducción y doblaje de los tres videojuegos estudiados.

5.1. Datos cuantitativos extraídos de la ficha de análisis del corpus

Como se ha indicado, el total de tiempo jugado recogido en las tres tablas de cálculo abarca algo más de 76 horas. De ellas, 26 se corresponden con las versiones original (13 horas) y doblada al español (13 horas) de *Batman: Arkham Knight* (BAK); 35 horas abarcan las versiones en inglés (17 horas) y en español (18 horas) de *Assassin's Creed Sydicate* (ACS) y, finalmente, 15 horas se extendió *Rise of the Tomb Raider* (RTR) en inglés (7 horas) y en español (8 horas).

La información extraída del corpus mediante las consultas con filtros en la hoja de cálculo nos devuelve datos cuantitativos sobre las distintas situaciones de juego que componen cada versión de cada videojuego y los tipos de ajuste observados en cada una de ellas. Para sintetizar esta información, se recogen a continuación en la figura 2 y en la tabla 2, numérica, los datos sobre los tipos de ajuste (libre, temporal, temporal exacto, sonoro y labial) detectados en cada situación de juego (tareas, acción, diálogos y cinemáticas) para BAK, ACS y RTR:

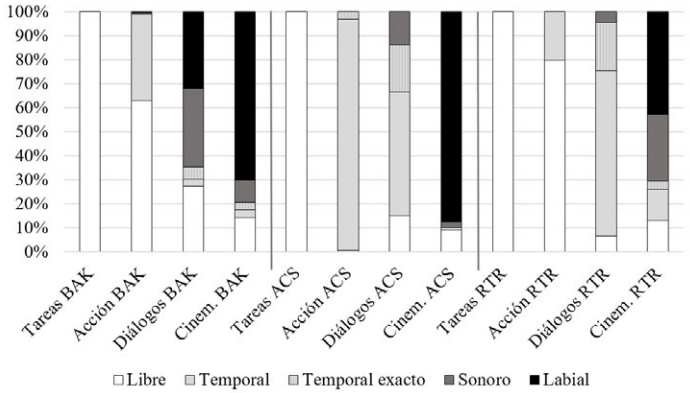

Figura 2. Tipos de ajuste en cada situación de juego de BAK, ACS y RTR

		Libre	Temporal	Temp. exacto	Sonoro	Labial
BAK	Tarea	100 %	0 %	0 %	0 %	0 %
	Acción	63,62 %	36,32 %	0 %	0,53 %	0,53 %
	Diálogos	27,59 %	2,87 %	5,17 %	32,76 %	32,18 %
	Cinemát.	14,29 %	3,06 %	3,06 %	9,18 %	69,39 %
ACS	Tarea	100 %	0 %	0 %	0 %	0 %
	Acción	0,62 %	96,27 %	3,11 %	0 %	0 %
	Diálogos	15,15 %	51,51 %	19,7 %	13,64 %	0 %
	Cinemát.	8,98 %	1,2 %	0 %	2,39 %	87,42 %
RTR	Tarea	100 %	0 %	0 %	0 %	0 %
	Acción	79,75 %	20,24 %	0 %	0 %	0 %
	Diálogos	6,67 %	68,9 %	20 %	4,44 %	0 %
	Cinemát.	13,04 %	13,04 %	3,49 %	27,83 %	42,61 %

Tabla 2. Tipos de ajuste en cada situación de juego de BAK, ACS y RTR

Como se observa en los datos anteriores, parece darse una relación entre determinadas situaciones de juego y tipos de ajuste en estos tres videojuegos de acción-aventura: las tareas, en los tres títulos, presentan siempre un ajuste

libre, es decir, sin restricciones, lo cual resulta lógico al tratarse siempre de mensajes que se transmiten al jugador de forma diegética mediante voces en *off*. También en las cinemáticas de los tres juegos parece haber una preferencia clara por el ajuste labial, el más restrictivo y preciso, lo cual enlaza con la naturaleza audiovisual tradicional de este tipo de escenas en las que el videojuego suprime por completo la interactividad y nos traslada a los recursos cinematográficos no interactivos. También hay algunos ejemplos de ajuste libre en las cinemáticas, dado que en todos los juegos se suceden narraciones con voces en *off*.

La acción de juego es una situación algo más variada: en BAK y RTR hay una clara preferencia por el ajuste libre y bastante presencia del ajuste temporal, mientras que, en el caso de ACS, el uso del ajuste temporal es totalmente predominante. Este dato puede deberse a la propia naturaleza del juego: en BAK y RTR, los personajes emplean continuamente *walkies-talkies* e intercomunicadores para hablar entre ellos mientras el jugador está actuando, lo cual motiva el uso preferente del ajuste libre, pues se trata de voces en *off*. En el caso de ACS, al situarse en un medio[1] histórico, la Londres victoriana de mediados del siglo XIX, la comunicación ha de ser en persona, por lo que el ajuste temporal representa la sincronía más útil en la configuración dinámica e interactiva de la acción con personajes relativamente visibles (aunque no con la precisión que permiten las cinemáticas).

Por último, los diálogos responden a su naturaleza híbrida: en los tres juegos se dan alternativamente con ciertas restricciones de la interacción, cuando se trata de diálogos que el jugador debe escuchar necesariamente, o

1. Debe señalarse que *medio* no se concibe aquí como el espacio o formato en el que se desarrolla un hecho (televisivo, interactivo, etc.), sino como uno de los distintos criterios que pueden emplearse para clasificar videojuegos (Wolf 2005). Así, entendemos el *medio* como las convenciones estilísticas, narrativas, temáticas e iconográficas en las que se sitúa la acción (*ibid*.: 114). Esta concepción podría equipararse a la división tradicional de géneros audiovisuales o narrativos, cuya clasificación no es plenamente extrapolable a los videojuegos, pues estos añaden la interacción entre sus múltiples características. De ahí que los videojuegos puedan clasificarse de forma más precisa según criterios diversos, como el *género interactivo* (en nuestro caso, acción-aventura, como podría ser también combate, carreras, simulación, estrategia, etc.), el modo de juego (1.ª o 3.ª persona, o mixto, e individual o multijugador) y el medio (bélico, deportivo, fantástico, terror, etc.), pero no exactamente según el género narrativo tradicional (Wolf 2005; Mejías-Climent 2019: 51-66).

sin restricción alguna, con interacción plena, cuando se trata de conversaciones de menor relevancia. Por ello, no hay un patrón claro en esta situación de juego cambiante, sino que los cinco tipos de ajuste pueden apreciarse en ellos.

Este análisis cuantitativo representa un punto de partida para continuar explorando en corpus de mayor tamaño. Dadas las tendencias que se apuntan, sería interesante, en primer lugar, ampliar el corpus a más títulos de acción-aventura para comprobar si las tendencias se afianzan. Del mismo modo, podrían analizarse juegos pertenecientes a otros géneros interactivos para comprobar si hay tendencias similares o completamente diferentes.

En segundo lugar, en una selección de videojuegos de acción-aventura podría introducirse la variable del medio en el corpus y comprobar si las tendencias en el uso de las sincronías varían, además, en función del medio en el que se sitúen los videojuegos de acción-aventura (histórico, actual, fantástico o de superhéroes, etc.), pues puede suponerse, a partir de los datos apuntados, cierta relación entre las restricciones del videojuego y la situación contextual y narrativa.

5.2. Datos cualitativos extraídos de las entrevistas con los agentes

Por otra parte, como se ha explicado, se quiso comprobar si existe una intención explícita por parte de los traductores o de algún agente de la cadena de traducción y doblaje de estos tres videojuegos a la hora de aplicar distintas sincronías en el texto doblado al español. De este modo, se recurrió a una triangulación metodológica de las herramientas que aportó resultados complementarios (Taylor y Marchi 2018: 7), es decir, que completan la información cuantitativa que se ha obtenido con el análisis empírico.

Las entrevistas semiestructuradas arrojaron luz sobre cada uno de los agentes del proceso de traducción y doblaje (a saber: gestores del proyecto, traductores, revisores, actores de doblaje, directores artísticos y técnicos de sonido) y las tres grandes etapas en las que se lleva a cabo el proceso: preparación y gestión, traducción y revisión lingüística y, por último, doblaje en sala.

En los tres videojuegos, el material que se recibe en la primera etapa se reduce a hojas de cálculo con las cadenas de texto que compondrán los contenidos doblados finalmente en sala, además de cierto material documental

sobre el videojuego (argumento, personajes y características, mecánicas de juego, especificaciones técnicas, contexto y algunas capturas de pantalla en determinados casos, entre alguna otra información).

Dado el material disponible y tal y como comparten gestores, traductores y revisores, no es posible aplicar sincronías específicas de doblaje en un texto para el cual aún no existen ni siquiera vídeos a los que ajustar la traducción. Por ello, en las etapas de preparación y traducción, las sincronías del doblaje se limitan a restricciones de espacio en función del tipo de cadena, identificadas no exactamente como situaciones de juego, según la terminología investigadora que aquí empleamos, sino, simplemente, como contenido de audio y cinemática (la traducción debe asemejarse lo máximo posible a la extensión del original) y texto en pantalla (las restricciones dependerán exactamente del espacio que se indique en caracteres o palabras, pues se trata de texto que aparecerá en la interfaz del juego).

En la etapa de doblaje en sala es cuando se identifican los cinco tipos de ajuste que analizamos en nuestro corpus, aunque, curiosamente, parece recaer sobre el último agente del proceso, el técnico de sonido, la principal responsabilidad de ajustar al máximo posible las ondas de audio dobladas a las originales, según la situación de la que se trate. Cabe mencionar que tampoco en esta fase se dispone de vídeos finales para los audios que se locutan, aunque sí, en algún caso, pudieron acceder a vídeos de captura de movimiento o *motion capture* (Kines 2000), o bien, a vídeos degradados para tener una mejor idea de las condiciones en las que se dan ciertas conversaciones, sin poder disponer de los vídeos definitivos.

Los directores artísticos son conscientes de la importancia de ajustar más o menos las ondas de audio a las originales en función del tipo de texto del que se trate, aunque, de nuevo, no emplean exactamente la terminología de las situaciones de juego clasificadas según tareas, acción, diálogos y cinemáticas, sino, más bien, las califican de acción y bloques de onomatopeyas, y escenas cinemáticas, y sí se diferencian los diálogos en RTR. La aplicación de las sincronías parece responder, en buena medida, sobre todo a la experiencia previa en el sector, con proyectos similares, de actores, directores y técnicos de sonido.

La referencia básica en sala de doblaje son las ondas de audio, al no existir vídeo, como sí se tiene al doblar películas o series. El tipo de ajuste

lo marcan los directores de doblaje según el nivel de restricción del texto: la acción y las onomatopeyas han de asemejarse a las ondas originales con cierto margen (ajuste temporal); en algunos casos, las restricciones son más estrictas (temporal exacto) y, si son cinemáticas reconocidas o diálogos pertenecientes a estas, se procura reproducir con absoluta precisión la onda de audio original, incluyendo pausas, entonación específica o incluso articulación labial (ajuste sonoro y labial).

De nuevo, según su experiencia y el tipo de onda de audio, los técnicos de sonido terminan por ajustar las ondas dobladas al máximo y devuelven a los encargados de la gestión del proyecto una estructura de carpetas y audios idéntica a la que recibieron, pero nunca llegan a ver el resultado final del videojuego doblado y completo, como sí sucede en películas o series, en las que el producto meta sale del estudio de doblaje tal y como se comercializará.

Toda esta información, además de completar los datos de los que ya disponíamos, corrobora la cierta relación que se ha detectado en el corpus entre situaciones de juego y tipos de ajuste, aunque la terminología empleada entre el análisis y estudios anteriores y los profesionales no parece ser exactamente la misma en cuanto a situaciones de juego. Sí se reconocen los cinco tipos de ajuste analizados, aunque, en el caso de los tres videojuegos que componen el corpus, no serán nunca responsabilidad directa del traductor, que está al comienzo de la cadena de traducción y doblaje en videojuegos y solo dispone de texto, nunca de imagen, a diferencia del doblaje cinematográfico.

6. Conclusiones

En estas páginas se ha expuesto de forma sintética el análisis de un corpus multimodal (Soffritti 2018) compuesto por videojuegos de acción-aventura que se ha confeccionado expresamente para responder a la pregunta de qué tipos de ajuste se emplean en cada situación de juego. Como se ha explicado, al tratarse de un corpus de estas características, la complejidad en el diseño y la gestión del material es mayor, puesto que se pretende registrar y dar cuenta no únicamente de lo que sucede en el código lingüístico, sino, más bien, en cada segmento (en este caso, situación de juego) analizado en función de su configuración audiovisual completa, transmitida a través de distintos modos semióticos (Kress y Van Leeuwen 2001). En concreto, nos hemos centrado

en el tipo de ajuste que se emplea en el doblaje de cada una de las cuatro situaciones de juego que representan las unidades de análisis del corpus.

El análisis empírico ha arrojado datos que apuntan a una cierta relación entre las situaciones de juego, que implican diversas formas de interacción con el usuario, y las sincronías del doblaje en videojuegos, concebidas como distintos niveles de restricción. En concreto, parece aplicarse siempre el ajuste libre para las tareas y el ajuste labial en las cinemáticas, con algunas voces en *off* con ajuste libre. La acción de juego parece depender de varios factores de la configuración del propio juego y del medio en el que se sitúe, pues la preferencia por el ajuste está dividida entre el temporal y el libre. Por último, los diálogos no presentan ninguna preferencia clara, tratándose de una situación híbrida cuya configuración depende por completo de cada juego.

Estos datos se han completado con la información cualitativa extraída de entrevistas semiestructuradas con los principales agentes involucrados en la cadena de traducción y doblaje de los tres videojuegos estudiados. Los resultados apuntan a que las sincronías dependen sobre todo de la última fase de la producción, la locución en sala, y se asocian con distintos momentos del videojuego no identificados como situaciones de juego, sino diferenciados entre acción y onomatopeyas, y algunos diálogos y cinemáticas.

Cabe señalar, como se ha ido mencionando, que este análisis aún ofrece varias limitaciones y representa tan solo un punto de partida para futuras ampliaciones del corpus. Con respecto a las limitaciones, se identifican algunas de las habituales de un corpus situado en el ámbito de la TAV (Baños, Bruti y Zanotti 2013): el registro ha de ser en vídeo apoyado en hojas de cálculo, de forma que se genera una cantidad ingente de datos que hay que administrar convenientemente y que, por el momento, no pueden analizarse mediante ningún tipo de *software*, más allá de emplear filtros en las hojas de cálculo creadas. En este caso, la segmentación del texto se ha basado en situaciones de juego, cuya alineación entre la versión original y la versión meta puede no ser del todo precisa según el nivel de aleatoriedad que presente el juego como producto interactivo. Esta limitación, no obstante, representará una proporción muy pequeña en corpus de gran tamaño.

Además de ampliar el corpus con videojuegos pertenecientes al mismo género interactivo, la acción-aventura, sería enriquecedor confeccionar

corpus amplios en otros géneros de videojuegos, de forma que pueda identificarse si en ellos varían o se mantienen las tendencias en el uso de las sincronías del doblaje. De hecho, un estudio de caso preliminar que se está realizando en un videojuego perteneciente al subgénero interactivo de la aventura gráfica (Mejías-Climent 2020) ya apunta a que pueda haber ciertas diferencias, no solo en el uso de las sincronías, sino en la cantidad de situaciones de juego en las que se estructura el corpus, según la naturaleza del videojuego.

Por último, cabe señalar que, para este tipo de estudios de corpus, sería de enorme ayuda contar con las hojas de cálculo originales que se emplearon para las traducciones, pero esta disponibilidad, en especial en el ámbito de los videojuegos, ha de descartarse por completo, dados los estrictos acuerdos de confidencialidad bajo los que las desarrolladoras y empresas de localización trabajan. La cooperación entre industria y academia, una vez más, sería enormemente beneficiosa para asegurar corpus de calidad sobre los que llevar a cabo estudios cuyos resultados fueran prácticos, a su vez, para el mundo profesional. Dada esta carencia, por ahora, el diseño del corpus estudiado empleando situaciones de juego para segmentarlo pretende ofrecer una metodología adecuada a la idiosincrasia multimodal e interactiva del producto, que se irá comprobando con futuras ampliaciones y variaciones en el análisis.

Referencias bibliográficas

BAKER, Mona. (1996) "Corpus-based translation studies: The challenges that lie ahead." En: Somers, Harold (ed.). *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager*. Ámsterdam y Filadelfia: John Benjamins, pp. 175-86.

BAÑOS, Rocío, Silvia Bruti y Serenella Zanotti. (2013) "Corpus linguistics and Audiovisual Translation: In search of an integrated approach." *Perspectives: Studies in Translatology* 21:4, pp. 483-90.

BERNAL MERINO, Miguel Ángel. (2006) "On the Translation of Video Games." *JoSTrans - The Journal of Specialised Translation* 6. Versión electrónica: <https://www.jostrans.org/issue06/art_bernal.php>.

BERNAL MERINO, Miguel Ángel. (2015) *Translation and localisation in video games Making entertainment software global*. Nueva York: Routledge.

BERNARDINI, Silvia y Dorothy Kenny. (2020) "Corpora." En: Baker, Mona y Gabriela Saldanha (eds.). *Routledge Encyclopedia of Translation Studies*. Londres y Nueva York: Routledge, pp. 110-15.

CADIEUX, Pierre y Bert Esselink. (2004) "GILT: Globalization, Internationalization, Localization, Translation." *Globalization Insider* 11:1.5. Versión electrónica: <http://www.i18n.ca/publications/GILT.pdf>.

CAPLE, Helen. (2018) "Analysing the multimodal text." En: Taylor, Charlotte y Anna Marchi (eds.). *Corpus Approaches to Discourse. A Critical Review*. Londres y Nueva York: Routledge, pp. 85-109.

CHAUME, Frederic. (2004a) *Cine y traducción*. Madrid: Cátedra.

CHAUME, Frederic. (2004b) "Synchronization in dubbing: A translational approach." En: Orero, Pilar (ed.). *Topics in Audiovisual Translation*. Ámsterdam y Filadelfia: John Benjamins, pp. 35-52.

CHAUME, Frederic. (2007) "Quality standards in dubbing: a proposal". *Tradterm* 13, pp. 71-89.

CHAUME, Frederic. (2012a) *Audiovisual translation: dubbing*. Manchester: St. Jerome Publishing.

CHAUME, Frederic. (2018a) "An overview of audiovisual translation: Four methodological turns in a mature discipline." *Journal of Audiovisual Translation* 1:1, pp. 40-63.

CHAUME, Frederic. (2018b) "Is audiovisual translation putting the concept of translation up against the ropes?" *JosTrans – The Journal of Specialised Translation* 30, pp. 84-104. Versión electrónica: <http://jostrans.org/issue30/art_chaume.php>.

CROSIGNIANI, Simone y Fabio Ravetto. (2001) "Localizing the Buzz! Game Series (Or how to successfully implement transcreation in a multi-million seller video game)." *Trans. Revista de Traductología* 15 (Special issue on games localization), pp. 29-38. Versión electrónica: <http://www.trans.uma.es/pdf/Trans_15/29-38.pdf>.

DÍAZ CINTAS, Jorge y Gunilla Anderman. (2009) *Audiovisual Translation: Language Transfer on Screen*. Houndmills, Hampshire y Nueva York: Palgrave Macmillan.

DIETZ, Frank. (2007) "How Difficult Can that be? The Work of Computer and Video Game Localisation." *Tradumàtica* 5. Versión electrónica: <http://www.fti.uab.es/tradumatica/revista/num5/articles/04/04art.htm>.

ENSSLIN, Astrid. (2012) *The Language of Gaming*. Nueva York: Palgrave Macmillan.

ESSELINK, Bert. (2000) *A Practical guide to software localization*. Amsterdam: John Benjamins.

FERNÁNDEZ COSTALES, Alberto. (2012) "Exploring Translation Strategies in Video Game Localisaton." *MonTI* 4, pp. 385-408.

FERNÁNDEZ TORNÉ, Anna. (2007) "Anàlisi de la localització de Codename: Kids Next Door - Operation V.I.D.E.O.G.A.M.E." *Tradumàtica* 5. Versión electrónica: <http://www.fti.uab.es/tradumatica/revista/num5/articles/08/08art.htm>.

GAMBIER, Yves. (2003) "Screen transadaptation: perception and reception." *The Translator* 9:2, pp. 171-89.

HERMANS, Theo. (2020) "Descriptive translation studies." En: Baker, Mona y Gabriela Saldanha (eds.). *The Rotutledge Encyclopedia of Translation Studies*. Londres y Nueva York: Routledge, pp. 143-47.

HURTADO ALBIR, Amparo. (2011) *Traducción y Traductología. Introducción a la traductología*. 5.ª edición. Madrid: Cátedra.

JAKOBSON, Roman. (2000) "On Linguistic Aspects of Translation." En: Venuti, Lawrence (ed.). *The Translation Studies Reader*. Londres y Nueva York: Routledge, pp. 111-18.

JEWITT, Carey, Jeff Bezemer y Kay O'Halloran. (2016) *Introducing Multimodality*. Oxon y Nueva York: Routledge.

JIMÉNEZ-CRESPO, Miguel A. (2009) "El uso de corpus textuales en localización." *Tradumàtica* 7. Versión electrónica: <http://www.fti.uab.cat/tradumatica/revista/num7/articles/05/05.pdf>.

JIMÉNEZ-CRESPO, Miguel A. (2013a) "Crowdsourcing, Corpus Use, and the Search for Translation Naturalness: A comparable corpus study of Facebook and non-translated social networking sites." *Translation and Interpreting Studies* 8:1, pp. 23-49.

JIMÉNEZ-CRESPO, Miguel A. (2013b) *Translation and web localization*. Oxon: Routledge.

JIMÉNEZ-CRESPO, Miguel A. (2015) "Testing Explicitation in Translation: Triangulating corpus and experimental studies." *Across Languages and Cultures* 16:2, pp. 257-83.

JIMÉNEZ-CRESPO, Miguel A. (2020) "Localization." En: Baker, Mona y Gabriela Saldanha (eds.). *The Routledge Encyclopedia of Translation Studies*. Londres y Nueva York: Routledge, pp. 299-304.

KAINDL, Klaus. (2013) "Multimodality and translation." En: Millán, Carmen y Francesca Bartrina (eds.). *The Routledge handbook of translation studies*. Londres: Routledge, pp. 257-69.

KENT, Steven L. (2001) *The Ultimate History of Video Games*. Nueva York: Three Rivers Press.

KINES, Melianthe. (2000) "Planning and Directing Motion Capture for Games." *Gamasutra* (19 enero). Versión electrónica: <https://www.gamasutra.com/view/feature/131827/planning_and_directing_motion_.php>.

KRESS, Gunther y Theo Van Leeuwen. (2001) *Multimodal discourse. The modes and media of contemporary communication*. Londres y Nueva York: Oxford University Press.

KVALE, Steinar y Svend Brinkmann. (2009) *Interviewing: learning the craft of qualitative research interviewing*. Los Ángeles: Sage.

LAVIOSA, Sara. (2002) *Corpus-based translation studies: Theory, findings, applications*. Ámsterdam y Atlanta: Rodopi.

LAVIOSA, Sara. (2012) "Corpora and Translation Studies." En: Hyland, Ken; Chau Meng Huat y Michael Handford (eds.). *Corpus Applications in Applied Linguistics*. Londres y Nueva York: Continuum International, pp. 67-83.

LÓPEZ REDONDO, Isaac. (2014) *¿Qué es un videojuego? Claves para entender el mayor fenómeno cultural del siglo XXI*. Sevilla: Héroes de Papel.

LOUREIRO PERNAS, María. (2007) "Paseo por la localización de un videojuego." *Tradumàtica* 5. Versión electrónica: <http://www.fti.uab.es/tradumatica/revista/num5/articles/03/03art.htm>.

MAIETTI, Massimo. (2004) *Semiotica dei videogiochi*. Milán: Unicopli.

MALAMATIDOU, (Sofia) 2018. *Corpus Triangulation. Combining Data and Methods in Corpus-Based Translation Studies*. Londres y Nueva York: Routledge.

MANGIRON, Carme. (2010) "The Importance of not being Earnest: Translating Humour in Video Games." En: Chiaro, Delia (ed.). *Translation, Humour and the Media*. Londres: Continuum, pp. 89-107.

MANGIRON, Carme. (2017) "Research in game localisation." *The Journal of Internationalization and Localization* 4:2, pp. 74-99.

MATA PASTOR, Manuel. (2005) "Localización y traducción de contenido web." En: Reineke, Detlef (ed.). *Traducción y localización: mercado, gestión y tecnologías*. Las Palmas: Anroart Ediciones, pp. 187-252.

MAXWELL-CHANDLER, Heather y Stephanie O'Malley Deming. (2012) *The Game localization handbook*. 2.ª edición. Sudbury, Massachussets: Jones.

Mejías-Climent, Laura. (2017) "Multimodality and dubbing in video games: A research approach." *Linguistica Antverpiensia, New Series: Themes in Translation Studies* 17, pp. 99-113. Versión electrónica: <https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/463>.

Mejías-Climent, Laura. (2018) "El ajuste en videojuegos: el doblaje de *Assassin's Creed Syndicate*." *Trans - Revista de traductología* 22, pp. 11-30.

Mejías-Climent, Laura. (2019) *La sincronización en el doblaje de videojuegos. Análisis empírico y descriptivo de los videojuegos de acción-aventura*. Castellón: Universitat Jaume I. Tesis doctoral inédita.

Mejías-Climent, Laura. (2020) "La evolución de las tecnologías en la confluencia de la interacción y el cine. El doblaje en una aventura gráfica." *inTRAlinea. online Translation Journal* 22. Versión electrónica: <http://www.intralinea.org/archive/article/2509>.

Méndez González, Ramón y José Ramón Calvo-Ferrer. (2017) *Videojuegos y [para]traducción: aproximación a la práctica localizadora*. Granada: Comares.

Müller Galhardi, Rafael. (2014) "Video Games and Fan Translations: A Case Study." En: Mangiron, Carme; Minako O'Hagan y Pilar Orero (eds.). *Fun for all: translation and accessibility practices in video games*. Berna: Peter Lang, pp. 175-95.

Muñoz Sánchez, Pablo. (2017) *Localización de videojuegos*. Madrid: Síntesis.

Neves, Joselia. (2005) *Audiovisual translation: Subtitling for the deaf and hard of hearing*. Londres: Roehampton University. Tesis doctoral.

O'Hagan, Minako y Carme Mangiron. (2013) *Game localization: translating for the global digital entertainment industry*. Ámsterdam: John Benjamins.

Van Oers, Annelies. (2014) "Translation Strategies and Video Game Translation: A Case Study of Beyond Good and Evil." En: Mangiron, Carme; Minako O'Hagan y Pilar Orero (eds.). *Fun for all: translation and accessibility practices in video games*. Berna: Peter Lang, pp. 129-48.

Olohan, Maeve. (2004) *Introducing corpora in translation studies*. Londres: Routledge.

Orero, Pilar (ed.) (2004). *Topics in Audiovisual Translation*. Ámsterdam y Filadelfia: John Benjamins.

Pavesi, Maria. (2018) "Corpus-based audiovisual translation studies. Ample room for development." En: Perez-González, Luis (ed.). *The Routledge Handbook of Audiovisual Translation*. Londres y Nueva York: Routledge, pp. 315-333.

Pérez-González, Luis (2020). "Multimodality." En: Baker, Mona y Gabriela Saldanha (eds.). *The Routledge Encyclopedia of Translation Studies*. Londres y Nueva York: Routledge, pp. 346-351.

Pruys, Guido Marc. (2009) *Die Rhetorik der Filmsynchronisation*. Colonia: Eigenverlag.

Pujol Tubau, Miquel. (2015) *La representació de personatges a través del doblatge en narratives transmèdia. Estudi descriptiu de pel·lícules i videojocs basats en El senyor dels anells*. Barcelona: Universitat de Vic - Universitat Central de Catalunya. Tesis doctoral.

Pym, Anthony. (2014) *Exploring Translation Theories*. Londres y Nueva York: Routledge.

Soffritti, Marcelo. (2018) "Multimodal Corpora and Audiovisual Translation Studies." En: Pérez-González, Luis (ed.). *The Routledge Handbook of Audiovisual Translation*. Londres y Nueva York: Routledge, pp. 334-349.

Taylor, Charlotte y Anna Marchi (eds.) (2018) *Corpus Approaches to Discourse. A Critical Review*. Londres y Nueva York: Routledge.

Toury, Gideon (1995) *Descriptive translation studies-- and beyond*. Filadelfia: John Benjamins.

Vázquez Rodríguez, Arturo (2018) *El error de traducción en la localización de videojuegos. Estudio descriptivo y comparativo entre videojuegos* indie *y no* indie. Valencia: Universitat de València. Tesis doctoral inédita.

Wolf, Mark J. P. (2005) *The Medium of the Video Game*. 3.ª edición. Austin: The University of Texas Press.

Zanettin, Federico. (2014) *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. Oxon y Nueva York: Routledge.

NOTA BIOGRÁFICA / BIONOTE

Laura Mejías-Climent es doctora por la Universitat Jaume I (UJI) y licenciada en Traducción e Interpretación por la Universidad Pablo de Olavide. Trabaja en la UJI como docente e investigadora mediante una beca posdoctoral y forma parte del grupo TRAMA. Ha cursado másteres en Traducción Audiovisual, Traducción y Nuevas Tecnologías y enseñanza secundaria e idiomas (MAES). Además, ha impartido clases en la Universidad Pablo de Olavide y en ISTRAD, como profesora del Máster en TAV (Universidad de

Cádiz) y los másteres en Traducción Especializada (Universidad Internacional Menéndez Pelayo). Participa en el Experto en Traducción y Localización de Videojuegos (ISTRAD) y en el Máster en TAV de la Universidad Europea de Valencia. Trabajó en la University of St. Thomas, en Estados Unidos, mediante una beca Fulbright, y reúne experiencia como traductora profesional y como gestora de proyectos de traducción, especializada en la traducción audiovisual.

Laura Mejías-Climent holds a PhD in Translation by the Universitat Jaume I (UJI) and a Bachelor's degree in Translation and Interpreting by the Universidad Pablo de Olavide (UPO). She teaches at UJI, where she also works as a postdoctoral researcher and member of the research group TRAMA. She holds three Master's Degrees: in Audiovisual Translation, Translation and New Technologies, and Secondary Education and Languages (MAES). She taught at UPO and ISTRAD, as a lecturer for the Master's Degree in AVT (Universidad de Cádiz) and the master programs in Specialized Translation (UIMP). She participates in the Expert Diploma in Video Game Translation and Localization (ISTRAD) as well as the Master's Degree in AVT at Universidad Europea de Valencia. She also taught in the USA thanks to a Fulbright scholarship, and worked as a translation project manager and a professional translator, specialized in the field of audiovisual translation.

Recibido / Received: 20/05/200
Aceptado / Accepted: 10/08/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.08

Para citar este artículo / To cite this article:
Moreno-Pérez, Leticia & Belén López-Arroyo. (2021) "Atypical corpus-based tools to the rescue: How a writing generator can help translators adapt to the demands of the market." En: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection*. *MonTI* 13, pp. 251-279.

ATYPICAL CORPUS-BASED TOOLS TO THE RESCUE: HOW A WRITING GENERATOR CAN HELP TRANSLATORS ADAPT TO THE DEMANDS OF THE MARKET[1]

LETICIA MORENO-PÉREZ
leticia.moreno@uva.es
Universidad de Valladolid - ACTRES

BELÉN LÓPEZ-ARROYO
mariabelen.lopez@uva.es
Universidad de Valladolid - ACTRES

Abstract

Corpus studies have become an undisputed aid for the evolution of translation, transferring knowledge from the academia to develop tools that have invaluably helped the profession. Nevertheless, the demands of the market require translators to improve their efficiency in order to adapt to its hectic pace. The aim of this paper is to present a possible solution through the use of corpus-based tools that are usually neglected in translation: writing aids. First, the reality of the translation market will be studied to understand the current context and translators' needs. Then, we will analyze some of the existing tools derived from corpus studies available for translators, both the most and less usual. Finally, we will focus on a booming sector of the market, that of

1. This study was carried out within the research project *Producción textual bilingüe semiautomática inglés-español con lenguajes controlados: parametrización del conocimiento experto para su desarrollo en aplicaciones web 2.0 y 3.0*, financially supported by Ministerio de Ciencia e Innovación since 2016 (FFI2016-75672-R).

oenology, to exemplify how one of the less typical tools, the writing generator, may be helpful for translators in terms of cost, time, and quality given the current demands.

Keywords: Corpus-based tools; Knowledge transfer; Translation market; Writing generator; Oenology.

Resumen

Los estudios de corpus han supuesto una inestimable ayuda para la evolución de la traducción, al generar herramientas que facilitan a los profesionales su labor mediante la transferencia de conocimiento desde la investigación. Pero el mercado exige un nuevo ritmo a los traductores, quienes necesitan mejorar su eficiencia. El objetivo de este estudio es presentar como posible solución unas herramientas basadas en corpus habitualmente desatendidas en traducción: los asistentes de escritura. Tras analizar la realidad del mercado para comprender el contexto actual y las necesidades de los traductores, expondremos algunas de las herramientas derivadas de los estudios de corpus de que estos disponen, tanto habituales como atípicas; por último, nos centraremos en un sector del mercado en auge, la enología, para ejemplificar cómo el generador de escritura puede resultar de ayuda para los traductores en aspectos como costes, tiempo y calidad, de acuerdo con las demandas actuales.

Palabras clave: Herramientas basadas en corpus; Transferencia de conocimiento; Mercado de traducción; Generador de escritura; Enología.

1. Introduction

Corpus-based studies became an undisputed turning point for the evolution of pure and applied translation studies (Laviosa 2002: 4) by incorporating, among others, quantitative methods of research to describe patterns of behavior in discourse. The combination of quantitative and generalizable data with qualitative insights into dimensions of discourse has provided a fruitful alliance for researchers (Marchi & Taylor 2018: 4). However, the global market we are in needs this knowledge to be transferred in the form of useful and usable tools and aids[2] (Rabadán Álvarez 2008: 105) to help translators improve their efficiency in their daily profession.

2. In this paper, *usefulness* refers to "the extent to which tools (technological, conceptual or otherwise) are relevant to the actual needs of a user" (Rabadán Álvarez 2008: 106; Landauer 1995: 4), whereas *usability* is "the extent to which a product can be used by

But, as it has been reported (Sinclair 2004, Rabadán Álvarez 2005-2008; 2008, among others), usually the academic research community does not think they have to supply solutions; it is the translator who has to derive them according to the researcher's conclusions. In this sense, Rabadán Álvarez already reported that commercially available tools and aids were not as popular or widely used as they might be expected, and the reason for this is that the user feels they are not useful since they do not supply solutions to problems (2008: 105). This situation still persists nowadays as, in the long run, translators end up developing their own research protocols and custom-made strategies to satisfy the demands of the translation market. Even though these demands are in constant change, this misunderstanding between the user needs and the researcher work should necessarily be solved in this era of globalization, technology and ICTs.

To change this situation, it is necessary to implement translation methodologies that increase value for money by supplying translators with tools that satisfy the needs of the market, helping them be more accurate and quicker; this would translate into lower rates clients are willing to pay for, without diminishing the translators' job and life quality. The aim of this paper is to explore paths to reach this result by applying corpus-based research products that already exist but are not originally designed for translators, and therefore are rarely used by them: writing aids. These neglected corpus-based tools ensure a high level of accuracy and could make their work more efficient in the terms mentioned above. To prove their usefulness, first we will analyze the reality of the translation market, to see which demands might be challenging and need to be improved. Then, we will review the typical corpus-based translation tools and we will contrast them with atypical corpus-based tools to help define how the latter could be more helpful for the current market demands. Finally, we will illustrate the pertinence and applicability of a specific atypical corpus-based tool, a writing generator, in a specific field of the market with a high demand of translation, the flourishing sector of oenology.

specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (Quesenbery 2001; also Kreitzberg & Little 2009).

2. The reality of the translation market with a focus on the business sector

Since the beginning of the 21st century, researchers have used different adjectives to describe the translation market, but all of them seem to agree on its changing nature, as translation is "highly dependent on external factors" (Sosoni & Rogers 2013: 7). Back in 2005, the research group Aula.int from the University of Granada, Spain, provided an overview of the translation market taking into account past trends and future prospects, and defined the following characteristics (Aula.int 2005: 133-135):

- Global, as the growth of international trade has resulted in an increased necessity of international communication, thus translation.
- Decentralized, as there are no geographical boundaries, increasing competition between translators who find it more complicated to focus their work on a certain language combination, field of expertise, or market.
- Specialized, not only in 'conventional' fields, such as legal, financial, medical, etc., but also in areas that emerged later, such as localization, emerging technologies, digital contents, etc.
- Dynamic, as new fields and translation aids emerge, so translators need to adapt and update their knowledge on a regular basis.
- Virtual, as relationships between translators and clients, agencies, experts or colleagues are developed mostly online.
- Demanding, as, on the one hand, deadlines are becoming shorter and the rates lower, resulting in translators being forced to accept a higher workload than desired. On the other hand, clients' demands in terms of quality and methodology to be followed (particular terminology, specific software, etc.) are increasingly restricting translators' freedom and working systems.

This group's perspective has been confirmed later on by researchers such as Dunne (2012) or Moorkens (2017), who noted how these characteristics have made the translation market evolve into a project-driven industry, in which stakeholders associate for a specific job and do not usually build strong client-provider relationships. In these terms, translation is progressively seen, especially by clients, as a commodity that a number of providers can supply

based on "availability, productivity, cost, or perceived translation quality" (Moorkens 2017: 469), leading to an even higher competition within the market.

Nevertheless, competition among translators might not be the most concerning issue arising from this trend. Moorkens (2017: 467) highlights two important external threats: the popularization of machine translation and postediting, and the emergence of non-professional translation practices. The former trend has taken over a significant part of the market (Robinson 2020: 32), as it is usually cheaper and faster, to the point that some researchers believe that "the translator's function can be expected to shift to linguistic postediting" (Pym 2013: 487) in the future. In some concerning cases, translation clients resort to online automatic translation and postediting to save time and money leaving the quality discussion aside, even when research has shown that decisions along this line can entail economic losses (Robinson 2020; Sosoni & Rogers 2013; Hennecke 2017). As regards non-professional translation, this trend is quite diverse in nature, but involves clients resorting to untrained mediators to perform translation tasks. It was confirmed as the most concerning perceived threat for the Spanish translation market stakeholders by Rico Pérez and García Aragón (2016: 35). Some may argue that the use of non-professional translation is limited to certain fields and contexts, as is the case of crowd-sourced translation, a common one when translating [non-official] social media or audiovisual contents (Katan 2016: 379). However, this reality becomes an important issue when it affects official communications, especially in sectors where success highly depends on translation, as is the case of business.

Translation in the business sector is still in continuous expansion given the dominant paradigm of globalization, which produces an increasing number of texts requiring linguistic mediation (Morón Martín & Medina Reguera 2016: 227). However, companies tend to resort to employees with knowledge of foreign languages to translate documentation from and into other languages, instead of specifically hiring professional translators (Varona 2002; Mayoral 2006; Aguayo 2013; Albuquerque & Costa 2018). The reason behind this practice is that companies see translation as a "'non-core' business" (Sosoni & Rogers 2013: 7), just "a means to an end –a mediation resource to meet goals in a business transaction between client and

supplier" (Albuquerque & Costa 2018: 150), so minimizing costs by resorting to employees seems only logical. But is it impossible to make quality and economy meet? If translators make their work more cost-efficient, the distance between the numbers of businesses and the numbers of translators may narrow, and companies may stop seeing translation as a burden. Therefore, translators need the tools to help them be more efficient, and this is where the results of academic research come into play.

3. Helping translators adapt to the market: typical and atypical corpus-based tools

As mentioned in the previous section, translation efficiency does not only have to do with time and money, but also with quality. Clients, specifically those in specialized sectors, require target texts that comply with the standards and expectations of the target community. This usually involves specific terminology and genres which are known and shared by the given sector, members of which expect to receive texts–translated or not–that reflect a certain language and content. For that purpose, translators

> need to be completely sure that the unit or expression they are employing in a specific target text is the best option to translate the source term or expression, not only regarding meaning but also register, style, geographical variant, etc. (Durán-Muñoz & Corpas Pastor 2020: 164).

If there exists a methodology that is able to meet these requirements, that is Corpus Linguistics, as it is founded on the real use of language. In fact, it is a widespread methodology both in translation theory and practice: according to previous research, most translators build their own corpora (comparable and parallel[3]) as part of their documentation and translation process (Durán-Muñoz 2012: 164; Zanettin 2013: 27); they consider the use of this methodology very effective despite the amount of time spent. Its appropriateness is also supported by the fact that a relevant number of studies have applied this methodology to the research on linguistics and translation with relevant results. However, although the quantitative nature

3. In this paper we understand comparable corpora as those multilingual corpora written in their original language, while parallel corpora are those multilingual corpora that are made up of texts in a source language and their translations.

of Corpus Linguistics improves the analysis and the translation as a product, it has to be combined or integrated with the qualitative techniques of other methodologies which consider the context. The reliability of statistical analysis, necessary for generalization, is combined with the precision and richness of qualitative analysis. As Marchi & Taylor state (2018: 9), "mixing methods is a form of triangulation" because they allow not only the researcher but also the professional translator to look into the data from many different windows (Baker 2018; Malamatidou 2018, among others).

Taking into account the aforementioned necessities of translators and the demands of the market, Corpus Translation Studies and Discourse Studies have been a fruitful source of tools that have helped not only translators, but also scientific and professional writers in a foreign language. There is a set of resources derived from corpus-based research that have been typically used by translators to improve efficiency and/or acceptability, especially:

- Lexicographical resources: dictionaries are the classic source of "equivalents and linguistic information" (Durán-Muñoz & Corpas Pastor 2020: 164). Lately, many lexicographers rely on corpora in the process of dictionary creation (Tarp & Fuertes-Olivera 2016) as an aid "for completing the dictionary structures they need when making a real dictionary" (Fuertes-Olivera 2012: 51).
- Ontologies: these knowledge bases are useful resources of terminology, more specifically in specialized fields, and many times their construction is based on a computerized analysis of a collection of texts (Carrero & Gomez 2008; Bautista Zambrana 2019). Ontologies, such as Oncoterm or EcoLexicon, are examples of framed-based terminologies created on a model of semantic analysis, based on the creation of lexical templates derived from corpus and dictionary analysis.
- Translation memories: "a specific type of dynamic parallel corpora" that have become "a standard tool of the trade" (Zanettin 2013: 20); they are used or built by translators through the alignment of parallel texts to "easily observe the original and translated segment" (Durán-Muñoz & Corpas Pastor 2020: 164) as a source to assist the translation process.

- Machine translation: "systems which rely largely on corpus-based statistical machine translation techniques" (Zanettin 2013: 20); translators can be both users and victims, as mentioned above (see section 2).
- Integrated corpus annotating and management tools: they include different tools, from POS concordance searches to semantic relations, to analyze DIY or existing specialized and reference corpora, e.g. Lancsbox, Termostat or Lextutor, to name a few.
- Online corpora: large and reliable compilations of texts that are accessible and manageable online by users, and which provide translators with "a wide range of linguistic and pragmatic information" (Durán-Muñoz & Corpas Pastor 2020: 167). COCA, BNC or Iweb are some of the most popular on-line corpora used by universities and translators containing millions of words divided by genres.
- Web crawlers: "tools that employ the Internet (the Web) as a direct source of information to launch linguistic queries or compile corpora automatically" (Durán-Muñoz & Corpas Pastor 2020: 167) and Corpus Manager, as search engines or online concordancers. Crawlers such Web BootCat or Webcorp, among others, use the web as a source of automatic corpus compilation, offering as well concordance and KWIC tools to analyze them.

The amount of commonly used tools based on corpus research proves to be large and heterogeneous. However, despite the development of such a number of tools aimed at improving the documentation process, some researchers and professionals have noted a relevant problem: translators are "still forced to consult a plethora of resources during the translation task" (Durán-Muñoz & Corpas Pastor 2020: 163). This does not seem to match the necessities of the market explained in the previous section, as the more time is spent, the less efficient the job. But a possible solution for this problem might be close at hand.

There is a different set of corpus-based tools that are most commonly used by writers in foreign languages, but that are atypical for translators, and which could be extremely useful given the current professional context: writing aids. Some of the most frequently used aids are:

- Writing assistants: these tools provide users with recommendations, mostly regarding terminology (terms, collocations, etc.) or style (grammar, discursive elements, etc.), to help them improve texts written in a foreign language so that the writing sounds natural; these tools vary in terms of operation, as some provide recommendations through the analysis of what has been written so far, while others require the user to launch a specific query. The solutions provided are based on an internal source of the software, as a dictionary or a corpus, and users must choose the option they consider to be the most adequate from the ones displayed and detailed in the interface. Examples of these are the different writing tools of Termium or the Write Assistant launched by Ordbogen.
- Templates: they are models, more specifically skeletal frameworks for given genres or text types. A writing template helps writers organize material and also helps them develop the kinds of sentence, paragraph, and structure that good writers display. Templates do not simply give writers advice on how to write; they show exactly *how to do it*. But while they provide step-by-step guidance in writing a given text type, they take for granted that the template user is fluent in the language being written (López-Arroyo & Roberts 2015: 150). In other words, the template shows how to write a specialized text type but it does not show the user how to write the language in which the text type is written (King, 2007; Supatranont 2012). More specifically, an English writing template is intended for English speakers who are called upon to write a specialized text in a genre with which they are not familiar.
- Writing generators: they are writing applications that allow users to produce full texts in a foreign language following the rhetorical particularities, norms and conventions of a given genre; they guide the user through the format of the genre in question, suggesting full semantic units and phrases, rather than terms or individual elements. The units offered to the user are based on quantitative and qualitative corpus analysis of that specific genre, so the resulting text will not only be correct in grammar, structure and format, but also reflect the particularities of the genre in the language being

used. The ACTRES research group is one of the most prolific on this matter, and has developed a few, as a generator for cheese description or for scientific abstract writing.

In general terms, the added value of these corpus-based tools is that they consist on several tools in one: in writing assistants, the translator can improve and proofread while producing the target text; templates offer a guide to write a particular text type; like writing templates, generators provide step-by-step guidance in writing a given text type. But since generators are in principle intended for non-native speakers of the language of the text, they provide guidance not only in rhetorical structure (text sections) and stock phrases and sentences found in such a text type, but also in the overall vocabulary and structures required for a given text type (López-Arroyo & Roberts 2015: 151). The translator is directly producing a proofread text with the words, style and format that the target community expects. Furthermore, since these tools require users to have a high knowledge of the languages involved, translators are a perfect profile for their proper use.

Given their characteristics, these tools do seem to fit the needs of translators in terms of efficiency. Focusing on specialized texts, as those in the business sector mentioned above, writing generators would probably be the most useful. Specialized genres tend to be more restrictive in terms of terminology and format, so in very closed specific genres, a previously reviewed set of structures would accelerate the process without losing quality. This is something that other translation aids, as translation memories, can also help with. However, generators provide solutions for some of the drawbacks of these other tools: the dubious quality (Doherty 2016: 954) and accuracy (Bowker 2005: 19) of the translations that sometimes feed translation memories should not be a problem in generators, as they are guaranteed by a thorough process of compilation and analysis. Also, the fact that the text is considered as a whole unit in writing generators eliminates the lack of "syntagmatic cohesion" (Pym, 2011: 3) and the problematic consequences of working with split segments (Pym 2013: 496) that sometimes derive from the use of translation memories. Furthermore, writing generators guarantee the representativeness of the specialized genres involved, which are usually

the best paid, so translators using this tool would be earning more money in less time.

This is supported by the fact that a writing generator partially skips the process of previous documentation/specialization of the translator in the subject matter; the tool provides an *ad hoc* corpus for a specific translation task, that is, a compilation aimed at creating a given specialized genre in a specific language pair. When a translator uses this tool, the representativeness and adequacy of the texts have already been checked, and the compilation analyzed and organized by the expert linguists who built the generator. This saves translators the significant amount of time that entails creating their own corpus for the documentation process, since it involves a building phase (including text search, quality check, origin verification, classification of texts, format conversion…) and an analysis phase (including determination of search settings, retrieval of data, selection and checking of candidates, advanced search of unaligned information…). The use of these tools also allows for learning and specializing while producing acceptable texts, that is, while getting paid.

Writing generators can be profitable for translators not only at an efficiency level, it can also be profitable at a competitive level, as translators would be able to offer their services to new markets more quickly. As an example, let us focus on the food sector. Food industries many times start producing a specific set of products, and later on decide to diversify. For example, many wineries in Spain have started producing olive oil recently. In this case, both products share the same or similar text genres, namely tasting notes (Sanz Valdivieso & López-Arroyo 2020:27). What is more, this text genre is common to other food products, such as spirits or cheese. Although each product has its own specificities, they share the same function and target audience, as well as a common rhetorical structure.

By acquiring a set of food-related writing generators as translation aids to ensure the quality and consistency, an important market niche would open for the translator, who could offer his or her services to the same or different companies with similar needs within one sector; this could be especially relevant in industries where there is a lack of specialized language professionals, as is the case of wine or oil (see section 4.1). The same situation would apply to other sectors or fields, as engineering, law, etc.

To illustrate how writing generators can help translators adapt to the reality of the translation market, in the following section we will focus on the example of the wine sector.

4. An atypical corpus-based tool applied to the translation market: a writing generator in the oenology sector

4.1. The market of oenology

In the last decades, the market of oenology has become highly relevant in international trade, and more specifically in Spain, the country we have chosen to illustrate our proposal. This boom has increased the need for multilingual communication in the field, but there are not many language mediators who have become experts. In fact, previous research has shown that translation in Spanish international trade companies in the food and oenology sectors is mostly done by workers within the companies (Ibáñez Rodríguez et al. 2010; Medina Reguera & Álvarez García 2014). However, the importance of this sector in the international market calls for an accurate management of language mediation, as there are millions in profits at risk. A more in-depth analysis of the market of oenology will help visualize its potential, both at an international and a local level.

The 2019 Statistical Report on World Vitiviniculture (International Organisation of Vine and Wine 2019) reveals meaningful figures about the wine industry in 2018: 292m hl of wine were produced worldwide; the consumption of wine reached 246m hl; and the import-export of wine reached 108m hl, which translates into €31,000m. Spain was the country with the largest number of hectares under vine, and was the fourth major grape producer in the world, with 96 per cent of its grape production dedicated to wine. As a result, it was the third major wine producer after Italy and France. Despite being the eighth wine consumer in the world (preceded by the USA, France, Italy, Germany, China, the United Kingdom and the Russian Federation, in this order), Spain did not appear on the list of main importers. Nevertheless, it was the main wine exporter worldwide, with 21.1m hl. An interpretation of the previous data leads to two revealing conclusions. Firstly, the main producers mostly consume their own wine, while the rest of the most important consumers import it. Secondly, as Spain is not

one of the biggest consumers, exporting is the way to monetize its surplus production. This places Spain at a relevant position in the market of oenology at a global scale.

Analysing the market at a local level (Salvador Insúa 2016: 429-435), more than 80 per cent of wine industries in Spain are microenterprises with very limited resources and production capacity to access the international marketplace; this fact hampers the export process, even though companies in this sector heavily depend on foreign trade to survive. Another side of the wine market in Spain is that of tourism (López-Arroyo & Fernández Antolín 2011). In addition to wine production, the offer of many wineries ranges from wine tastings to meals at their restaurants, lodging at their hotels, and even wine-based beauty treatments at their spas. These vacation packages are not only addressed to tourists at a national level, but also worldwide.

From both the international and national contexts, it can be inferred that language mediation should be a key element in the market of oenology, and more specifically in Spain, given the importance of this country in the market at a global scale. However, the specific situation of the sector in the country could encourage a reduction of costs in the commonly non-core use of translation, which would explain the aforementioned trend of resorting to employees for language mediation. But, why not investing in translation when studies confirm that one out of ten companies have suffered order or project cancellations due to the lack of foreign language expertise (Hennecke 2017: 23)? These companies are putting their income at stake, confirming the need in this sector for efficient translation that is acceptable for the target community, affordable, quick and of high quality. At this point, it can be observed that the situation coincides with that of the translation market in general; therefore, it is necessary to find a way to make translation more cost-effective, thus attractive for companies who actually need it. A writing generator for the oenology sector could help both parts meet.

4.2. The ACTRES oenology writing generator

As introduced in section two, a generator is a tool that helps a non-native speaker produce a given type of specialized text in another language. As main general characteristics, we can highlight: (i) they are doubly language-bound,

since they are designed to be used by speakers of another language, and therefore have to take the users' native language into account; in the case described in this paper, for instance, the generator is designed to be used by speakers of Spanish as a first language who need to write in English, which means that the generators take the user's native language as a starting point; (ii) they are genre bound; and (iii) they offer semantic units and phrases based on quantitative and qualitative corpus analysis, which guarantees that the resulting text is both correct and acceptable to the target specialized community. These characteristics make generators a reliable and accurate tool for translators to produce specialized texts in a foreign language, the main need of exporting companies, in this case.

To focus on the sector of wine, we will specifically describe the writing generator developed by ACTRES (Contrastive Analysis and Translation English-Spanish in its Spanish acronym), a research group devoted to the design of writing applications and to the construction of Controlled Natural Languages for the international promotion of products and services, namely in the food sector. This writing generator has been developed by IT engineers, expert linguists and translators following a corpus-based methodology, ensuring the use of authentic, accurate language. In the next sections we will describe how these researchers have set the foundations of the tool using Corpus Linguistics, to later explain and exemplify how the generator works, taking into account the way it can help translators.

4.2.1. Methodological framework of the generator

The ACTRES oenology generator uses comparable corpora, since they allow to describe the differences in the structure of the genres in the two languages under study and the results of that initial contrastive study are used to feed the writing generators in the following stage. These are *ad hoc* domain-specific corpora (Corpas Pastor & Seghiri 2009: 78) in English and Spanish compiled using pragmatic text selection criteria: representativeness, to ensure a representative sample of the language of expert members of the discourse community; and availability, taking into account the ease to obtain the texts constituting the corpus.

The corpus includes wine tasting notes from specific websites, such as those of the Appellations of Origin in Spain and North America, which give direct and restricted access to the information written by winery oenologists. Only the wine tasting notes included in wine tasting technical sheets released by wineries were used, so as to ensure that the writer was an expert and that the audience being addressed also consisted of experts. This methodological adjustment allowed for more parallelism between the English and Spanish texts, hence for more accurate interlingual comparison of the wine tasting notes.

The final corpus includes 750 wine tasting notes in Spanish and 716 wine tasting notes in English, which amount to 54,545 and 55,339 words respectively. The resulting corpus is not big in size, but it meets the appropriate criteria to fulfill the purpose of the study, taking into account that a smaller corpus may be called for when rhetorical tagging is used, which can only be done semi-automatically (Flowerdew 2005: 329). That is the case of this corpus. First, the files were tagged to provide pragmatic information for the texts, such as the winery or where the notes come from and the date of publication. Then, texts were labeled to identify the different rhetorical moves. The labeling process has to be necessarily manual, since it implies a process of constant decision-making. At the same time, the corpora samples are labeled and managed using certain software that has been integrated in an application suite that was especially designed for the creation of the writing generator, as the ACTRES Browser and Tagger, which will be described in the next section.

4.2.2. The generator development process: corpus analysis

The ACTRES generator specifically includes three elements of analysis that follow a top-down methodology.

The first element included in the generator is the prototypical rhetorical structure showing the moves and steps to be included when writing the genre in English. To obtain it, texts are manually labelled using rhetorical labels that help setting up the semantic units (*moves* and *steps*, according to Swales 1990, 2004) that serve as common ground to describe the prototypical semantic units in each language (Bondarko 1984), their cross-linguistic

juxtaposition and their contrast in order to obtain the prototypical structure for the genre under study for the target discourse community. This qualitative analysis is complemented by a quantitative one, following Suter's criteria (1993), to distinguish the most recurrent moves from the secondary ones by the frequency of occurrence of each rhetorical move. The most frequently recurrent moves, which range between a frequency of 40% and 100%, are considered conventional (Biber et al. 2007: 24) or compulsory (Suter 1993: 119). This category includes Suter's compulsory *high-priority* and *medium-priority* moves and steps. The moves occurring the least frequently (<40%) are deemed to be of low priority and occasional and are called *optional*. In this case, only compulsory high-priority and medium-priority moves are incorporated; that is to say, only moves with more than 40% frequency are included. This part of the process is carried out with help of three of the components of the ACTRES application suite: the Filemanager, an online tool that allows researchers to manage, store and use their corpus; the Tagger builder, an online tool that allows researchers to develop a semantic tree with all the moves and steps identified in the genre under study; and the Tagger, an on-line tool designed to manually label the texts (i.e. files) once a move and/or step has been identified, as well as to manage and store the files.

Secondly, the generator includes the lexico-grammatical patterns most frequently used in each move and step so as to solve problems on how to string words together, not only correctly and acceptably, but also idiomatically. Once the moves and steps are identified, the top-down methodology identifies those lexico-grammatical resources typical in the genre under study. More specifically, it focuses on specialized phraseology or terminological word combinations, also called *phraseological units*, which occur frequently in the technical language of wine. These include collocations, irreversible binomials, idioms, routine formulae and combinatorial patterns (Roberts 1998; Andrades 2014). The underlying assumption is that, by identifying key technical and subtechnical terms in each move and step in both language corpora, we are able to detect some relevant phraseological units that structure the information and are valuable for describing patterns of behaviour in the grammar used by the target discourse community when writing this specific specialized genre. The description and analysis of these units lead to the description of recursive lexico-grammatical patterns in

each move and step. This part of the process is carried out with help of another component of the ACTRES application suite: the Browser, where researchers can analyse and contrast rhetorical structures and obtain the frequency of occurrence of moves and steps, thus making it easy to identify the prototypical structures.

The third element included in the generator is the terminological and phraseological glossary (with examples extracted from the corpus) that will provide not only terms, but also their most common collocations. For this step, researchers also resort to the Browser, since it displays a word list and functions as a monolingual and bilingual concordancer, so the researcher obtains phraseological and terminological information that can be described and compared to identify lexico-grammatical patterns of behaviour.

And how does the information extracted from these elements help build the generator? As an example, a qualitative and quantitative analysis carried out with the first element showed, among other findings, that the central moves (those occurring between 70% and 100% of the cases) in the two languages correspond to the different tasting phases; in other words, colour, nose and palate are the compulsory moves for the construction of a wine tasting note in English and in Spanish. The results of the lexico-grammatical analysis showed different recurrent structures with different levels of complexity that will be reflected in the generator. Finally, the terminological and phraseological searches allowed describing and defining subtechnical terms and collocations in the two languages, which are stored in a dictionary connected to the generator. These and all the relevant data obtained in the analysis phase are extracted, and the resulting structure, patterns and terminology (dictionary) are linked to the generator software following the same top-down structure.

4.2.3. Operation of the generator

Technically speaking, generators are computer-friendly applications that guide the user through the writing process. The wine writing generator of ACTRES is a web-based tool that leads the user through the common rhetorical structure of the genre of wine tasting notes:

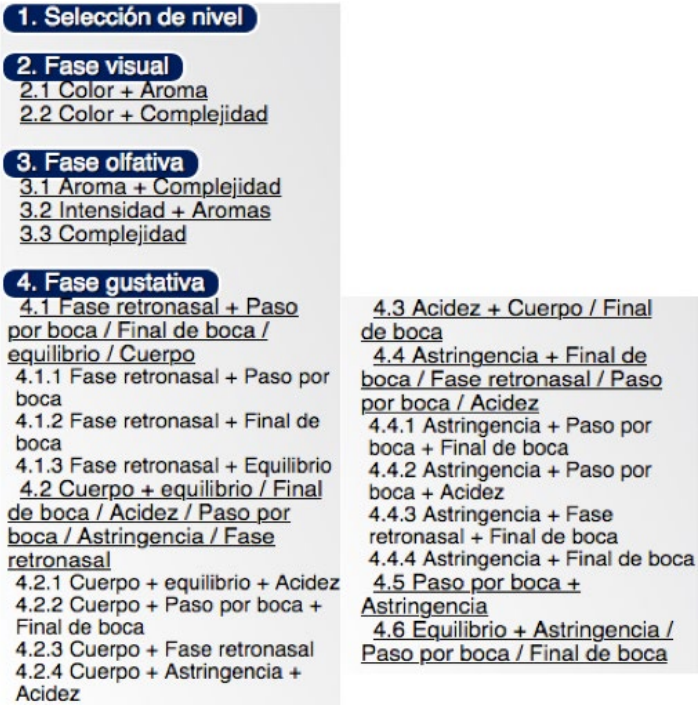

Figure 1. Generator interface: presentation of rhetorical structure.

First, the user has to choose between three levels of complexity, option 1 being the simplest, and option 3 being the most complex. Assuming that the potential users of the generator will have different levels of knowledge of the specialized language of wine tasting in English and assuming as well that, linguistically and pragmatically speaking, there are different options to express the same meaning in a given language, the generator will provide different structures for each level. Once chosen, the options offered to the user will correspond to that level of complexity throughout the process.

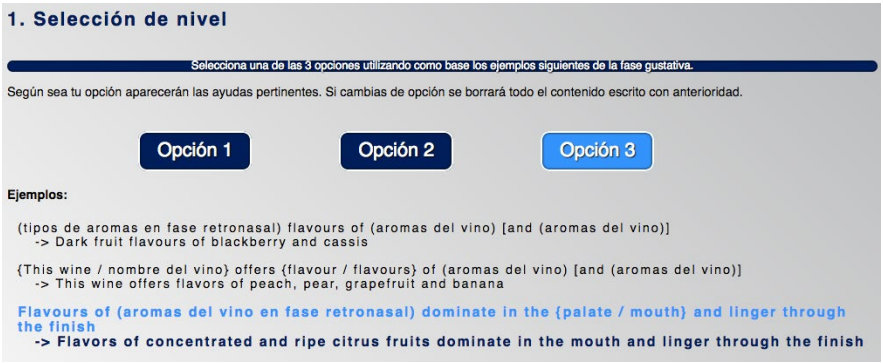

Figure 2. Generator interface: levels of difficulty.

Then, the generator presents moves and steps in Spanish to the user, who can then begin the writing phase. The user is offered different common structures in English for each step, so the translator can choose the steps and structures that appear in the source text. Since the structures in the generator are the prototypical ones according to the corpus, it is highly probable that the structure of the source text is among the ones suggested by the tool. Otherwise, the proposed structures are constituted by different parts that can be edited during the process.

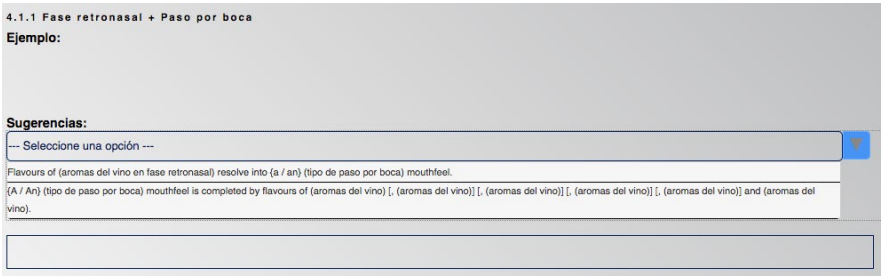

Figure 3. Generator interface: suggested structures within a step.

Each structure contains three different elements marked with different orthographic symbols:

{The wine / (nombre del vino)} is (nombre de color) in colour, {displaying / showing} aromas of (aromas del vino) [, (aromas del vino)] [, (aromas del vino)] [, (aromas del vino)] [, (aromas del vino)] and (aromas del vino).

First, there are fixed elements that form the basic skeleton of the structure and offer no changing options, although they can be altered if deemed necessary to comply with the customer's demands. Second, there are dependent elements that are required for reasons of grammar or use; they appear between keys, and the user is offered the most common options in the target language (English in this case) to choose the one that will fill that given gap by clicking on it. Finally, there are changing elements, parts in parentheses that suggest the kind of lexical information they could be filled with. When a list of elements of the same kind is usual, these changing elements are replicated one after the other and placed between square brackets, to highlight that they are optional. The suggestions of these changing elements are in the source language (Spanish) so the writer can insert the term(s) of the source text here, and the generator's dictionary will display the entries that include that term along with the English equivalents. The user can then choose the most adequate depending on the information in the source text.

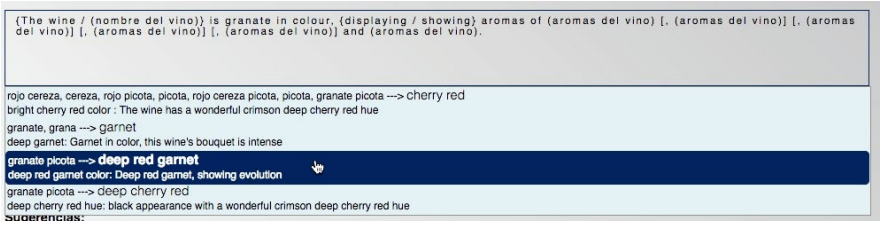

Figure 4. Generator interface: dictionary entries containing the term *granate*.

This process is repeated to fill all the moves and steps of the wine tasting note that are necessary with the lexical information desired. As a result of the thorough analysis carried out by expert linguists on the genre, the resulting text will contain the accepted structure, lexico-grammatical patterns and specific terminology of wine tasting notes in the target language and community, requiring little or no documentation/drafting on the translator's side. Furthermore, the user will only have to carry out a minimal proofreading

phase (e.g., in case there is something very specific that a given client wants to include outside the norm).

Additionally, the user can upload their own pictures (as the logo or the bottle), and other common details (website, email, address of the winery, etc.) to the generator, and download the final document in different formats (PDF, MS Word and xml). These final steps could not only save time for translators, but would also give them the chance to offer the client the full finished product, and not just the text; this would definitely add value to the translator's work in this highly competitive market.

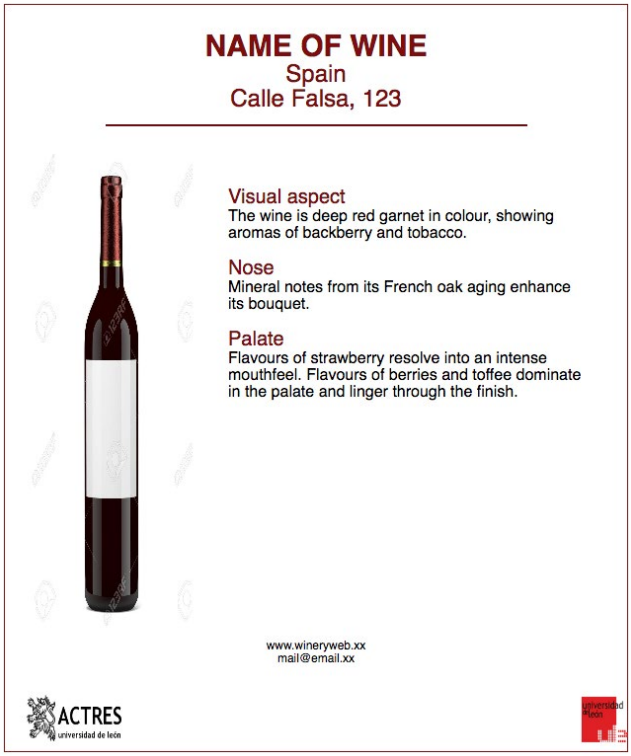

Figure 5. Preview of a finished wine tasting note.

5. Conclusions

The translation market has become a minefield for translators: the high competition makes it complicated for professionals to specialize in one field of expertise or market, so translators have to accept high workloads at low prices and despite the tight deadlines. All that competition is increased by the fact that clients, especially businesses, understand translation as a commodity that can be offered by any supplier that meets their availability, productivity, cost and perceived translation quality requirements (Moorkens 2017: 465).

In this context, Corpus Translation Studies can still help translators through the creation of tools that are adapted to the market, implying a true knowledge transfer between the profession and the academia, a process that is many times neglected in this and other areas of research. More specifically, writing aids (namely generators), are tools that can help increasing productivity, as they shorten the translation process and allow translators to handle a variety of genres from different fields of specialization; they improve costs, as specialized genres are usually better paid and using this tool translators can start producing acceptable texts from the very beginning; and they ensure quality, as they are tools built by expert linguists in the field with the information extracted from the analysis of a compilation of real texts. All this translates into high efficiency.

The generator described here is only an example, but there are more writing aids built by linguists and developed in the academia available, as the ones for the fields of tourism, medicine or public administration by the arText Project (cf. http://sistema-artext.com), or the other writing generators by the ACTRES group, some also related to the food industry (cheese, biscuits, herbal teas, and dried meats), some focused on other fields, such as tourist promotion, electronic products, company documentation, etc. (cf. https://actres.unileon.es/wordpress/?lang=en).

Although our study proposes that generators can be positive for translators to overcome some of the key problems in the translation industry, they also have a few possible drawbacks. Currently, the writing aids available are limited to certain genres and language pairs. However, if the use of generators became a trend among translators or companies, demand would

foster research in this area, so more genres and language pairs could be added to the existing sources. A second disadvantage is that users who want to use certain writing aid tools have to pay a license, as is the case of the tool described here. Although an initial investment might be seen as a burden, the aforementioned efficiency improvement in terms of time, costs and quality would help recover that investment in a short period of time. Finally, the fact that writing aids are based on corpora could help solving a common problem of specialized sources: updating "newly coined specialized units" (Durán-Muñoz & Corpas Pastor 2020: 163). If the demand existed, developers could regularly feed the generators' dictionaries by including new texts in the corpora they are based on.

We hope to have opened a worthy path for translators, whose job is increasingly necessary and demanding, but decreasingly cost-effective. Even a slight improvement in efficiency can make the difference, since clients might stop seeing translation as a burden and start seeing it as value for money. Pedagogy about the importance of translation is still useful and necessary, but the reality calls for different approaches to make businesses and translators' points of view meet.

References

AGUAYO, Natividad. (2013) "El traductor-intérprete en el comercio exterior: ¿Realidad o necesidad?" *Entreculturas* 5, pp. 57-74.

ALBUQUERQUE, Alexandra & Rute Costa. (2018) "The Satisfactory Cycle of Terminology Management in Translation-Mediated Business Communication: Problems and Opportunities." In Gallego Hernández, Daniel & Éric Poirier (eds.) 2018. *Business and Institutional Translation: New Insights and Reflections*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 149-164.

ANDRADES, Arsenio. (2014) *Estudio contrastivo de unidades fraseológicas especializadas (UFE) en un corpus comparable bilingüe de contratos de derecho civil en lengua inglesa y española*. Madrid: UCM. Unpublished PhD.

Aula.Int. (2005) "Translator Training and Modern Market Demands." *Perspectives* 13:2, pp. 132-142.

BAKER, Paul (2018) "Reflecting on Reflecting Research." In Taylor, Charlotte & Anna Marchi (eds) 2018. *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, pp. 281-292.

Bautista Zambrana, Mª Rosario. (2019) *Terminología y ontologías: Un estudio alemán-inglés-español basado en corpus*. Granada: Comares.

Biber, Douglas; Ulla Connor & Thomas A. Upton (eds.). (2007) *Discourse on the Move. Using Corpus Analysis to Describe Discourse Structure*. Antwerp: John Benjamins.

Bondarko, Alexander. V. (1984) *Functional Grammar. A Field Approach*. Amsterdam: John Benjamins.

Bowker, Lynne. (2005) "Productivity vs Quality? A Pilot Study on the Impact of Translation Memory Systems." *Localisation Focus* 4 (1), pp. 13-20.

Carrero, Francisco; José Carlos Cortizo & José María Gómez. (2008) "Building a Spanish MMTx by Using Automatic Translation and Biomedical Ontologies." In: Fyfe, Colin; Dongsup Kim; Soo-Young Lee & Hujun Yin (eds.) 2008. *Intelligent Data Engineering and Automated Learning – IDEAL 2008*. Berlin/ Heidelberg: Springer, pp. 346-353.

Corpas Pastor, Gloria & Miriam Seghiri. (2009) "Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents". In Beeby, Allison; Patricia Rodríguez & Pilar Sánchez-Gijón (eds.) 2008. *Corpus Use and Translating*. Antwerp: John Benjamins, pp. 75-107.

Doherty, Stephen. (2016) "The Impact of Translation Technologies on the Process and Product of Translation." *International Journal of Communication* 10, pp. 947-969.

Dunne, Keiran J. (2012) "The Industrialization of Translation: Causes, Consequences and Challenges." *Translation Spaces* 1, pp. 143-168.

Durán-Muñoz, Isabel & Gloria Corpas Pastor. (2020) "Corpus-Based Multilingual Lexicographic Resources for Translators: an Overview." In: Domínguez Vázquez, María José; Mónica Mirazo Balsa & Carlos Válcarcel Riveiro (eds.) 2020. *Studies on Multilingual Lexicography, (Lexicographica, Series Maior)*. Berlin: De Gruyter, pp. 159-178.

Durán-Muñoz, Isabel. (2012) *La ontoterminografía aplicada a la traducción: Propuesta metodológica para la elaboración de recursos terminológicos dirigidos a traductores*. (Studien Sprachwissenschaft Und Interkulturel Kommunikation 80). Frankfurt am Main/New York: Peter Lang.

Flowerdew, Lynne. (2005) "An Integrated Approach of Corpus-Based and Genre-Based Approaches to Text Analysis in EAP/ESP: Countering Criticism." *English for Specific Purposes* 24, pp. 321-332.

FUERTES-OLIVERA, Pedro A. (2012) "Lexicography and the Internet as a (Re-) source." *Lexicographica* 28, pp. 49-70.

HENNECKE, Angelika. (2017) "El entorno actual del mercado y la necesidad de traducción especializada en Alemania." *Cuadernos de Lingüística Hispánica* 30, pp. 19-41.

IBÁÑEZ RODRÍGUEZ, Miguel; Jesús Bachiller Martínez & María Teresa Sánchez Nieto. (2010) "Comercio exterior y mediación lingüística en el sector vitivinícola de Castilla y León." *Hermeneus* 12, pp. 161-182.

KATAN, David. (2016) "Translation at the Cross-Roads: Time for the Transcreational Turn?" *Perspectives. Studies in Translatology* 24(3), pp. 365-381.

KING, Kevin B. (2007) *The Writing Template Book*. University of Michigan: USA.

KREITZBERG, Charles B. & Ambrose Little. (2009) "Usability in Practice: Useful, Usable and Desirable. Usability as a Core Development Competence." *MSDN Magazine*. Available at: <http://msdn.microsoft.com/en-us/magazine/dd727512.aspx>

LANDAUER, Thomas. K. (1995) *The Trouble with Computers: Usefulness, Usability, and Productivity*. Cambridge: The MIT Press.

LAVIOSA, Sara. (2002) *Corpus-Based Translation Studies: Theory, Findings, Applications*. Amsterdam: Rodopi.

LÓPEZ-ARROYO, Belén & Roda P. Roberts. (2015) "The Use of Comparable Corpora: How to Develop Writing Applications. In: Sánchez Nieto, Maria Teresa (ed.) 2015. *Corpus Based Translation and Interpreting Studies: From Description to Application*. Berlin: Frank & Timme, pp 147-156.

LÓPEZ-ARROYO, Belén & Martín Fernández Antolín. (2011). "Estudios basados en corpus y lexicografía bilingüe: aplicaciones en un diccionario de fichas de cata". In BAZZOCCHI, Gloria, Pilar Capanaga & Sara Piccioni. (eds.) 2011. *Turismo ed enogastronomiatra Italia e Spagna. Linguaggi e territorio da esplorare*. Milán: Franco Angeli, pp. 99-116.

MALAMATIDOU, Sofia. (2018) *Corpus Triangulation. Combining Data and Methods in Corpus Based Translation Studies*. London: Routledge

MARCHI, Anna & Charlotte Taylor. (2018) "Partiality and Reflexivity". In TAYLOR, Charlotte & Anna Marchi. (eds.) 2018. *Corpus Approaches to Discourse: A Critical Review*. London: Routledge, pp. 1-16.

MAYORAL, Roberto. (2006) "La traducción comercial." *Butlletí de la Associació de Traductors i Intèrprets Jurats* (diciembre), s.p. Available at: <https://www.ugr.es/~rasensio/docs/Traduccion_comercial.pdf>

MEDINA REGUERA, Ana & Carmen Álvarez García. (2014) "La relación empresa-traducción en el sector agroalimentario andaluz." *Skopos* 4, pp. 187-206.

MOORKENS, Joss. (2017) "Under Pressure: Translation in Times of Austerity." *Perspectives* 25(3), pp. 464-477.

MORÓN MARTÍN, Marian & Ana Medina Reguera. (2016) "La competencia del traductor que no 'traduce': el traductor en ámbitos de internacionalización empresarial." *MonTI* 8, pp. 225-255.

PYM, Anthony. (2011) "What Technology Does to Translating." *Translation & Interpreting* 3 (1), pp. 1-9.

PYM, Anthony. (2013) "Translation Skill-Sets in a Machine-Translation Age." *Meta* 3, pp. 487-503.

QUESENBERY, Whitney. (2001) "What Does Usability Mean: Looking beyond 'Ease of Use'. Available at: <http://www.digitalspaceart.com/projects/cogweb2002v2/papers/whitney/whitney1.html>

RABADÁN ÁLVAREZ, Rosa. (2008) "Refining the Idea of 'Applied Extensions'. "In Pym, Anthony; Miriam Schlesinger & Daniel Simeoni (eds.) 2008. *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury*. Antwerp: John Benjamins, pp. 103-118.

RABADÁN ÁLVAREZ, Rosa. (2005-2008) "Tools for English-Spanish Cross Linguistic Applied Research". *Journal of English Studies* 5-6: 309-324.

RICO PÉREZ, Celia & Álvaro García Aragón. (2016) *Análisis del sector de la traducción en España (2014-2015)*. Villaviciosa de Odón: Universidad Europea. Available at: <http://abacus.universidadeuropea.es/bitstream/handle/11268/5057/analisis%20sector%20traduccion%2014-15.pdf?sequence=2&isAllowed=y>

ROBERTS, Roda P. (1998) "Phraseology and Translation." In: Fernández Nistal, Purificación and José María Bravo Gonzalo (eds.) 1998. *La traducción: orientaciones lingüísticas y culturales*. Valladolid: Universidad de Valladolid, pp. 61-77.

ROBINSON, Douglas. (2020) *Becoming a Translator: An Introduction to the Theory and Practice of Translation*. London/New York: Routledge.

SALVADOR INSÚA, José Antonio. (2016) *Mercado internacional del vino: intentos de modelización y estrategias territoriales de comercialización en España*. Valladolid: Universidad de Valladolid. Unpublished PhD.

SANZ VALDIVIESO, Lucía & Belén López-Arroyo (2020). "On Describing Olive Oil Tasting Notes in English." *Fachsprache. Journal of Professional and Scientific Communication* 42.1-2, pp. 27-45.

Several Authors (International Organisation of Vine and Wine, OIV). (2019) *Statistical Report on World Vitiviniculture*. Available at: <http://www.oiv.int/public/medias/6782/oiv-2019-statistical-report-on-world-vitiviniculture.pdf>

SINCLAIR, John. (2004) *Trust the Text. Language, Corpus and Discourse*. London: Routledge.

SOSONI, Vilemini & Margaret Rogers. (2013) Translation in an Age of Austerity: from Riches to Pauper, or not? *mTm* 5, pp. 5-17.

SUPATRANONT, Pisamai. (2012) "Developing a Writing Template of Research Article Abstracts: A Corpus-Based Method." *Procedia Social and Behavioural Sciences* 66, pp. 144-156.

SUTER, Hans-Jürg. (1993) *The Wedding Report. A Prototypical Approach to the Study of Traditional Text Types*. Antwerp: John Benjamins.

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

SWALES, John. (2004). *Research Genres*. Cambridge: Cambridge University Press.

TARP, Sven & Pedro Fuertes-Olivera. (2016) "Advantages and Disadvantages in the Use of Internet as a Corpus: The Case of the Online Dictionaries of Spanish Valladolid-Uva." *Lexikos* 26, pp. 273-296.

VARONA, Lucema. (2002) "El traductor ante la micro y pequeña empresa pyme." In: Alcina Caudet, Amparo & Silvia Gamero Pérez (eds.) 2002. *La traducción científico-técnica y la terminología en la sociedad de la información*. Castelló de la Plana: Publicacions de la Universitat Jaume I, pp. 201-206.

ZANETTIN, Federico. (2013) "Corpus Methods for Descriptive Translation Studies." *Procedia - Social and Behavioral Sciences* 95, pp. 20-32.

BIONOTES / BIONOTAS

LETICIA MORENO-PÉREZ holds a PhD in Professional and Institutional Translation from Universidad de Valladolid (Spain), where she has lectured topics such as English for Specific Purposes or specialized translation since 2010. Her research focuses on specialized language and its translation from

a functional approach through corpus-based studies, specifically in the field of legal and business translation. She has published and presented at international conferences on this area, and she is a member of the international research group ACTRES (Contrastive Analysis and Translation English-Spanish in its Spanish acronym), also specialized in the field. She has combined her academic work with professional translation and language training in the business sector for 10 years.

Belén López-Arroyo is an Associate Professor in ESP at the University of Valladolid (Spain). She taught legal and business translation in the School of Translation and Interpreting from 1997 to June 2013 and she currently teaches legal and business translation and Corpus Linguistics in the English Studies Degree. Her research interests include Discourse Analysis, Genre Analysis, Lexicography and Terminology Contrastive analysis and Translation. She is author of several articles and books related to contrastive analysis of scientific and professional genres and its implication for translation. In the ACTRES team she is in charge of the Rhetoric and Phraseology of Expert-to-Expert Discourse (in different areas) and its applications for developing writing aids in English for Spaniards.

Leticia Moreno-Pérez es Doctora en Traducción Profesional e Institucional por la Universidad de Valladolid (España), donde ha impartido docencia en áreas como Inglés para Fines Específicos o traducción especializada desde 2010. Especializada en el campo jurídico y de los negocios, su investigación se centra en el lenguaje especializado y su traducción desde una perspectiva funcional a través de los estudios basados en corpus. Ha publicado y participado como ponente en congresos internacionales sobre este campo, y es miembro del grupo ACTRES (Análisis Contrastivo y Traducción Inglés-Español), especializado en este área. Además de su trabajo en el ámbito académico ha ejercido como traductora y formadora de lengua inglesa en el ámbito empresarial durante 10 años.

Belén López-Arroyo es Profesora Titular de Universidad en el área de Filología Inglesa en la Universidad de Valladolid (España). Impartió docencia en la Facultad de Traducción e Interpretación (Campus de Soria) desde 1997 hasta el 2013. Actualmente, imparte Traducción especializada I: derecho y economía y Lingüística del Corpus en el Grado en Estudios Ingleses

en la Facultad de Filosofía y Letras. Su investigación incluye análisis del discurso, análisis textual, terminología y fraseología contrastiva y traducción. Es autora de diversos artículos y libros en el campo del análisis contrastivo de géneros científicos y profesionales. Miembro del grupo ACTRES desde su creación, donde se encarga de la retórica y fraseología contrastiva en el discurso experto-experto en diferentes áreas y sus aplicaciones para el desarrollo de herramientas de ayuda a la escritura en inglés de géneros especializados para hispano parlantes.

Recibido / Received: 11/04/2020
Aceptado / Accepted: 23/09/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.09

Para citar este artículo / To cite this article:
Santamaría Urbieta, Alexandra & Elena Alcalde Peñalver. (2021) "Autocrítica de publicaciones previas basadas en corpus: Análisis DAFO." En: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection. MonTI* 13, pp. 280-300.

AUTOCRÍTICA DE PUBLICACIONES PREVIAS BASADAS EN CORPUS:ANÁLISIS DAFO

Alexandra Santamaría Urbieta
alexandra.santamaria@unir.net
Universidad Internacional de la Rioja

Elena Alcalde Peñalver
e.alcalde@uah.es
Universidad de Alcalá

Resumen

El objetivo de este artículo consiste en reflexionar y hacer autocrítica sobre cuatro publicaciones previas elaboradas conjuntamente por las autoras y en las que se utilizó una metodología de corpus. Para ello, se empleará la metodología DAFO (Debilidades, Amenazas, Fortalezas y Oportunidades), que nos permitirá analizar, por un lado, la dirección hacia la que se dirigen los estudios de traducción basados en corpus y, por otro, lo que sería necesario mantener, mejorar o modificar en las publicaciones de esta tipología. A través de un análisis primeramente teórico, en el que se realizará un recorrido por los estudios basados en corpus, pasaremos a presentar la metodología empleada para analizar, desde un punto de vista crítico, las cuatro publicaciones de las autoras. Consideramos que el presente artículo tendrá un doble efecto a nivel docente e investigador y que por tanto redundará en beneficio de los Estudios de Traducción.

Palabras clave: Autocrítica; Corpus; DAFO; Reflexión; Traducción.

Abstract

The objective of this article is to reflect and make self-criticism on four previous publications elaborated jointly by the authors and in which a corpus methodology

was used. For this purpose, the SWOT methodology (Strengths, Weaknesses, Opportunities and Threats) will be used, which will allow us to analyze, on the one hand, the direction towards which corpus-based translation studies are heading and, on the other, what would need to be maintained, improved or modified in publications of this type. Through a first theoretical analysis, in which a tour of corpus-based studies will be conducted, we will move on to present the methodology used to analyze, from a critical point of view, the four publications of the authors. We consider that this article will have a double effect at the teaching and research levels and that it will therefore benefit Translation Studies

Keywords: Self-criticism; Corpus; SWOT; Reflection; Translation.

1. Introducción

Podríamos decir que existen tres tipos de estudios de traducción que se llevan a cabo con el uso de corpus. Por un lado, nos encontramos con los estudios de corpus aplicados, por otro, los teóricos y, por último, los descriptivos. Los primeros se centran en el uso de este recurso en las actividades traslativas para el proceso de documentación, de toma de decisiones, de revisión o incluso de evaluación. Los segundos se centran en la "descripción, modelado y representación" (Corpas Pastor 2008: 12). Los últimos, en cambio, analizan y estudian el concepto de la equivalencia en el campo de la traducción. Corpas Pastor subrayó su importancia al apuntar que "la lingüística de corpus ofrece un marco idóneo para la búsqueda de la equivalencia en traducción" (2008: 83). Estos estudios descriptivos o Estudios de Traducción con Corpus (ETC), que tienen sus inicios de la mano de Baker (1993) en los años 90, inciden en la necesidad y utilidad de este tipo de corpus no solo para el estudio de la traducción desde el punto de vista meramente teórico, sino también desde su perspectiva más práctica, puesto que la compilación de un corpus, ya sea este paralelo o comparable, ofrece la posibilidad al traductor de disponer de los equivalentes lingüísticos más cercanos durante el proceso de traducción. Desde el punto de vista teórico, el corpus puede ser empleado en el aula de traducción con el objetivo de formar a traductores e intérpretes, así como para brindarles una herramienta de gran utilidad para su futuro como traductores. Es por ello que los Estudios de Traducción han sido sacudidos por la metodología de corpus (Corpas Pastor 2008) y por

sus múltiples usos tanto desde el punto de vista teórico como didáctico y profesional. Baker (1993), tomando como inspiración los estudios de Sinclair (1991), da algunas de las primeras pinceladas de las implicaciones teóricas de esta metodología, así como del concepto de equivalencia. Asimismo, la autora sugiere el uso profesional de los estudios de corpus y la utilidad de estos en un entorno informático y, más concretamente, en la tecnología de la traducción (Bowker y Corpas Pastor 2014), esto es, "un tipo de tecnología lingüística, monolingüe o multilingüe, diseñada para formar parte de los entornos de trabajo del traductor" (Corpas Pastor 2012: 2). Esta última ha experimentado grandes avances a consecuencia del uso del corpus. Ejemplos de ello son, por un lado, los sistemas de traducción asistida por ordenador (TAO), los cuales integran un gestor terminológico, y los sistemas automáticos de traducción basados en corpus (Statistical Based Machine Translation, SBMT). Estas herramientas forman parte del entorno profesional del traductor y todas ellas, en mayor o menor medida, integran el corpus en sus tareas, ya sea para compilar, como para gestionar o explotar el corpus, lo cual incide en la importancia de la metodología de corpus en los Estudios de Traducción. Como se observa, desde los primeros usos del corpus para la enseñanza de idiomas a finales de los años ochenta (Johns 1991) hasta la actualidad, el corpus se ha introducido en los Estudios de Traducción desde el punto de vista descriptivo, así como didáctico. Se trata, por tanto, de un recurso de análisis e investigación que ha ido evolucionando de la mano de la lingüística, la traducción, la formación y la tecnología. Los enfoques con corpus se han constituido como pilares fundamentales, más aún en la actualidad, al aliarse estos con las técnicas del Procesamiento de Lenguaje Natural.

El objetivo de este artículo consiste en analizar cuatro publicaciones previas, en las que se empleó la metodología de corpus, y que han sido elaboradas de manera conjunta por las dos autoras. Para ello, tras esta introducción, continuaremos con un breve recorrido de la trayectoria que han tenido los estudios basados en corpus en el ámbito de la Traducción desde su comienzo hasta la actualidad. Seguidamente, explicaremos la metodología DAFO (Debilidades, Amenazas, Fortalezas y Oportunidades) que aplicaremos para analizar las publicaciones objeto de nuestro estudio.

2. La evolución de los estudios basados en corpus en Traducción

El uso de corpus ha demostrado, a lo largo de los años, sus beneficios e implicaciones pedagógicas (Sánchez Ramos 2017). Partiendo de esta base, en primer lugar, debemos establecer lo que entendemos por corpus, que, en palabras del grupo EAGLES (1996) y considerando esta definición una de las más estandarizadas, está formado por "*a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language*". Resulta complicado quedarse con una única definición, puesto que son varios los ámbitos que analizan su utilidad: la lingüística de corpus, la lexicografía de corpus, así como la lingüística computacional. Lo que está claro y parte como sustento de cualquier corpus es que este debe tomar como base una colección de textos que surgen como resultado de una situación real y su compilación debe estar guiada "por una serie de criterios lingüísticos explícitos para asegurar que pueda usarse como muestra representativa de una lengua" (Pérez Hernández 2002).

Antes de ahondar en la evolución de los estudios de traducción basados en corpus, desde un punto de vista más general, es necesario remontar los estudios de corpus a los años 50. Durante esta década se experimentó un interés creciente por los estudios lingüísticos desde una perspectiva empírica (Firth 1957) hasta que Chomsky (1957) diera paso a otro tipo de estudios, estos basados en una perspectiva más mentalista. Como consecuencia de la gran cantidad de información a la que se tiene acceso desde ya hace unas décadas a través de medios informáticos, los estudios de corpus de corte empírico despertaron de su letargo y se pusieron a disposición de investigadores y lingüistas, que desarrollaron nuevos métodos de estudio en una amplia variedad de campos entre los que se encuentra el de la traducción e interpretación.

Como ya se ha apuntado con anterioridad, si echamos la vista atrás al pasado de los estudios basados en corpus en el ámbito de la traducción, nos encontraríamos con Baker, quien es considerada "la precursora de proponer y adaptar este enfoque en corpus a los propósitos de los estudios empíricos y descriptivos de la investigación en traducción" (De Felipe Boto 2007: 261). Esta metodología surge de la mezcla de la Lingüística de Corpus (LC) y los Estudios Descriptivos de la Traducción (EDT) y ha experimentado

en las últimas décadas un amplio crecimiento (Laviosa 2002), puesto que "los estudios basados en corpus bien diseñados y organizados conducen a un desarrollo cualitativo y cuantitativo de la disciplina" (Sanz, Zubillaga y Uribarri 2015: 211). La clave de cualquier tipo de estudio basado en corpus, como apuntan estos autores, está en la calidad de su compilación, así como en su organización, que determinan la utilidad de los resultados tras el análisis. Dependiendo, por tanto, de la organización que apliquemos, estaremos ante tipos de corpus diferentes: corpus monolingües, corpus plurilingües, corpus comparables o corpus paralelos. Podríamos aunar todos estos corpus en tres grandes grupos: (1) corpus que se centran en el producto, (2) corpus que se centran en el proceso de traducción y (3) corpus orientados a la función de las traducciones.

Al igual que a lo largo de los años las investigaciones en traducción basadas en estudios de corpus han ido evolucionando, las herramientas empleadas para este fin también se han visto ampliadas. Sin ir más lejos, la misma definición del término "corpus" se ha visto ligeramente alterada como indica García Ferrer (2013: 94), puesto que por "corpus" se entiende "un conjunto de textos recogidos según unos criterios determinados para ser utilizados con unos propósitos específicos y en un formato legible por el ordenador". Se podría decir, por tanto, que este tipo de estudios se han beneficiado de la evolución de las nuevas tecnologías y que han aumentado aquellas investigaciones que sacan provecho de ellas para hacer búsquedas más exhaustivas, precisas y con un corpus de mayor tamaño y complejidad. Ya en el 2002 Bowker y Pearson describieron los cuatro criterios fundamentales que los corpus debían reunir: (1) que sean auténticos, (2) que estén recogidos en formato electrónico, (3) que sean lo suficientemente cuantiosos y (4) que los criterios de selección sean rigurosos. El tercer criterio establecido por estos dos autores incide en la relación, ya casi incuestionable, que existe entre el corpus y su almacenamiento en formato digital. Un corpus, como afirma Villayandre Llamazares (2008: 340) "para ser una herramienta útil al lingüista, debe estar informatizado, es decir, los textos de que consta tienen que estar en formato electrónico (corpus informatizado o automatizado)". Son cuatro las razones detrás de esta afirmación. Según la autora, el ordenador permite: (1) buscar información de una manera rápida, (2) recuperarla, (3)

calcular la frecuencia de aparición de una o varias palabras y (4) clasificar los datos obtenidos de acuerdo con diferentes criterios.

El gran tamaño de los corpus como, por ejemplo, el Corpus de Referencia del Español Actual, elaborado por la RAE, y el British National Corpus en inglés, que alcanzan los 100 millones de palabras, subrayan el impacto del soporte informático en los estudios de corpus en general y en los estudios de traducción en particular. Asimismo, la lingüística computacional también ha favorecido el renacer de la lingüística de corpus, que se aleja de una perspectiva meramente teórica para acercarse a aquella que se centra en el estudio y análisis de textos reales producidos por hablantes, esto es, surge la necesidad de desarrollar sistemas prácticos con el objetivo de crear gramáticas y léxicos computacionales (Villayandre Llamazares 2008). Del mismo modo, la traducción ha experimentado en los últimos años un avance hacia el desarrollo de la traducción automática (TA), y algunos de estos sistemas se basan en corpus lingüísticos y toman como base el análisis y comparación de textos bilingües y multilingües (TA basada en estadística), así como otros corpus trabajan con ejemplos (TA basada en ejemplos). Esta situación da muestra de que los estudios de corpus basados en traducción evolucionan con las nuevas herramientas tecnológicas para permitir estudios más reales desde el punto tanto teórico como práctico.

Con el fin de observar la evolución de los estudios de corpus basados en la rama de la traducción, tanto en inglés como en español, se ha hecho una búsqueda en Google Académico de las investigaciones publicadas en el área. Aunque somos conscientes de las limitaciones de esta herramienta, conviene destacar que no existe un buscador que ofrezca "una cobertura completa de las citas que se emiten" (Torres-Salinas, Ruiz-Pérez y Delgado-López-Cázar 2009: 508) y que en este artículo se presentan estos datos desde el punto de vista informativo y a nivel micro, esto es, "como ayuda a los autores e investigadores concretos en la búsqueda rápida, fácil y directa de documentos a texto completo" (2009: 510).

En primer lugar, hemos querido observar los resultados que lanza este buscador sobre los estudios de corpus y traducción desde 1990, primer año al que remonta la búsqueda el buscador, hasta la actualidad. Para ello, hemos optado por buscar los términos "corpus estudios traducción" y "corpus translation studies" porque la ausencia de nexos aporta mejores y más precisos

resultados (Villegas 2003). El número de estudios asciende a 65.600 resultados en español y 206.300 en inglés. Para precisar un poco más la búsqueda, hemos acotado los intervalos de búsqueda en periodos de 5 y 10 años, opción que permite Google Académico. A continuación, se muestran ambas gráficas:

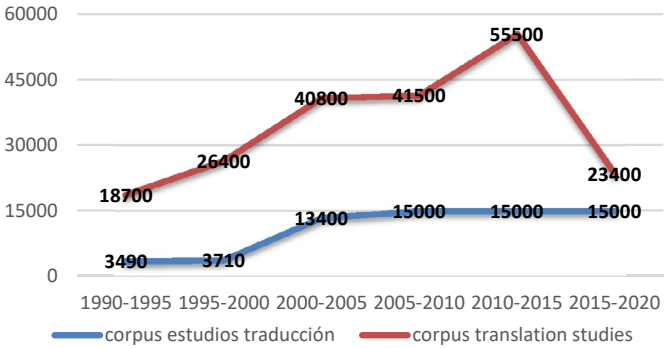

Ilustración 1: Intervalo de búsqueda en periodos de 5 años de los términos "corpus estudios traducción" y "corpus translation studies".

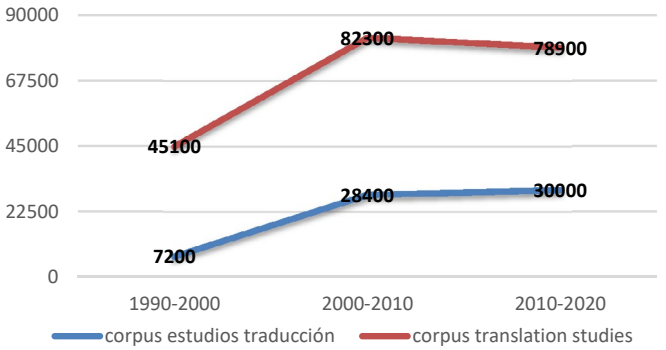

Ilustración 2: Intervalo de búsqueda en periodos de 10 años de los términos "corpus estudios traducción" y "corpus translation studies".

En la primera gráfica (Ilustración 1) en la que la búsqueda se ha hecho en intervalos de cinco años, se observa que en los dos primeros periodos (1990-1995 y 1995-2000) los resultados, aunque abundantes, son el principio de

una tendencia al alza que aumenta exponencialmente en el tercer periodo (2000-2005) y que continúa así en los siguientes años. En lo que respecta a la búsqueda en inglés, esta aumenta cada periodo de cinco años hasta el último periodo, en el que la presencia de este tipo de estudios disminuye. En cuanto a la búsqueda en español, y salvando las diferencias en cuanto a la mayor presencia del inglés en este ámbito de investigación, la tendencia es al alza y a mantener un ritmo constante de publicación de este tipo de estudios en los últimos años.

En la segunda gráfica (Ilustración 2), que muestra los resultados obtenidos en búsquedas de periodos de diez años, se observa una tendencia muy parecida en ambos idiomas, puesto que la gráfica es muy similar. Se podría decir que el periodo que comprende los años 2000 y 2010 ha sido el más fructífero en cuanto a la producción de estudios relacionados con el corpus y los estudios de traducción.

Los estudios basados en corpus en traducción van de la mano de las nuevas investigaciones que surgen en el campo y, por lo tanto, irán evolucionando con los avances del mundo de la traducción. Desde los primeros estudios que inspiraron la lingüística de corpus, allá por los años 50 hasta la actualidad, pasando por los primeros estudios propiamente de corpus expuestos por Baker en el año 1993, estos han tomado en el sector de la traducción diferentes caminos para ofrecer resultados válidos a la comunidad científica, esto es, al traductor profesional, al investigador, al docente en el aula de traducción y al alumno.

3. Metodología

Un análisis DAFO es una herramienta que permite detallar los aspectos positivos y negativos de un plan de empresa. Se trata de una forma estratégica de hacer un diagnóstico de la situación actual de una organización para tomar así las decisiones oportunas que permitan mejorar en el futuro (Espinosa 2020). De esta forma, este análisis nos permite examinar el contexto competitivo de un plan de empresa desde dos vertientes (externa e interna) y siempre sobre la base de hechos objetivos (Sisamón Gil 2012). Tal y como indica esta autora, la vertiente externa permite analizar las amenazas y oportunidades del sector o industria donde se sitúa la empresa para así anticiparse a ellas y

poder superarlas o aprovecharlas según las circunstancias que se desarrollen. Es importante señalar que en esta vertiente se deben definir las fronteras y los competidores a los que se enfrenta la empresa. Por otro lado, la vertiente interna analiza las fortalezas y debilidades de la empresa en sí misma, es decir, en función de la competencia, pero solo teniendo en cuenta factores individuales internos.

En el ámbito de la traducción, existen artículos previos que han utilizado esta herramienta para el análisis de contenidos, pero con otros objetivos investigadores, como en el caso de Morón Martín (2009) o Plaza Lara (2019). En la primera publicación, la autora realiza un análisis DAFO de la formación universitaria en Traducción e Interpretación en España en función de la perspectiva profesional y de la cobertura de necesidades de los egresados. Por su parte, Plaza Lara analiza en su estudio los másteres que forman parte de la red de Másteres Europeos en Traducción (EMT) con el objetivo de determinar si incluyen contenidos de posedición y traducción automática en sus planes de estudio y el grado de especialización en función de las asignaturas y módulos que se ofertan. Mediante el análisis DAFO, la autora identifica los factores internos y externos que pueden influir en la enseñanza de estos contenidos en los programas de enseñanza identificados.

De forma general, y para empezar nuestro estudio, estas serán las dos preguntas que nos plantearemos para el análisis de las publicaciones seleccionadas en este artículo:

- ¿Cuáles son los puntos negativos de la publicación? Esto nos permitirá detectar las amenazas y debilidades de cada publicación.
- ¿Cuáles son los puntos positivos de la publicación? Esto nos permitirá detectar las oportunidades y fortalezas de cada publicación.

De estas dos primeras preguntas obtendremos el punto de partida para realizar el análisis pormenorizado de cada publicación en función de la metodología detallada que aplicaremos y que recogemos en la Tabla 1, adaptada de Casero Gimón (2019). Además, es necesario señalar que tal y como lo hicieran Aliaga et al. (2018), realizaremos un análisis DAFO narrativo de cada publicación, es decir, aportaremos la información necesaria para responder a cada pregunta con el objetivo de "razonar más detalladamente y presentar esas reflexiones para el debate" (565).

Debilidades. Análisis interno	Fortalezas. Análisis interno
¿Qué factores pueden ser percibidos como una debilidad/carencia por parte de los lectores de la publicación? ¿Qué factores podrían reducir el éxito de la publicación?	¿Qué ventajas competitivas tiene esta publicación? ¿Qué resultados obtenemos que no obtienen otras publicaciones? ¿Qué recursos se han analizado y que no lo han hecho otras publicaciones? ¿Qué factores pueden ser percibidos como un beneficio por parte de los lectores de la publicación?
Amenazas. Análisis externo	**Oportunidades. Análisis externo**
¿A qué obstáculos se enfrenta la publicación? ¿Qué están haciendo sobre el tema otros académicos? ¿Pueden surgir nuevas publicaciones sobre el mismo tema? ¿Qué datos sobre el tema existen actualmente? ¿Puede alguna de las debilidades afectar la transferencia o replicabilidad de la publicación?	¿Cuáles son las tendencias de la investigación sobre este tema? ¿Es un tema sobre el que es necesario seguir investigando? ¿Qué cambios se están presentando que puedan propiciar que la publicación tenga una mayor transferencia o replicabilidad?

Tabla 1. Metodología DAFO para el análisis (Adaptada de Casero Gimón, 2019: 40).

4. Aplicación del análisis DAFO a publicaciones previas

En este apartado aplicaremos la metodología de análisis DAFO a cuatro publicaciones de 2019 realizadas de forma conjunta por las dos autoras. Hemos limitado la muestra a cuatro para poder hacer un análisis pormenorizado. Su selección se ha hecho sobre la base de la reciente fecha de publicación, lo que nos permite reflexionar acerca del trabajo elaborado, pero con la mirada fija en la evolución futura de la disciplina. Aplicaremos la metodología que explicamos en el apartado anterior en un formato de tabla en el que iremos dando respuesta a cada una de las preguntas mencionadas.

4.1. Publicación 1. Translation and Communication in the Promotion of Business Tourism: Emerging Research and Opportunities

Esta aportación es un libro publicado por la editorial IGI Global en el año 2019 en el que las autoras se adentran en el mundo de la traducción turística de negocios y en el tipo de lenguaje que se emplea para su promoción. A lo largo de los ocho capítulos que conforman el libro se ofrece una perspectiva teórica de lo que es el turismo de negocios, el discurso turístico a partir del análisis y descripción de un corpus en inglés y en español, y cómo se desarrolla la traducción de las guías promocionales de inglés a español, a partir de un estudio de corpus paralelo. Por otro lado, y desde una perspectiva práctica, el libro ofrece una batería de actividades y una clasificación de estrategias de transferencia para llevar al aula de traducción.

Debilidades. Análisis interno
Uno de los factores que pueden percibirse como una debilidad en esta publicación es el corpus analizado, puesto que este podría haberse ampliado. Por un lado, el corpus comparable estaba compuesto por 12.321 palabras y el paralelo por 186.407, 96.559 palabras en español y 89.848 en inglés. Probablemente un corpus más amplio podría haber aportado datos más exhaustivos sobre este tipo de discurso, principalmente en lo que respecta al corpus comparable.
Amenazas. Análisis externo
La amenaza que debemos reseñar se centra en la rapidez con la que avanzan ambos campos de investigación, el de las finanzas y el turismo, y en el nacimiento de nuevos discursos que podrían hacer que la publicación se convirtiera en obsoleta rápidamente. Sin embargo, como veremos más adelante, esta amenaza, bien analizada y delimitada, puede ser también una oportunidad para continuar en este campo de investigación.
Fortalezas. Análisis interno
Una de las grandes fortalezas de esta publicación está en el último apartado del libro, puesto que no son frecuentes los estudios que, además del análisis teórico de la cuestión, ofrecen a sus lectores un apartado práctico para que, en este caso, el docente de traducción lleve al aula actividades que le pueden servir al alumno para adentrarse en el mundo de la traducción turística de negocios. Asimismo, aunque las estrategias de transferencia se han venido estudiando a lo largo de los años, esta publicación presenta una clasificación práctica y útil desde el punto de vista científico y académico, que podría aplicarse al análisis de otro tipo de discursos. Por último, debemos destacar que, a pesar de que la traducción turística es un tema de investigación habitual, no es frecuente encontrarlo junto al mundo de los negocios. El número de publicaciones sobre este campo es escaso (Napu 2016; Sulaiman & Wilson 2018), por lo que esta publicación es de gran interés, así como novedosa.

Oportunidades. Análisis externo
El tema objeto de estudio en este libro ofrece amplias oportunidades en el ámbito de la investigación científica, puesto que el turismo avanza con la sociedad, surgen nuevas formas de viajar y diferentes maneras de hacerlo. Esto implica que el discurso turístico también vaya evolucionando y sea objeto de estudio tanto en la actualidad como en el futuro. Asimismo, el mundo de los negocios no se estanca y está inmerso en un proceso constante de cambio que implica que nazcan nuevos géneros y discursos. Estos dos campos, al juntarse, crean oportunidades de análisis lingüístico que investigadores del área aprovecharán y que, por tanto, este primer volumen podría ampliarse en un segundo número e incluir otros géneros turísticos o avances que hayan surgido en el sector. Como podemos observar, lo que hemos destacado en el apartado de amenazas sobre la rapidez de evolución y desarrollo de la temática de la publicación constituye también la principal oportunidad.

Tabla 2. Análisis de la publicación 1. Fuente: Elaboración propia.

4.2. Publicación 2. Estudio contrastivo de dos guías promocionales sobre turismo de negocios (inglés-español)

Este artículo fue publicado en el número 18 de la revista *Hikma* y en él se presenta un estudio donde se identifican y describen las estrategias empleadas en la traducción de los elementos léxicos de dos guías promocionales del turismo de negocios. El corpus objeto de estudio está conformado por cuatro subcorpus equilibrados: dos subcorpus en español y sus correspondientes traducciones al inglés. El primer corpus paralelo describe los beneficios de llevar a cabo reuniones y congresos en México y el segundo corpus, por otro lado, hace lo correspondiente con la provincia española de Asturias. En ambos casos se trata de guías promocionales en soporte electrónico. El subcorpus en español tiene un total de 71.651 palabras y el subcorpus en inglés está formado por 64.863 palabras. Asimismo, uno de los objetivos de esta investigación era reflexionar sobre la docencia en el ámbito de la traducción turística de negocios a través de la explotación de corpus en el aula.

Debilidades. Análisis interno
Uno de los factores que podrían reducir el éxito de la publicación es que no se incluyen datos cuantitativos para cada una de las estrategias de traducción analizadas. Incluir estos datos habría aportado al artículo mayor rigor científico. Se tiende a la generalidad y a dar datos demasiado vagos para tratarse de una investigación con un corpus tan extenso. Habría sido necesario aportar porcentajes y cifras para que el lector se hiciera una idea más real sobre la presencia de elementos léxicos en el corpus objeto de estudio. Otra de las debilidades del estudio, aunque por razones de espacio, es que este se centra en un análisis del léxico y no en otros aspectos que también podrían ser de interés. Además, el título podría llevar al lector a pensar que se va a encontrar con un análisis completo de las guías promocionales cuando, en realidad, las autoras solo se centran en un grupo de elementos.
Amenazas. Análisis externo
La principal amenaza a la que se enfrenta la publicación está estrechamente relacionada con la debilidad anteriormente descrita, ya que no haber incluido porcentajes para cada una de las estrategias de transferencia estudiadas puede afectar a la replicabilidad del estudio, así como a su uso como base para futuras investigaciones sobre el tema.
Fortalezas. Análisis interno
Una de las fortalezas principales de este estudio es el número de palabras analizado puesto que, aunque se trataba de un corpus paralelo, se estudiaron 136.514 palabras (71.651 en español y 64.863 en inglés). El apartado incluido al final del artículo, denominado "Reflexiones didácticas", es considerado por las autoras como otro punto fuerte de la publicación, ya que incide en la falta de profesionalización que existe actualmente en los Grados en Traducción e Interpretación, por ejemplo, en el ámbito de la traducción turística. Asimismo, en ese mismo apartado, se incide en la necesidad de mostrar la utilidad del corpus paralelo, así como de herramientas de compilación de corpus y concordancia para suplir la ausencia de terminología técnica en el campo de la traducción de negocios. Por último, cabe destacar que la novedad del estudio lo convierte en un documento de gran interés para el mundo científico. No existen en la actualidad investigaciones que aúnen el estudio de la traducción turística y de negocios aparte de los llevados a cabo por las autoras.
Oportunidades. Análisis externo
La forma en la que se plantea la investigación y el hecho de que las autoras se hayan centrado únicamente en guías promocionales en soporte electrónico permite continuar con el análisis y compararlo con otros documentos relacionados con el turismo de negocios que se publiquen en otro tipo de formatos. Asimismo, pueden servir como guía para futuros análisis de otro tipo de documentación del sector turístico, en general, o del de negocios, en particular.

Tabla 3. Análisis de la publicación 2. Fuente: Elaboración propia.

4.3. Publicación 3. Compliance or cumplimiento normativo? A corpus study with professional and didactic purposes in the Spanish press

En esta publicación se analiza la frecuencia de uso en prensa en inglés o español de un término de gran relevancia en el ámbito jurídico-económico en 2017 y 2018 como es el caso de *compliance*. Este artículo fue publicado en el volumen 14, del año 2019, de la Revista de Lingüística y Lenguas Aplicadas. El corpus se elabora a partir de textos de dos periódicos generalistas (*El País* y *La Vanguardia*) y dos especializados (*Expansión* y *El Economista*). Se realizó para los cuatro periódicos una búsqueda a través de sus webs de los artículos que incluyeran la palabra *compliance*. Dicha búsqueda arrojó datos interesantes sobre la frecuencia de aparición de la palabra en periódicos españoles en el periodo comprendido entre enero de 2017 y los diez primeros meses de 2018 sumando un total de 1.152 apariciones entre los cuatro periódicos. Asimismo, también se realiza la búsqueda entre el primer mes del año 2010 y el primer mes del año 2018 para poder comprobar la frecuencia de uso del término en ocho años y compararlo con su aparición en los periódicos durante un año y diez meses. Los datos muestran que durante ocho años el termino apareció 1.248 veces, mientras que en año y diez meses se empleó en 1.152 ocasiones.

Debilidades. Análisis interno
Los factores que se pueden percibir como una carencia por parte de los lectores y que igualmente podrían reducir el éxito de la publicación son el periodo de análisis del término (2017 y 2018), lo que hace que con el paso del tiempo los resultados dejen de estar actualizados. Esto tiene especial relevancia en el ámbito económico-financiero, donde la rapidez con la que se transmite la información y se generan nuevos conceptos hace que sea necesario contar con datos actualizados. Del mismo modo, la muestra se reduce a cuatro periódicos (dos especializados y dos generalistas) y para ampliar la validez de los resultados hubiera sido necesario llevar a cabo un análisis más amplio que por las limitaciones de espacio de la publicación no se realizó. Por último, el corpus puede utilizarse como recurso para la traducción o docencia.
Amenazas. Análisis externo
Los obstáculos a los que se podría enfrentar la publicación tendrían relación con la posible aparición de publicaciones sobre el mismo tema que incluyeran un corpus más amplio y que abarcara incluso periódicos de otros países y otros idiomas. Esto permitiría obtener un mayor número de resultados y analizar el uso del término con más detalle desde la perspectiva de la lingüística computacional (Villayandre Llamazares 2008).

Fortalezas. Análisis interno
Existen estudios previos que analizan el uso de anglicismos en el campo económico-financiero, pero ninguno se ha detenido hasta la fecha en el término *compliance,* el término de moda en el ámbito empresarial en el año 2017 y 2018 (*Expansión* 2018; *El País* 2018). Por tanto, al centrar el análisis a estos dos años, los resultados dan muestra de su uso en un periodo en el que por el contexto económico de cambios que se producían en las empresas, el *compliance* tenía especial relevancia. Del mismo modo, la calidad de su compilación tal y como señalan Sanz, Zubillaga y Uribarri (2015), permite indicar que los datos obtenidos aportan de manera fiable información sobre un término del campo económico-financiero de reciente aparición y sobre el que no existen publicaciones previas. Asimismo, se compara la frecuencia de uso del término en esos dos años y en un periodo comprendido desde 2010 para observar la mayor recurrencia del mismo durante ese "periodo de moda". Los recursos utilizados de cuatro periódicos (dos generalistas y dos especializados) permiten igualmente comparar su uso en función del nivel de especialidad de la publicación y en los periódicos que cuentan con el mayor número de lectores según los datos publicados hasta esa fecha por el Estudio General de Medios (EGM). Consideramos que el hecho de ser un análisis muy específico y a la vez comparativo permite al lector beneficiarse de esta publicación desde dos ángulos: el didáctico, al poder emplear el estudio como ejercicio de clase para este mismo término u otros que vayan surgiendo en este ámbito y, por otro lado, el profesional, al aportar información contextualizada sobre el uso de un término específico en un periodo determinado. Por último, no existen más publicaciones en el ámbito traductológico que aborden de forma específica el análisis de este término y no consideramos que ninguna de las debilidades mencionadas pueda afectar la transferencia o replicabilidad de la publicación, ya que solamente la limitan sin afectar esto a la validez de los resultados. En el ámbito académico se estudia el tema del *compliance* desde la disciplina del Derecho y la Economía, pero no se han encontrados estudios realizados desde una perspectiva traductológica.
Oportunidades. Análisis externo
La influencia del inglés en el ámbito empresarial y el uso de anglicismos es una realidad que debe ser estudiada en detalle por el peso del componente léxico en esta especialidad y su importancia en la traducción. No se aprecian tendencias similares recientes en este campo sobre el tema por lo que sería necesario seguir investigando al respecto para conocer la evolución del término y si progresivamente se opta por su traducción al español. De esta forma, consideramos relevante extender y ampliar su análisis y observar además su uso en manuales o casos jurídicos en los últimos años.

Tabla 4. Análisis de la publicación 3. Fuente: Elaboración propia.

4.4. La traducción en el ámbito de Arduino: propuesta de glosario inglés-español

En este artículo, publicado en *Quaderns de Filologia: Estudis Lingüístics* XXIV en el año 2019, se procede a la compilación de un corpus monolingüe en inglés, formado por 253.588 palabras, a partir de diez manuales de Arduino (plataforma de código abierto) para crear una propuesta de glosario al español de los términos que mostraran una mayor frecuencia de aparición.

Debilidades. Análisis interno
La principal carencia de este artículo y que podría igualmente reducir el éxito de la publicación es el número limitado de términos que se incluyen en el glosario, ya que por limitaciones de espacio se redujeron a veinte.
Amenazas. Análisis externo
El principal obstáculo es el rápido avance del ámbito de la informática, lo que supone que surjan nuevos términos y otros puedan caer en desuso. Por tanto, aunque el objetivo del artículo sea contribuir a la normalización terminológica, solo en unos años se podrá comprobar si los resultados obtenidos han resultado útiles a estos efectos. No existen actualmente recursos fiables sobre Arduino ni publicaciones desde la perspectiva traductológica al respecto. No obstante, la calidad de compilación del corpus supone que la metodología utilizada podría replicarse y transferirse sin ningún problema a otras publicaciones e incluso ampliarla para estudios futuros.
Fortalezas. Análisis interno
La principal ventaja competitiva de este estudio es que aporta un recurso fiable que permite al traductor contar con equivalentes de términos que, al tratarse de un ámbito de gran novedad como el de Arduino, no aparecen de forma extendida en glosarios o bases de datos. Se analizaron un total de 253.588 palabras procedentes de diez manuales técnicos sobre el tema que fueron escogidos de acuerdo con los expertos en telecomunicaciones con los que trabajamos. Esta colaboración traductor-experto aporta igualmente un valor añadido al glosario de la publicación del que se pueden beneficiar los usuarios.
Oportunidades. Análisis externo
No se observan en la actualidad investigaciones que se centren en la parte lingüística de Arduino, pero sí es un tema de gran relevancia y actualidad y que cada vez tiene más demanda en el mercado. Por ello, a medida que avance surgirán nuevas realidades lingüísticas a las que habrá que ir dando respuesta desde el ámbito de la traducción.

Tabla 5. Análisis de la publicación 4. Fuente: Elaboración propia.

5. Conclusiones

El objetivo de este estudio era realizar una reflexión y hacer autocrítica sobre cuatro publicaciones elaboradas por las autoras y cuya base metodológica era un estudio de corpus. Para ello, en primer lugar, en la introducción esbozamos de forma preliminar en qué consisten este tipo de estudios y su importancia y aplicación en el ámbito de la traducción. Posteriormente, en el marco teórico realizamos un recorrido a través de la evolución de los estudios de corpus en este campo y quedó constancia de la tendencia al alza en la publicación de estudios que emplean esta metodología entre los años 2000 y 2010, pero su actual estancamiento e incluso tendencia a la baja. En cuanto a la metodología que aplicamos para este artículo, consideramos que hoy en día, debido a la caracterización actual del mundo académico por el que se nos exige un gran número de publicaciones de calidad, era importante detenernos también a reflexionar sobre el trabajo realizado y determinar líneas futuras de actuación. Por ello, consideramos que la metodología DAFO adoptada nos permitía analizar con detalle los puntos positivos y negativos de cuatro trabajos previos realizados mediante el uso de corpus, de gran relevancia en el ámbito de la traducción por su utilidad tanto para la docencia como para la profesión. Entre las preguntas que nos planteábamos para las debilidades se encontraban los factores que podían ser percibidos como tal por parte de los lectores y los que podrían reducir el éxito de la publicación. En cuanto a las fortalezas, buscábamos analizar las ventajas competitivas de la publicación, los resultados obtenidos en comparación con otros estudios publicados, los recursos y los factores que podrían ser considerados ventajosos por parte de los lectores. Por otro lado, para las amenazas nos deteníamos en los obstáculos de la publicación, otros temas en los que otros académicos estuvieran trabajando, datos que existieran actualmente o debilidades que pudieran afectar a la transferencia o replicabilidad de la publicación. Por último, las oportunidades que suponía la publicación giraban en torno a las tendencias de investigación, si resultaba necesario seguir investigando en el tema y cambios que pudieran propiciar que la publicación tuviera una mayor transferencia o replicabilidad.

De esta forma, como hemos podido observar en el análisis de las cuatro publicaciones seleccionadas para este estudio, en todas encontramos

debilidades, amenazas, fortalezas y oportunidades que nos han permitido reflexionar sobre nuestra investigación para así darnos cuenta de qué aspectos considerábamos que habíamos realizado de forma correcta y cómo podemos a partir de ahora seguir avanzando, teniendo en cuenta lo que aún se puede mejorar. Como debilidades comunes cabe destacar el avance de los campos temáticos en los que hemos centrado nuestra investigación, lo que hace que los datos obtenidos puedan quedar rápidamente desactualizados con el paso del tiempo. Del mismo modo, el tamaño del corpus realizado en las publicaciones mencionadas podría seguir ampliándose y así obtener más datos sobre el discurso analizado. No obstante, estas mismas debilidades podrían convertirse en oportunidades de investigación para seguir avanzando en esta misma línea en el campo de la traducción, ya que los campos temáticos en los que se centran los estudios resultan de gran interés, originalidad y relevancia. Como fortalezas hemos destacado la utilidad de los datos obtenidos que pueden tener tanto una aplicabilidad práctica para la profesión como para la didáctica de la profesión, así como la novedad de los temas tratados, ya que la combinación de finanzas y turismo no se ha tratado hasta la fecha en profundidad.

De esta forma, podemos afirmar que la realización de una autocrítica a publicaciones propias mediante la aplicación de una metodología DAFO nos ha permitido ganar una perspectiva de trabajo que no habíamos alcanzado hasta la fecha y poder así afrontar retos futuros de investigación con mayor madurez y conocimiento sobre qué debemos hacer para que los próximos estudios de corpus en los que trabajemos contribuyan aún más a reforzar el avance de nuestra disciplina.

Referencias bibliográficas

ALIAGA, Francisco M.; Calisto GUTIÉRREZ-BRAOJOS & Antonio FERNÁNDEZ-CANO. (2018) "Las revistas de investigación en educación: Análisis DAFO." *Revista de Investigación Educativa* 36:2, pp. 563-579.

BAKER, Mona; Gill FRANCIS & Elena TOGNINI-BONELLI. (1993) *Text and Technology*. In honour of John Sinclair. Ámsterdam: John Benjamins Publishing Company.

Bowker, Lynne & Gloria Corpas Pastor. (2014) "Translation Technology." En: Mitkov, Ruslan (ed.) *Translation technology handbook of computational linguistics*. Oxford: Oxford University Press.

Bowker, Lynne & Jennifer Pearson. (2002) *Working with Specialized Language. A Practical Guide to Using Corpora*. Londres: Routledge.

Casero Gimón, José Luis. (2019) *De la idea de negocio al plan de empresa*. Madrid: UDIMA.

Chomsky, Noam. (1957) *Syntactic Structures*. The Hague: Mouton & Company.

Corpas Pastor, Gloria. (2012) "Corpus, tecnología y traducción." En: Casas Gómez, Miguel & María García Antuña (eds.) *Jornadas de Lingüística*, XII (Cádiz, del 30 de marzo al 1 de abril de 2009). Cádiz: Servicio de Publicaciones de la Universidad de Cádiz, pp. 75-98.

Corpas Pastor, Gloria. (2008) *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Madrid: Peter Lang.

De Felipe Boto, María del Rosario. (2007) "Los estudios basados en corpus en Traducción." *Interlingüística* 17, pp. 261-267.

EAGLES (1996) "Text Corpora Working Group Reading Guide." *Documento Eagles (Expert Advisory Group on Language Engineering)* EAG-TCWG-FR-2.

Espinosa, Roberto. (2020) "La matriz de análisis DAFO." Versión electrónica: <https://robertoespinosa.es/2013/07/29/la-matriz-de-analisis-dafo-foda>

Expansión (2017) "El 'compliance', tema jurídico del año.

Expansión." Versión electrónica <https://www.expansion.com/juridico/actualidad-tendencias/2017/06/26/5951438fe5fdea511e8b45f0.html>

Firth, John Rupert. (1957) *Papers in Linguistics 1934-1951*. Londres: Oxford University Press.

García Ferrer, Mercedes. (2016) "Diseño y construcción de un corpus de referencia de latín." *Methods* 3.

Johns, Tim. (1991) "From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning." *ELR Journal* 4.

Laviosa, Sara. (2002) *Corpus-based translation studies. Theory, findings, applications*. Ámsterdam/Nueva York: Rodopi.

Morón Martín, Marián. (2009) "Perfiles profesionales en Traducción e Interpretación: análisis DAFO en el marco de la sociedad multilingüe y multicultural." *La linterna del traductor*. Versión electrónica: <http://www.lalinternadeltraductor.org/n4/dafo-traduccion.html>

NAPU, Novriyanto. (2016) "Translating tourism promotional texts: translation quality and its relationship to the commissioning process." *Cultus, The Journal of Intercultural Mediation and Communication* 9:2, pp. 47-62.

PÉREZ HERNÁNDEZ, Chantal. (2002) "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento." *Estudios de Lingüística del Español* (ELiEs). Universidad de Málaga. Versión electrónica: <http://elies.rediris.es/elies18/index.html>

PLAZA LARA, Cristina. (2019) "Análisis DAFO sobre la inclusión de la traducción automática y la posedición en los másteres de la red EMT." *Journal of Specialized Translation* 31, pp. 260-280. Versión electrónica: < https://www.jostrans.org/issue31/art_plaza.pdf>

SÁNCHEZ RAMOS, María del Mar. (2017) "Metodología de corpus y formación en la traducción especializada (inglés-español): una propuesta para la mejora de la adquisición de vocabulario especializado." *Revista de Lingüística y Lenguas Aplicadas* 12, pp. 137-150.

SANZ, Zuriñe; Naroa ZUBILLAGA & Ibón URIBARRI. (2015) "Estudio basado en corpus de las traducciones del alemán al vasco." En: Sánchez Nieto, María Teresa (ed.). 2015. *Corpus-based Translation and Interpreting Studies / Estudios Traductológicos basados en corpus*. Frank & Timme.

SISAMÓN GIL, Rosa María. (2012) "El análisis "DAFO" aplicado a la intervención en casos de personas en situación de exclusión social." *Documentos de Trabajo Social* 51, pp. 469-487.

SINCLAIR, John. (1991). *Corpus Concordance Collocation*. Oxford University Press.

SULAIMAN, M. Zain & Rita WILSON. (2018) "Translating tourism promotional materials: a cultural-conceptual model." *Perspectives* 26:5, pp. 629-645.

TORRES-SALINAS, Daniel; Rafael RUIZ-PÉREZ & Emilio DELGADO-LÓPEZ-CÁZAR. (2009) "Google Scholar como herramienta para la evaluación científica." *El Profesional de la Información* 18:5, pp. 501-510.

VILLAYANDRE LLAMAZARES, Mika. (2008) "Lingüística con corpus." *EH Filología* 30, pp. 329-349.

VILLEGAS, Bayardo. (2003) "Rápida y pertinente búsqueda por internet mediante operadores booleanos." *Universitas Scientiarum* 8, pp. 51-54.

NOTA BIOGRÁFICA / BIONOTE

Alexandra Santamaría Urbieta es Doctora en Traducción Turística por la Universidad de Las Palmas de Gran Canaria y actualmente trabaja como profesora e investigadora en la Universidad Internacional de La Rioja (España). Es licenciada en Traducción e Interpretación por la Universidad del País Vasco y su faceta como traductora la ha llevado a impartir asignaturas de traducción especializada en varias universidades nacionales e internacionales. Asimismo, ha hecho públicos los resultados de las investigaciones realizadas en el campo de la traducción en varios congresos.

Alexandra Santamaría Urbieta holds a PhD on Tourist Translation from the University of Las Palmas de Gran Canaria and currently works as a lecturer and researcher at the International University of La Rioja (Spain). She has a degree in Translation and Interpreting by the University of the Basque Country and her facet as a translator has led her to teach specialized translations subjects in several national and international universities. Likewise, she has made public the results of the investigations carried out in the field of translation in several congresses.

Elena alcade peñalver es doctora en Traducción por la Universidad de Granada y es actualmente profesora e investigadora en la Universidad de Alcalá (Madrid, España). Pertence al grupo de investigación Fitispos (Formación e Investigación en Tradución e Interpretación en los Servicios Públicos) de la Universidad de Alcalá y tiene experiencia docente e investigadora en el ámbito de la traducción especializada a nivel internacional.

Elena alcade peñalver holds a PhD on Translation from the University of Granada (Spain) and works as a lecturer and researcher at the University of Alcalá (Madrid). She is part of the Fitispos (Training and Research in Public Service Translation and Interpreting) research group at the University of Alcalá. She has teaching and research experience at international level in the field of specialized translation.

Recibido / Received: 23/05/2020
Aceptado / Accepted: 27/07/2020

Para enlazar con este artículo / To link to this article:
http://dx.doi.org/10.6035/MonTI.2021.13.10

Para citar este artículo / To cite this article:
Buts, Jan & Henry Jones. (2021) "From text to data: mediality in corpus-based translation studies." En: Calzada, María & Sara Laviosa (eds.) 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS spring-cleaning: A critical reflection*. *MonTI* 13, pp. 301-329.

FROM TEXT TO DATA: MEDIALITY IN CORPUS-BASED TRANSLATION STUDIES

JAN BUTS
butsj@tcd.ie
Trinity College Dublin, Ireland

HENRY JONES
h.jones4@aston.ac.uk
Aston University, UK

Abstract

This paper seeks to promote deeper reflection within the field of corpus-based translation studies (CTS) regarding the digital tools by means of which research in this discipline proceeds. It explicates a range of possibilities and constraints brought to the analysis of translated texts by the keyword in context (KWIC) concordancer and other data visualisation applications, paying particular attention to the ways in which these technological affordances have actively shaped central theoretical hypotheses within CTS and related fields, as well as the general principles of corpus construction. This discussion is illustrated through a small case study which applies the suite of corpus analysis tools developed as part of the Genealogies of Knowledge project to the investigation of two English translations of the *Communist Manifesto*.

Keywords: Mediality; Digital technologies; Data visualisation; Corpus construction; KWIC concordancer.

Résumé

Cet article cherche à stimuler une réflexion plus approfondie dans la traductologie de corpus concernant les outils numériques au moyen desquels la recherche est menée

dans cette discipline. L'article explique diverses possibilités et contraintes de l'analyse des textes traduits, assistée par un concordancier « KeyWord In Context » (KWIC) et par d'autres outils de visualisation des données. Une attention particulière sera portée à la manière dont ces affordances technologiques ont façonné des hypothèses théoriques centrales ainsi que les grands principes de la construction de corpus, dans l'approche traductologique basée sur le corpus et dans d'autres domaines proches. Cette discussion est illustrée par une étude de cas appliquant la suite logicielle développée pour l'analyse de corpus dans le cadre du projet Genealogies of Knowledge, à deux traductions en anglais du *Manifeste communiste*.

Mots-clés : Medialité; Technologies numériques; Visualisation des données; Construction de corpus; Concordancier KWIC.

1. Introduction

The use of text corpora for the investigation of language predates the invention of the modern computer (Fenlon 1908; Svartvik 1992: 7). Nevertheless, the success of corpus-based methodologies across the humanities today is primarily associated with the application and assistance of digital technologies in the research process (Luz & Sheehan 2020: 2-3). Beginning with Roberto Busa's work on the *Index Thomisticus* in the 1940s, corpus analysts have exploited the processing power of computers to facilitate the investigation of linguistic patterns repeated across ever larger collections of text. Such techniques rely fundamentally on the core principle of digital media, namely, numerical representation: the ability of the computer to transform any media object into the standardised language of mathematics (Manovich 2001). In order to be interrogated using corpus analysis software, a text must first be digitised; the alphabetic characters or logograms through which its contents are expressed must be converted into a binary code of 1s and 0s, itself an abstract representation of voltage, to be stored and interpreted by the machine. Once the information is stored, each token – a delineated string of characters in the corpus, often corresponding to a word – must be indexed (Luz 2011: 137-139). Tokens are assigned a numerical value that records the items' exact location in the source material. Finally, the researcher's ability to interact with the corpus depends on visualisation tools, such as the classic keyword in context (KWIC) concordance display, which can convert this

mathematical information and reconstitute fragments of the original corpus texts on the screen.

The transition from an analogue to a digital work environment and the shift towards binary code as the lingua franca of the twenty-first century has inaugurated a paradigm change within several established scholarly disciplines, but also inspired the creation of hybrid, interdisciplinary approaches to knowledge production. In this respect, *digital humanities* has come to serve as an umbrella term for a variety of practices, including the digitization of texts and artefacts, the study of born-digital material, as well as the development of digital tools and the new methods they facilitate (Sheridan 2016). A key theme within this field of inquiry has been a focus on questions of materiality, in part because technological changes have eroded the seemingly self-evident qualities of previous objects of study and their interpretation: if the digital medium has rendered text more dynamic, its previously more static qualities gain in significance. Similarly, if a hyperlinked environment invites non-linear browsing, reading conventions require renewed attention. In this regard, changes in literacy demands are now central to semiotic debate (Kress 2003). Several scholars have additionally begun to address self-reflexively the emphasis placed on pattern recognition in much digital humanities research, and to critically examine the implications of our focus on patterns – rather than structures or narratives – as the primary objects of study (Dixon 2012; Berry 2011).

In this broad context, the field of translation studies too has shown growing interest in the semiotic and material media in and through which translations are stored, transmitted and – by extension – studied (Armstrong 2020; Pérez-González 2014). There is growing recognition within the discipline that media tools such as books, newspapers, websites and DVDs are not passive conduits for the transmission of information, nor are they inert containers for its storage. Rather, they have their own mediality, they offer their own unique sets of techno-social possibilities and constraints, and they can thus more accurately be considered 'environments' that shape every aspect of our engagement with a text (Jones 2018). Littau (2011, 2016), for example, has explored a series of media-induced transformations in reading, writing and translation practices throughout history, from the oral culture of Ancient Rome through to the network culture of today's digital world. Jones

(2018) has examined how changes in the tools used to produce, distribute and consume audiovisual products during the twentieth century have influenced not only the approaches to translation adopted by subtitlers and their commissioners, but also the ability of ordinary consumers to get involved in this process. Finally, Cronin (2013) has discussed in more general terms the value of recognising the transformative potential of computing as an activity that reconfigures relationships between languages, cultures and texts.

Surprisingly, however, the subdiscipline of corpus-based translation studies (CTS) has remained largely silent on these issues, an omission which may seem particularly striking given the extent to which scholars based in CTS depend upon and interact with technology to enact their research. Reasons for this lacuna may vary. Considerations of mediality in CTS research would have been too distracting if investigated at the outset. When Baker (1993: 243) introduced corpus techniques in translation studies in order to elucidate "the nature of translated text as a mediated communicative event", this left no immediate room to consider the method of study as a mediated communicative event in itself. In addition, technological hesitancy may prompt researchers to turn away from interrogating the tools they use, and to remain silent on the topic. Practitioners of close reading are seldom acquainted with the physics of vision, and even less are they expected to be, yet a lack of digital literacy can inspire scholarly shame and therefore lack of transparency in the dissemination of tools and methodological pathways. Whatever the cause, the convertibility of the sign and its attachment to the binary standard are yet to be consistently questioned.

This paper argues that CTS requires a sustained interrogation of its practices in relation to its conditions of existence: the transformation, by means of code, of text that can be read into data that can be queried. The field of translation studies has successfully demonstrated that expressions across languages are declared rather than found equivalent (Hermans 2007: 6; Tymoczko 2010: 3). We argue that the same holds for expressions across medial environments, and that this includes the setting in which such expressions are studied. Equivalence between a dataset and the objects of study it is made to represent depends upon a specific, situated agreement on perceptual conventions established within a given research culture. Building on these principles, we seek to encourage deeper reflection within CTS and

related fields on the mediality of corpus research. This disciplinary 'dusty corner' is addressed in the following section mainly with reference to the role of the KWIC display in CTS research, which illustrates the convergence of the technological and theoretical boundaries of the field of study. Reflections on the limitations of KWIC analysis and of alternative forms of visualisation are not absent from the field of corpus research (e.g. Anthony 2018). Yet in this article, we seek to shed light not only on the restrictive, but also on the transformative aspects of using a given tool. The logic of the medium is shown to inform not only the mode of analysis, but also various aspects of corpus construction and representation. In the final sections this general account of mediality in CTS is instantiated by means of a discussion of the *Communist Manifesto* as it appears in the Genealogies of Knowledge (GoK) corpora. We illustrate that, just like multiple retranslations of a single text can, even within the space of a phrase, produce widely divergent images that invite a range of different interpretations, varying digital representations of textual material are only equivalent to their source insofar as the medium's mutational qualities are left uninterrogated.

2. A Medium Shaping a Discipline: CTS and the KWIC concordance

The term *medium* typically denotes a channel of communication and does so in abstract fashion. The medium is 'television' rather than the television set. In this sense, any medium is intimately connected to a technology, of which particular tools are instantiations. Different compounds such as *print media* and *social media* highlight different characteristics of a medial environment – among other aspects, one can focus on material conditions (as in 'print') or pragmatics (as in 'social'). The divide between these main aspects of media technologies often gives rise to different perceptions of how a medium takes shape and influences its users. On the one hand, a focus on the constraints enforced by new media technologies may lead to technological determinism: the tools you use condition the actions you undertake, and ultimately, the thoughts you have. This line of thinking has come to be associated with McLuhan's (1964: 7) mantra that "the medium is the message". On the other hand, a focus on the way new media technologies are received and put to particular uses by their human users may foreground the social construction

of technology (Klein & Kleinman 2002): the nature of our adoption of and engagement with technological innovations is the result of choices made at the individual as well as group level. The distinction between the material and the pragmatic perspective in the study of media thus runs along the classic divide between agency and structure. As several commentators since McLuhan have noted (Cronin 2013; Littau 2006), the most productive accounts of our ultimately social reality will take both perspectives into account, thus treating a medium as a compromise: a consensus arrived at by agents to act in accordance with a constraint imposed by a tool that, on the whole, facilitates a common goal.

Within CTS, this common goal is fundamentally the ability to identify and interrogate recurring patterns within large collections of written or transcribed text, selected according to a specific set of criteria, held in machine-readable form, and compiled in order to investigate a hypothesis about the process or products of translation (Baker 1995: 225). Such a collection of texts is itself little more than a mute database. It is the mode of access that turns a collection into a corpus in the modern sense: collected text is to be "analysed automatically or semi-automatically using different types of software specifically created for linguistic research" (Malamatidou 2018: 43). A certain vagueness about the type of software to be used has been part of discipline-defining statements in CTS since its early days, but in actual fact the predominant mode of display throughout the last three decades has been the KWIC concordance, identified by Baker (1995: 226) as "the corpus analyst's stock in trade". Indeed, the concordance line is the commodity the analyst deals in, as well as the currency that guarantees, through the provision of textual evidence, trust and recognizability within a larger research community.

Thus, in the case of CTS, one should not just consider 'the computer' or 'the screen' as the medium through which research proceeds. The KWIC concordance interface, a specific application of the computer as environment, is the prime medial display. In the following paragraphs we argue that the concordancer as medium is not a neutral tool of representation: not only does its design reflect a specific set of concerns among researchers interested in language use, but its affordances have also actively shaped central theoretical hypotheses within CTS and related fields, as well as the general principles

of corpus construction. None of what follows is intended to suggest that the principles of corpus research are defective, or that KWIC analysis is not informative. Rather, the examples we offer serve to indicate that there is a myriad of ways to approach language, and the tools we use put in place specific constraints that narrow down the possibilities productively, but in the process also transform the qualities of the textual material in view as well as our intuitions about it.

2.1 Theory and Technology

Electronic KWIC retrieval predates CTS by quite a stretch and is typically traced back to the work of Luhn (1960). The prior development of this technique can itself be situated within a much longer history of producing analogue concordances recording every instance of an alphabetical list of keywords along with a snippet of the immediate co-text for each occurrence. In addition to providing a solution to specific problems (for example, the need to identify suitable passages from the Bible with which to illustrate a sermon – Fenlon 1908), such endeavours reflect a fundamental interest – shared among theologians, philologists and ultimately linguists – in the iterability of the sign, or the ways in which lexical repetition and contextual variation constitute meaning.

For early adopters such as the linguist and lexicographer John Sinclair, whose work provided the main inspiration for the use of corpora in translation studies, the assistance of the machine offered a powerful means of developing more objective and productive methodologies for linguistic analysis, in marked contrast with previously dominant introspective approaches to the study of language (McIntyre & Walker 2019: 6; Stubbs 1996: 24). In particular, the use of an electronic KWIC concordancer made vastly more accurate and efficient the study of collocation, understood as the frequent co-occurrence of linguistic elements (Firth 1968: 14). It is important to note, however, that in the Firthian tradition out of which Sinclair's work emerged collocation had been understood to depend upon a broad, situational understanding of context. For Firth collocational patterns could also consist of sonorous markers such as alliteration, or organisational features such as paragraph structure (Partington 1998: 16-17). Once the KWIC display became

the dominant interface for accessing corpora, however, this tool came to strongly inform the theoretical priorities of linguistic research, as can be illustrated with reference to hypotheses about extended units of meaning.

One of the major arguments of corpus linguistics is that it can offer proof that individual words cannot be considered clear and distinct units of meaning (Hunston 2007: 250; Sinclair 2008: 409). Notably, the alternative proposals for extended units of meaning have been made conveniently in tune with the constraints of the medium. Semantic prosody, for instance, is a corpus-based theoretical innovation that refers to the evaluative or attitudinal function shared by a sequence of frequently co-occurring lexical items (Stubbs 2009: 124-125). The first elaborate illustration of the phenomenon can be found in Louw (1993), and despite some valid criticism regarding the coherence of the concept (Hunston 2007; Stewart 2010; Whitsitt 2005), semantic prosody has received sustained attention in corpus-based translation studies (Munday 2011; Stewart 2009). Louw's (1993: 170) original argument was rich in examples, such as the attested use of *symptomatic of*, a phrase which tends to be followed by a lexical item with a negative overtone (e.g. *tensions, inadequacies*), and which is thus argued to be itself indicative of a negative attitude towards a given topic. At times, however, examples are found where speakers use *symptomatic of* followed by a supposedly positive reference, as in *symptomatic of our good reputation*. According to Louw, this suggests that the speaker might be ironic, insincere, or perhaps even unaware of their own attitude.

Throughout Louw's article one finds a depreciation of introspective linguistics, and a plea for the corpus-based method to be adopted (Louw 1993: 173). Louw provides the reader with sets of concordance lines at several stages, because the crucial factor in identifying a semantic prosody, identifying deviations from it, as well as coming to grips with the reason for these deviations, is the perusal of a sequence of similar phrases. If some theoretical innovations can only be revealed as well as illustrated through concordances, the medium becomes indispensable to linguistic research, and ultimately determines the discipline's view of language. Indeed, in response to the received lexicographical paradigm and its focus on the confines of the individual word as the primary unit of meaning in any given language, Sinclair (2004: 34) would ultimately come to argue that semantic prosody

was a more likely candidate to constitute "the boundary of the lexical item". The markers of semantic prosody are typically situated within a very limited span on either side of the keyword, a span that has more in common with the KWIC concordance than with sentence structure or features of text organization. Once lexis is severed from the writing system's main word divisions, there is no obvious reason to situate the boundaries elsewhere, whether at two, three, or twenty positions from the keyword, yet the concordance view strongly nudges one first to restrict collocational relevance to a limited span, and secondly to interpret this restriction as meaningful.

CTS research has examined whether there is a connection between the idea of an extended unit of meaning and the idea of the translation unit, or "the smallest segment of the utterance whose signs are linked in such a way that they should not be translated individually" (Kenny 2011; Vinay & Darbelnet 1995: 21). Tognini-Bonelli (2001: 133) argues that the two notions are different: the unit of meaning is a mere linguistic convention, while a translation unit is strategic, and "the result of explicit balancing decisions", including broad contextual considerations. Linguistic conventions, however, are equally context-dependent formalizations of speakers' strategies, and it is highly likely that the difference posited is an effect of the medium: looking at text in a different manner may suggest that it has been produced according to different motivations, but what suggests itself as a theoretical discrepancy is in fact a technological one. Every utterance expresses concern for macro-structural features, even though the decontextualised lines of a concordance may temporarily suggest otherwise.

Apparent disregard for integral texts as communicative units relates not just to corpus analysis, but in some cases extends to corpus construction, as is evident from the use of sample corpora, which are made up of parts of originally longer texts, and which reveal a belief in the primacy of patterns over narrative or text structure as objects of study in CTS and across the broader field of the digital humanities. Sampling is a method developed to ensure representativeness, a factor that became important in relation to the claim that corpora provide a view of natural language in use. Early lexicographical work dictated that potential variety must be investigated, meaning that corpus representativeness was equated with a search for either comprehensiveness (the corpus should contain as many different texts as possible)

or balance (the corpus should contain a comparable number of tokens for each kind of text selected). However, neither balance nor comprehensiveness necessarily produce good intimations of linguistic and ultimately cognitive reality. Effective propaganda or advertising, for instance, typically produces limited unique textual output. In any given language, Coca-Cola slogans throughout history would barely fill a page, but they are read and heard many times over, with significant psychological and economic consequences (see also Baker *et al*. 2008: 283). Despite many corpora today focusing on matters of ideology and influence, textual balance still overrides repetition and textual impact in the selection of corpus materials. Thus, despite aiming to represent language in use, no corpora take actual use into account as a construction principle.

Research into the connection between speech and cognition, such as Hoey's (2007, 2011) work on lexical priming, attaches great importance to repetition. Words are encountered in a certain context, come to be associated with it, and thus are produced again when a similar context is encountered. The field of corpus-assisted discourse studies (e.g. Baker 2006) presents repetition as a means of persuasion and a hallmark of ideology. Lexical priming and discourse analysis have both found ample application in CTS, and generally, identifying repeated collocations or other lexical patterns can be seen as the central objective of corpus research. However, while repetition is sought after in corpus analysis, variation is central to corpus construction. If texts were represented multiple times within a corpus, the concordance would not be helpful, as it would return mantras of identical lines, meaningless without the immediate provision of adequate context outside of the medium. When Luhn introduced the electronic KWIC concordance, he devised it as a tool to query indexes of technical literature in order to find material relevant to one's study. In this context, duplicating entries would not have made sense. The expansion of the use of the concordance in CTS, as in other disciplines, to include matters of social and political impact means that the absence of duplicate texts is now a mere convention, partly sustained by the affordances of the tool.

2.2 Description and Representation

It is not only the theoretical priorities of linguistic analysis that are shaped by the mediality of the CTS research environment. The processes of preparing, describing and representing corpus texts are also influenced to a large extent by the affordances of the digital medium, beginning, for instance, with the use of a document type definition (DTD). The DTD is a common means of determining how documents stored in a corpus database are to be interpreted by an extensible markup language (XML) application such as a concordance browser (Luz 2011: 133; Zanettin 2011: 112). The DTD explicates the guidelines for producing a valid XML document for a given document type. Typically, a DTD will resemble a minimal grammar: it consists of a skeletal set of elements complemented with a list of potential attributes. XML is used for markup, meaning that its tags are normally not displayed in the concordance output. Nonetheless, all concordance lines returned in a corpus search will have been matched to a defined category, and potentially have undergone structural alterations for this purpose. These are made in addition to the inevitable changes in material texture, font, size, imaging, colour, location and so on that already affect each element of a publication prepared for inclusion in a corpus. Consequently, a concordance line is always a back-translation. The text it contains has first been translated into a form that conforms to a markup syntax that the software can interpret, and then is subsequently returned to the concordance user as a representation of the original text. This representation is declared equivalent to its source but at least implicitly operates along different linguistic as well as material constraints.

The conventional separation between text and paratext may serve as an example. Typically, the bottom of a page is reserved, where relevant, for footnotes in small font. Footnotes do not belong to the main argument and are therefore placed outside the main verbal sequence. As per publishing conventions, footnotes may be provided either by authors or by additional contributors such as translators or editors. When constructing a corpus, it might be advisable to indicate what constitutes the main text of a document, and what constitutes paratext, so the DTD will need to specify a

syntax fit to represent this distinction. However, are footnotes essentially different from endnotes, and should this distinction be encoded? Are they like introductions, meaning that a 'paratext' element can cover both? What about marginalia, written, like a footnote, by a supplemental contributor and apart from the main chain of information, but lacking a footnote's expected markers, such as the use of ordinal numbering?

The latter issue may be partly responsible for the relatively small number of diachronic studies in CTS (Malamatidou 2018: 51), although sophisticated examples are available (e.g. Gabrielatos *et al*. 2012). Languages change, but so do the documents and discursive conventions through which they do so. This makes the construction of diachronic corpora particularly challenging. How much information should be provided to adequately characterise the context of an expression, and how can there be consistency when the potential correspondences between different historical situations are limited? This problem is naturally present in CTS, as it mediates between different cultures, yet in studies covering a broad time span the issue recurs on multiple fronts. Can the same set of metadata be applied to a vellum scroll and to a born-digital blogpost? If so, which characteristics should take precedence? A document type is a conventional constraint in place to categorize relevant information that would otherwise be lost in the structural limitations imposed by adaptation to a KWIC concordance interface. When using a DTD for this purpose, there can be many specifications, but little nuance.

Multimodal corpora (e.g. Baldry & Thibault 2008; Jiménez Hurtado & Soler Gallego 2013) operate partly in response to the radical recontextualization presented by textual concordances, and they can include features such as the layout of manuscripts or the situational environment of spoken utterances. Multimodal corpora can be seen as extensions of the textual paradigm, but they also remind us that a corpus does not have to consist of text, as other communicative modalities are available. And even if a corpus consists of text, textual representation may not be the most efficient or productive way to study its contents. Indeed, a concordance interface is not a mandatory mode of access to a corpus. Most browsers come with facilities such as frequency list tools which can provide information about a corpus without ever having to produce a KWIC view. Such basic mathematical

operations are widely used, as are more complex statistical procedures that can provide multifaceted information about the constitution of a given set of translated texts (e.g. Oakes & Meng 2012). Often, statistical manipulation produces information that can be visualised in tables, graphs or charts, without any need for a concordance.

This does not mean, however, that the influence of the concordance constraint is no longer present. A great number of statistical operations applied to corpora today are variants on collocational measures such as z-score, t-score, and mutual information (Cantos, Pascual & Sánchez 2001: 202). A collocate, or co-occurring lexical element, is statistically significant when it accompanies another lexical element more often than can be expected given a degree of randomness assumed in linguistic exchange. Typically, collocation is calculated "within a specified linear distance or span" (Cantos, Pascual & Sánchez 2001: 202; Sinclair 2004). As discussed in the previous subsection, the pre-machinic Firthian notion of collocation was not necessarily restricted to a specified linear distance, nor to the lexical item. These constraints were imposed by the view presented in a KWIC concordance, and thus even when this does not form part of one's research methodology, the medium continues to exert its influence.

The landscape of the discipline is rapidly changing, and the medium around which its research practices have for a long time converged is no longer a given. Precisely at this point it is necessary, as has been attempted in this section, to indicate how an inherited medium has partly shaped practices and principles in CTS. In the second half of this paper, we will discuss mediality in corpus-based translation studies with specific reference to the construction and analysis of a series of corpora built as part of the Genealogies of Knowledge (GoK) project (genealogiesofknowledge.net/about). We will illustrate the negotiation of the issues addressed in Section 2 in conjunction with the development and use of a set of visual tools that accompany a dedicated concordance interface. We begin in Section 3 with a brief overview of this project's aims and resources, provided to set the discussion that follows in Section 4 in its proper context.

3. Genealogies of Knowledge: Aims and Resources

Genealogies of Knowledge was an interdisciplinary research project led by the Centre for Translation and Intercultural Studies at the University of Manchester and funded by the UK Arts and Humanities Research Council from April 2016 to the end of March 2020. Going forward, the Genealogies of Knowledge team continues to develop and expand its activities through a dedicated Research Network (genealogiesofknowledge.net/research-network/). The core objective of this endeavour has been to explore the role of translation and other forms of mediation in negotiating the meanings of key political and scientific concepts as they have travelled across time and space (Baker & Jones, *forthcoming*). The team is interested, for example, in how translators, commentators and other cultural mediators – including editors, historians, philosophers, citizen journalists and bloggers – have contributed to the ongoing evolution and contestation of concepts such as democracy, citizenship, truth, proof and fact when interpreting and adapting their sources for new audiences (Baker 2020; Jones 2019, 2020; Karimullah 2020). To this end, five non-parallel but closely interconnected corpora have been built, of which the largest is the Modern English corpus. This contains over 350 translations, commentaries and original writings by authors as diverse as Aristotle, Cicero, Rousseau, Marx, Wittgenstein, Foucault and Balibar, and totals in excess of 21 million tokens. The other corpora include an ancient Greek corpus (3.3 million tokens), a Latin corpus (1.5 million tokens), a medieval Arabic corpus (3.3 million tokens) and an Internet English corpus (5.6 million tokens). These all comprise similarly diverse collections of texts, written and/or translated at other moments in time over the past 2,500 years, under very different social, cultural, political and ideological conditions, and with the aim of fulfilling a diversity of philosophical, scientific and political purposes.

Of note here is the fact that the corpora, given the lengthy timespan covered, contain material originally drawn from a variety of media. In principle this variety is greater than in practice. The text of ancient manuscripts, for instance, was not collected for the corpus in its first documented form, but mostly through copies digitized from relatively recent prints, comparable in most respects to the monographs and edited volumes in our Modern

English corpus. The Internet English corpus texts, on the other hand, have been extracted from the highly dynamic, hyperlinked habitat where they, for the most part, first appeared. This difference in textual transmission history has implications for the number of transformations the material underwent before its inclusion in the corpus, and also influences the process of corpus compilation, as principles and practices of access may differ highly between online and offline publishing cultures. The internet provides a media environment increasingly dominated by a rejection of existing copyright laws through models such as creative commons and copyleft licensing, which attempt to assert the "fundamental human right to access our shared knowledge" (Nesson 2012: ix). Such evolutions bear witness not just to a cybercultural ideal of solidarity, but also to a logistic reality requiring a new property logic: a communicative environment constructed around hyperlinks and subject to a clipboard with instant copying capacity cannot incorporate effective controls against copyright infringement. The memetic internet economy is one of sharing and repurposing content. Access issues thus proved much less problematic when designing and building the Internet English corpus: while the research team did request permission from the site administrators of the 36 online media outlets currently represented, this was in most cases freely granted, in marked contrast with the response of the majority of copyright holders contacted during the construction of our print-based corpora.

The GoK resources are made available to the wider research community by means of a suite of open-source corpus analysis tools, which can be downloaded either from the project website (genealogiesofknowledge.net/software/) or via SourceForge (https://sourceforge.net/projects/modnlp/). This software package includes familiar interfaces such as a KWIC concordancer alongside a collection of more experimental data visualisation 'plugins', some of which have been designed specially with the aims and interests of the Genealogies of Knowledge project in mind. The features of these tools and their material implications for research in this field will be discussed in the following section.

4. Genealogies of Knowledge: Medial Environment

In this section, the medial environment constructed by the Genealogies of Knowledge software is illustrated with reference to two English versions of the *Communist Manifesto*. As Marx and Engels wanted the Manifesto's call to arms to be disseminated rapidly, and on a global scale, the *Communist Manifesto* is known for its 'obsession with its own translations', which are continually called for in successive prefaces (Puchner 2006: 3). The text has consequently been translated and retranslated many times, and in the Modern English corpus we have included both the first English translation (by Helen MacFarlane in 1850) and the most widely distributed one (by Samuel Moore in 1888), both of which are now freely available in the public domain. While MacFarlane's translation was first published in *The Red Republican*, it is the reprint produced in *Woodhull and Claflin's Weekly* in 1871 from which the text in the GoK corpus derives. Samuel Moore's 1888 version was approved by Friedrich Engels himself and remains the canonical version to this day. The copy of this text in the corpus derives from a collected volume of Marx's writings, published by Hackett (Simon 1994).

In Moore's version, the first sentence of the Manifesto reads: "A spectre is haunting Europe – the spectre of Communism" (1888/1994: 158). The phrase is immediately recognizable, and remains creatively productive – apart from the phrase being repeated as is, the word *Communism* has often been replaced in blog posts, newspaper features, academic articles and internet memes with a range of topical phenomena: from "the spectre of authoritarian capitalism" (Macfarlane 2020) to the "the spectre of the Unionised Jazz Musician" (Weidler 2013). Passed on through Derrida's *Specters of Marx* (1993), the phrase also helped lay the foundations for the interdisciplinary study of 'hauntology', a term applied in turn to numerous artistic efforts, particularly in the domain of music (Sexton 2012). Despite this lasting cultural prominence, the spectre never recurs in the Manifesto after the first page. Searching for the term *spectre* in both MacFarlane's and Moore's versions using the Genealogies of Knowledge concordance browser therefore returns just four lines (Figure 1).

Examining this concordance, here sorted by the R2 collocate, reveals that all cases of *spectre* derive from one version of the work, as can be understood

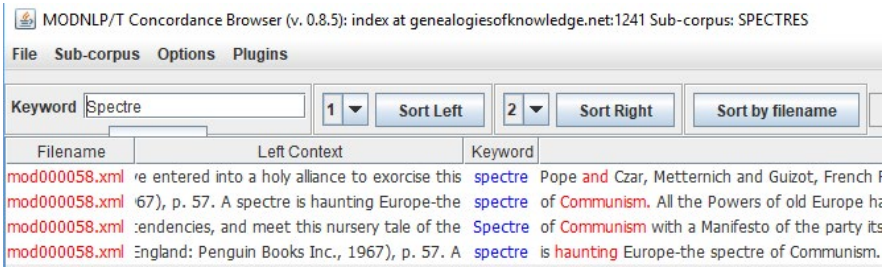

Figure 1: Concordance of **spectre** in MacFarlane's and Moore's translations of the *Communist Manifesto*, sorted by the R2 collocate.

from the identical file numbers on the left of Figure 1. Clicking on the interface's Metadata button reveals that these lines are all drawn from Moore's translation, and that the word *spectre* does not occur in MacFarlane's text. MacFarlane's version commences the treatise as follows: "A frightful hobgoblin stalks throughout Europe. We are haunted by a ghost, the ghost of Communism." A hobgoblin is a folkloristic, mischievous spirit. Moore's spectre simply haunts, but Macfarlane's goblin, or 'bugbear', as she later calls it, stalks. The impersonal Europe is replaced with a personal 'we' as the recipient of the ghost's attention, stressing the ambiguity of the spectral presence in the Manifesto: particularly in MacFarlane's text, it seems to haunt both the communists and their enemies. The possibility of this double reading stems from the fact that both parties intend to expel the ghost, but while their opponents seek to exorcise it, the communists seek to incarnate the spectre.

Such inferences are more difficult to establish in a concordance browser than by simply reading the texts themselves. The rendering of the German *Gespenst* by the English triumvirate of *hobgoblin*, *ghost*, and *bugbear*, or as a single *spectre*, immediately catches the eye in a linear reading, as do the activities pursued by these figures, but the nature of a concordance makes the sequence challenging to process. This is not just because a concordance breaks the narrative of the text, but also because the combination of a set co-textual span (in this case, 130 characters) and a keyword-centred view may produce multiple representations of single tokens, thus skewing the relation between text and data. In the concordance above, for instance, one can see that because the distance between separate occurrences of *spectre* is

shorter than the span, there is a duplication of spectres in the KWIC display. Only four are in the text, while one can spot six in the concordance. Through fragmentation and reordering, the browser conjures up supplementary spectres and amplifies the lexical patterns present in the text (Buts 2019: 93-108). Consequently, as discussed in section 2, the KWIC concordance is designed to spot repetition but disorients when repetition is ubiquitous throughout a corpus or locally concentrated in a specific section of a single text.

Other threads involving more dispersed and numerically significant patterns of repetition can be investigated in a manner more suited to the corpus software. For instance, the frequency list for both translations combined shows that the terms *bourgeois* and *bourgeoisie* are very common in this corpus. At 153 and 141 occurrences in a corpus of only 25,000 words, they take up the nineteenth and twenty-first positions in the list, and both items together occur more than frequent stop words such as *this* and *that*. However, the GoK Metafacet visualisation tool indicates that Moore (100 hits) uses *bourgeois* more than twice as much as MacFarlane (41 hits) (Figure 2).

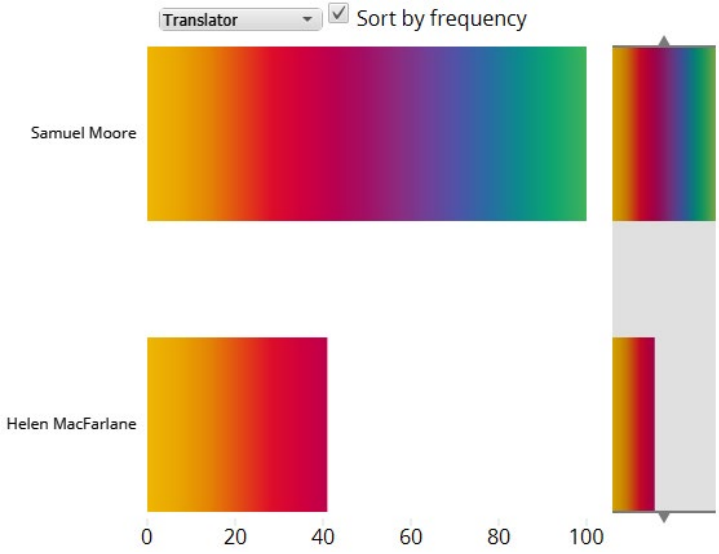

Figure 2: Metafacet visualisation comparing the frequency of **bourgeois** in MacFarlane's and Moore's translations

Metafacet allows for a comparison between the texts included in a corpus selection on the basis of any characteristic recorded in the metadata, such as year of publication, translator or outlet. Metafacet interacts directly with the concordance, and upon request removes lines associated with, for instance, one translator's output from view. It thus allows the user to alternate rapidly and dynamically between concordances generated from the two different translations. Although the Metafacet tool stands in direct connection with the concordance, it is a purely numerical, frequency-based application providing a count of concordance lines without relying upon or indicating what is presented within these lines. This distinguishes the tool from the Mosaic, a different visualisation tool meant to represent lexical items and their co-text occurring within the concordance. The Mosaic display is informed by principles of visualisation theory, as well as by the didactic work of Sinclair, who used similar diagrams in his example analyses (Luz & Sheehan 2014, 2020; Sinclair 2003). The user can request a Mosaic based on frequency, or on several variants of popular collocation measures such as mutual information and z-score. Restricting the search to MacFarlane's translation using the Metafacet tool, and then switching to the Mosaic visualisation's Column Frequency view reveals that *bourgeois* commonly precedes the words *regime*, *society*, *freedom*, *property*, and *socialism* in MacFarlane's translation (Figure 3).

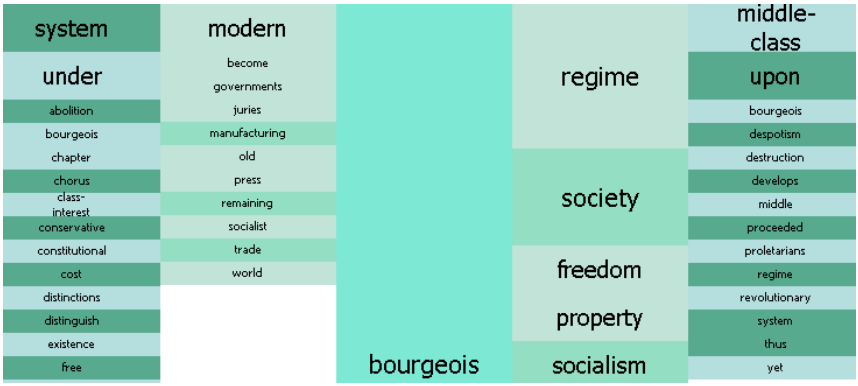

Figure 3: Column Frequency (No stopwords) view of the Mosaic tool displaying the most frequent L2-R2 collocates of **bourgeois** in MacFarlane's translation

Mosaic provides an alternative means of engaging with a keyword in context by grouping the lexical items that occur in each word position to the left and right of the search term and allocating them a differently shaded rectangle in columns placed immediately adjacent to this keyword. The bigger the word tile in the Column Frequency view, the more frequent the collocation. Figure 3 shows that many collocates listed in the R1 column are nouns. The concordance lines confirm that this pattern derives from MacFarlane using the term *bourgeois* almost exclusively as an adjective. By contrast, Moore, while generally conforming to this pattern, also uses the word nominally, as in 'the individual bourgeois'. One could perhaps further infer from Figure 3 that the term *bourgeois* is intimately associated with *socialism* and *freedom* for the writers of the Manifesto. However, references to 'bourgeois socialism' in the Manifesto are imbued with a strongly negative aura of meaning, given that in Marx's view this form of socialism seeks to redress social grievances only "in order to secure the continued existence of bourgeois society" (Marx & Engels 1994: 181).

The text as a whole presents an antagonistic dichotomy between proletariat and bourgeoisie, and vocabulary is part of the battlefield. Both parties may use terms such as *socialism*, but supposedly one class produces false representations, while the other holds onto true aspirations. Closer scrutiny of the co-text is necessary here to adequately interpret the visual representations, meaning that, at the very least, use of the Mosaic tool must be integrated within a workflow that additionally exploits the affordances of the concordance display. Nevertheless, the Mosaic counts, orders and highlights with remarkable efficiency, and its ability to shift between several measures of collocational importance can strongly impact the user's interaction with the data. It is important to note here that some functionalities of the GoK software package such as Frequency List are based upon the whole corpus under investigation, while tools such as Metafacet and Mosaic only take into account data present in the concordances retrieved for a specific keyword. Furthermore, whereas Metafacet gives information about the contextual provenance of the lines returned in a search, the Mosaic presents an alternative visualisation of their textual constitution. In short, statistics and their visualisations can function independently, but often depend on the generation of KWIC concordances, or at least on the selection of a keyword.

It should further be noted that the *keyword* in *keyword in context* is itself a flexible designation. In the GoK browser environment, as in many corpus tools, a variety of orthographic sequences corresponding to a search pattern can be retrieved by means of a sequencing grammar, as well as more complex regular expressions. When corpora are queried through such metalanguages, one makes use of a means of interpretation that is heavily dependent upon the affordances of the medium. In what way the metalanguages facilitating digital research correspond to the data being processed and queried, and ultimately to the text studied, is a matter requiring continued attention. Similarly, in what way the output of visualisation tools designed for corpus analysis, such as KWIC, Metafacet, and Mosaic discussed above, can be considered mutually equivalent representations, and to what extent they can function independently, is a question CTS must dare to ask.

5. Conclusion

Translation studies is accustomed to the importance of presentation, style, and rhetoric, and does not question the idea that different translatorial strategies lead to different translation solutions, and thus to different interpretations. In this respect, corpus-based studies of translation have facilitated the meticulous investigation of small shifts that aggregate into paradigm changes. Nevertheless, despite this sensitivity to matters of transmission and transformation, CTS has not brought into focus its own medial qualities. The transformation of text into data allows for many competing approaches and representations, none of which should individually be accepted uncritically as equal and accurate renderings of the object of study. The representation of text, for instance, is not a mandatory constraint for the interpretation of text. Yet at present, given the human familiarity with this form of expression and the influence KWIC design has exerted on the development of alternatives, the triangulation of multiple software tools, often integrated with the KWIC concordance, seems to be a sensible approach for CTS to pursue. Yet, as tools become more abundant and easier to work with, one risks losing sight of the particular choices that govern their inception and implementation. Efficient research tools tend to draw attention away from themselves, and from the choices they impose on the form that an object of analysis takes.

One may be reminded here that Sinclair, speaking about the lexicon, argued that "there is no distinction between form and meaning" (Sinclair 1991: 7). The statement illustrates a functional view of language, one in which use is the determining factor: some words 'fit' in certain situations. Corpus research, a mode of enquiry that finds its origins in the declaration of equivalence between meaning and form, cannot leave the research environment out of its purview. From a similarly reflective perspective, translation studies, like translation, cannot merely repeat, copy, or reproduce information. When interpretation takes place in a customized medial environment, the affordances of this environment should be interrogated. This article has attempted to promote wider awareness of this issue in CTS, and to illustrate the influence of the medium on corpus analysis with reference to the representation of the *Communist Manifesto* in the Genealogies of Knowledge Modern English corpus. We have drawn particular attention to the interdependence between textual analysis, theoretical development, and corpus construction, for instance with reference to the convention that a corpus should contain ample textual variation, rather than consist of repeated utterances. The short case study analysis illustrated that one of the causes for the avoidance of duplication in corpus construction may be that the concordance view is well suited to call attention to dispersed repetition, but less so to represent concentrated repetition. Such examples go beyond the established critique that corpus analysis tends to disregard the integrity of the text as a communicative unit. The fruitful alliance between corpus and discourse studies has paid ample attention to the balance sought between closer and more distant forms of reading, and has indeed recently turned a critical eye to methodological choices and their relation to the tools used for research (Taylor & Marchi 2018). This article has argued that CTS should be at the forefront of this ongoing critical engagement, as a translational perspective may aid in explicating the various transformations that turn text into data, and that make data suitable for interpretation.

Examining this dusty corner of the discipline is all the more important as the gap between analysis and presentation widens. When concordance evidence was the incontestable 'stock-in-trade' of the corpus analyst, research articles could reproduce a large part of the investigative flow, thus ensuring transparency and replicability. Today, the use of very variable corpora,

statistical operations and visualisation tools is rapidly multiplying, and it is often a struggle to represent methodological pathways in the classic format of research papers. In effect, this procedure often requires back-translation: textual information processed to facilitate corpus research is once again adapted to representation on the page, be it printed or digital. Finally, then, we suggest that further research into the question of how different media shape our interaction with textual data must also begin to consider the demands of publishing cultures and the extent to which existing models for publication may need to evolve in order to meet the requirements of the expanding digital humanities.

References

ANTHONY, Laurence. (2018) "Visualisation in Corpus-based Discourse Studies". In: Taylor, Charlotte & Anna Marchi (eds.) *Corpus Approaches to Discourse: A critical review*, London: Routledge, pp. 1-15.

ARMSTRONG, Guyda. (2020) "Media and Mediality." In: Baker, Mona & Gabriella Saldanha (eds.) 2020. *Routledge Encyclopedia of Translation Studies*. London: Routledge, pp. 310-315.

BAKER, Mona. (1993) "Corpus Linguistics and Translation Studies: Implications and applications." In: Baker, Mona, Gill Francis & Elena Tognini-Bonelli (eds.) 1993. *Text and Technology: In honour of John Sinclair*. Philadelphia & Amsterdam: John Benjamins, pp. 233-250.

BAKER, Mona. (2020) "Rehumanizing the Migrant: The translated past as a resource for refashioning the contemporary discourse of the (radical) left." *Palgrave Communications* 6:1.

BAKER, Mona & Henry Jones. (*forthcoming*) "Genealogies of Knowledge: Theoretical and methodological issues." *Palgrave Communications*.

BAKER, Paul. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.

BAKER, Paul, Costas Gabrielatos, Majid KhosraviNik, Michal Krzyzanowski, Tony McEnery & Ruth Wodak. (2008) "A Useful Methodological Synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK Press." *Discourse & Society* 19:3, pp. 273-306.

Baldry, Anthony & Paul J. Thibault. (2008) "Applications of Multimodal Concordances." *Hermes – Journal of Language and Communication Studies* 41, pp. 11-41.

Berry, David. (2011) "The Computational Turn: Thinking about the digital humanities." *Culture Machine* 12, pp. 1-22.

Buts, Jan. (2019) *Political Concepts and Prefiguration: A corpus-assisted enquiry into democracy, politics and community*. Manchester: University of Manchester. Unpublished PhD Thesis.

Cantos, Pascual & Aquilino Sánchez. (2001) "Lexical Constellations: What collocates fail to tell." *International Journal of Corpus Linguistics* 6:2, pp. 199-228.

Cronin, Michael. (2013) *Translation in the Digital Age*. London & New York: Routledge.

Derrida, Jacques. (1994) *Specters of Marx: The state of the debt, the work of mourning and the new international*. Peggy Kamuf (trans.). New York & London: Routledge.

Dixon, Dan. (2012) "Analysis Tool or Research Methodology: Is there an epistemology for patterns?" In: Berry, M. David (ed.) 2012. *Understanding Digital Humanities*. Basingstoke: Palgrave Macmillan, pp. 191-209.

Fenlon, John F. (1908) "Concordances of the Bible." In: Herbermann, Charles G., Edward A. Pace, Condé B. Pallen, Thomas J. Shahan & John J. Wynne (eds.) 1908. *The Catholic Encyclopedia, Volume 4*. New York: Robert Appleton Company, pp. 195-196.

Firth, John R. (1968) "Linguistic Analysis as a Study of Meaning." In: Palmer, Frank R. (ed.) 1968. *Selected Papers of J. R. Firth*. London & Harlow: Longman, 12-26.

Gabrielatos, Costas, Tony McEnery, Peter J. Diggle & Paul Baker. (2012) "The Peaks and Troughs of Corpus-based Contextual Analysis." *International Journal of Corpus Linguistics* 17:2, pp. 151-175.

Hermans, Theo. (2007) *The Conference of the Tongues*. Manchester: St. Jerome Publishing.

Hoey, Michael. (2007) "Lexical Priming and Literary Creativity." In: Hoey, Michael, Michaela Mahlberg, Michael Stubbs & Wolfgang Teubert (eds.) 2007. *Text, Discourse and Corpora: Theory and analysis*. London & New York: Continuum, pp. 31-56.

HOEY, Michael. (2011) "Lexical Priming and Translation." In: Kruger, Alet, Kim Wallmach & Jeremy Munday (eds.) 2011. *Corpus-Based Translation Studies: Research and Applications*. London & New York: Bloomsbury, pp. 153-168.

HUNSTON, Susan. (2007) "Semantic Prosody Revisited." *International Journal of Corpus Linguistics* 12:2, pp. 249-286.

JIMÉNEZ Hurtado, Catalina & Silvia Soler Gallego. (2013) "Multimodality, Translation and Accessibility: A corpus-based study of audio description." *Perspectives* 21:4, pp. 577-594.

JONES, Henry. (2018) "Mediality and Audiovisual Translation." In: Pérez-González, Luis (ed.) 2018. *Routledge Handbook of Audiovisual Translation*. London & New York: Routledge, pp. 177-191.

JONES, Henry. (2019) "Searching for Statesmanship: A corpus-based analysis of a translated political discourse." *Polis: The Journal for Ancient Greek and Roman Political Thought* 36, pp. 216-241.

JONES, Henry. (2020) "Retranslating Thucydides as a Scientific Historian." *Target* 32:1, pp. 59-82.

KARIMULLAH, Kamran. (2020) "Editions, Translations, Transformations: Refashioning the Arabic Aristotle in Egypt and metropolitan Europe, 1940–1980." *Palgrave Communications* 6:3.

KENNY, Dorothy. (2011) "Translation Units and Corpora." In: Kruger, Alet, Kim Wallmach and Jeremy Munday (eds.) 2011. *Corpus-Based Translation Studies: Research and Applications*. London & New York: Bloomsbury, pp. 76-102.

KLEIN, Hans K. & Daniel Lee Kleinman. (2002) "The Social Construction of Technology: Structural considerations." *Science, Technology & Human Values* 27:1, pp. 28-52.

KRESS, Gunther. (2003) *Literacy in the New Media Age*. London & New York: Routledge.

LITTAU, Karin. (2006) *Theories of Reading: Books, bodies, and bibliomania*. Cambridge: Polity Press.

LITTAU, Karin. (2011) "First Steps Towards a Media History of Translation." *Translation Studies* 4:3, pp. 261-281.

LITTAU, Karin. (2016) "Translation and the Materialities of Communication." *Translation Studies* 9:1, pp. 82-96.

LOUW, Bill. (1993) "Irony in the Text or Insincerity in the Writer: The diagnostic potential of semantic prosody." In: Baker, Mona, Gill Francis & Elena

Tognini-Bonelli (eds.) 1993. *Text and Technology: In honour of John Sinclair*. Philadelphia & Amsterdam: John Benjamins, pp. 157-176.

LUHN, Hans Peter. (1960) "Key Word-in-Context Index for Technical Literature (Kwic Index)." *American Documentation* 11:4, pp. 288-295.

LUZ, Saturnino. (2011) "Web-Based Corpus Software." In: Kruger, Alet, Kim Wallmach & Jeremy Munday (eds.) 2011. *Corpus-Based Translation Studies: Research and Applications*. London & New York: Bloomsbury, pp. 124-149.

LUZ, Saturnino & Shane Sheehan. (2014) "A Graph Based Abstraction of Textual Concordances and Two Renderings for their Interactive Visualisation." In: 2014. *Proceedings of the International Working Conference on Advanced Visual Interfaces*. New York: ACM, pp. 293-296.

LUZ, Saturnino & Shane Sheehan. (2020) "Methods and Visualization Tools for the Analysis of Medical, Political and Scientific Concepts in Genealogies of Knowledge." *Palgrave Communications* 6:49, pp. 1-20.

MACFARLANE, Laurie. (2020) "A Spectre is Haunting the West – The spectre of authoritarian capitalism." *Open Democracy*. <https://www.opendemocracy.net/en/oureconomy/a-spectre-is-haunting-the-west-the-spectre-of-authoritarian-capitalism/>

MALAMATIDOU, Sofia. (2018) *Corpus Triangulation: Combining data and methods in corpus-based translation studies*. London & New York: Routledge.

MANOVICH, Lev. (2001) *The Language of New Media*. Cambridge, MA & London: MIT Press.

MARX, Karl and Friedrich Engels. (1850) "German Communism – Manifesto of The German Communist Party." Translation by Helen MacFarlane. *The Red Republican* 21:1.

MARX, Karl and Friedrich Engels. (1850/1871) "German Communism - Manifesto of The German Communist Party." Translation by Helen MacFarlane. *Woodhull and Claflin's Weekly* 4:7. Online version: <http://iapsop.com/archive/materials/woodhull_and_claflins_weekly/>

MARX, Karl & Friedrich Engels. (1888/1994) "The Communist Manifesto." Translation by Samuel Moore. In: Simon, Lawrence H. (ed.) 1994. *Karl Marx: Selected writings*. Indianapolis: Hackett Publishing Company, pp. 157-186.

MCINTYRE, Dan & Brian Walker. (2019) *Corpus Stylistics: Theory and practice*. Edinburgh: Edinburgh University Press.

MCLUHAN, Marshall. (1964) *Understanding Media: The extensions of man*. New York: McGraw-Hill.

MUNDAY, Jeremy. (2011) "Looming Large: A cross-linguistic analysis of semantic prosodies in comparable reference corpora." In: Kruger, Alet, Kim Wallmach & Jeremy Munday (eds.) 2011. *Corpus-Based Translation Studies: Research and Applications*. London & New York: Bloomsbury, 169-186.

NESSON, Charles R. (2012) "Foreword." In: Dulong de Rosnay, Melanie & Juan Carlos de Martin (eds.) 2012. *The Digital Public Domain: Foundations for an Open Culture*. Cambridge: Open Book Publishers, pp. xi-xiii..

OAKES, Michael P. & Ji Meng. (eds.) (2012) *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*. Amsterdam & Philadelphia: John Benjamins.

PARTINGTON, Alan. (1998) *Patterns and Meanings: Using corpora for English language research and teaching*. Amsterdam & Philadelphia: John Benjamins.

PÉREZ-GONZÁLEZ, Luis. (2014) "Multimodality in Translation and Interpreting Studies: Theoretical and methodological perspectives." In: Bermann, Sandra & Catherine Porter (eds.) 2014. *A Companion to Translation Studies*. Chichester: Wiley Blackwell.

PUCHNER, Martin. (2006) *Poetry of the Revolution: Marx, manifestos, and the avant-gardes*. Princeton: Princeton University Press.

SEXTON, Jamie. (2012) "Weird Britain in Exile: Ghost Box, hauntology, and alternative heritage." *Popular Music and Society* 35:4, pp. 561-584.

SHERIDAN, Mary P. (2016) "Recent Trends in Digital Humanities Scholarship." In: Dejica, Daniel, Gyde Hansen, Peter Sandrini & Iulia Para (eds.) 2016. *Language in the Digital Era: Challenges and perspectives*. Warsaw & Berlin: De Gruyter Open, pp. 2-13.

SIMON, Lawrence H. (ed.) (1994) *Karl Marx: Selected writings*. Indianapolis & Cambridge: Hackett Publishing Company.

SINCLAIR, John. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

SINCLAIR, John. (2003) *Reading Concordances: An introduction*. London: Pearson Longman.

SINCLAIR, John. (2004) *Trust the Text: Language, corpus and discourse*. Ronald Carter (ed.) London & New York: Routledge.

SINCLAIR, John. (2008) "The Phrase, the Whole Phrase, and Nothing But the Phrase." In: Granger, Sylviane & Fanny Meunier (eds.) 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam & Philadelphia: John Benjamins, pp. 407-410.

STEWART, Dominic. (2009) "Safeguarding the Lexicogrammatical Environment: Translating semantic prosody." In: Beeby, Allison, Patricia Rodríguez Inés & Pilar Sánchez-Gijón (eds.) 2009. *Corpus Use and Translating*. Amsterdam & Philadelphia: John Benjamins.

STEWART, Dominic. (2010) *Semantic Prosody: A critical evaluation*. New York & London: Routledge.

STUBBS, Michael. (1996) *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Cambridge: Blackwell.

STUBBS, Michael. (2009) "The Search for Units of Meaning: Sinclair on empirical semantics." *Applied Linguistics* 30:1, pp. 115-137.

SVARTVIK, Jan. (1992) "Corpus Linguistics comes of Age." In: Svartvik, Jan (ed.) 1992. *Directions in Corpus Linguistics*: *Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*. Berlin & New York: Mouton de Gruyter, pp. 7-13.

TAYLOR, Charlotte and Anna Marchi (eds.) 2018. *Corpus Approaches to Discourse: A critical review*. London: Routledge.

TOGNINI-BONELLI, Elena. (2001) *Corpus Linguistics at Work*. Amsterdam & Philadelphia: John Benjamins.

TYMOCZKO, Maria. (2010) "Translation, Resistance, Activism: An overview." In: Maria Tymoczko (ed.) 2010. *Translation, Resistance, Activism*. Amherst: University of Massachusetts Press, 1-22.

VENUTI, Lawrence. (1998) *The Scandals of Translation*: *Towards an Ethics of Difference*. London & New York: Routledge.

VINAY, Jean-Paul & Jean Darbelnet. (1995) *Comparative Stylistics of French and English: A methodology for translation*. Amsterdam: John Benjamins.

WHITSITT, Sam. (2005) "A Critique of the Concept of Semantic Prosody." *International Journal of Corpus Linguistics* 10:3, pp. 283-205.

WEIDLER, Arndt. (2013) "A Spectre Is Haunting Europe – The spectre of the unionised jazz musician!" Translation by Paul McCarthy. *Goethe Institute Blog*. <https://www.goethe.de/en/kul/mus/gen/jaz/jah/20454965.html>

ZANETTIN, Federico. (2011) "Hardwiring Corpus-Based Translation Studies: Corpus encoding." In: Kruger, Alet, Kim Wallmach & Jeremy Munday (eds.) 2011. *Corpus-Based Translation Studies: Research and Applications*. London & New York: Bloomsbury, pp. 103-123.

BIONOTE

JAN BUTS is a postdoctoral researcher attached to the QuantiQual project (https://adaptcentre.ie/projects/quantiqual/) at Trinity College Dublin, and a co-coordinator of the Genealogies of Knowledge Research Network (https://genealogiesofknowledge.net/research-network/). He works at the intersection of translation theory, conceptual history, corpus linguistics, and online media.

HENRY JONES is a lecturer in translation and intercultural studies at Aston University, UK. He is a co-coordinator of the Genealogies of Knowledge Research Network (https://genealogiesofknowledge.net/research-network/) and co-editor of the *Routledge Encyclopedia of Citizen Media* (2021). His current research interests include corpus-based translation studies, translation history, media theory and online translating communities.

NOTICES BIOGRAPHIQUES

JAN BUTS est chercheur postdoctoral attaché au projet QuantiQual (https://adaptcentre.ie/projects/quantiqual/) au Trinity College de Dublin, et un des coordinateurs du Genealogies of Knowledge Research Network (https://genealogiesofknowledge.net/research-network/). Il travaille à l'intersection de la théorie de la traduction, de l'histoire conceptuelle, de la linguistique des corpus et des médias en ligne.

HENRY JONES est maître de conférences à Aston University, Royaume-Uni. Il est un des coordinateurs du Genealogies of Knowledge Research Network (https://genealogiesofknowledge.net/research-network/) et un des éditeurs du *Routledge Encyclopedia of Citizen Media* (2021). Ses intérêts de recherche comprennent la traductologie de corpus, l'histoire de la traduction, la théorie des médias et les communautés virtuelles.

AIMS / OBJETIVOS / OBJECTIUS

MonTI (*Monographs in Translation and Interpreting*) is an academic, peer-reviewed and international journal fostered by the three public universities with a Translation Degree in the Spanish region of Valencia (Universitat d'Alacant, Universitat Jaume I de Castelló and Universitat de València).

Each issue will be thematic, providing an in-depth analysis of translation- and interpreting-related matters that meets high standards of scientific rigour, fosters debate and promotes plurality. Therefore, this journal is addressed to researchers, lecturers and specialists in Translation Studies.

MonTI will publish one issue each year, first as a hard copy journal and later as an online journal.

In order to ensure both linguistic democracy and dissemination of the journal to the broadest readership possible, the hard-copy version will publish articles in German, Spanish, French, Catalan, Italian and English. The online version is able to accommodate multilingual versions of articles, and it will include translations into any other language the authors may propose and an attempt will be made to provide an English-language translation of all articles not submitted in this language.

Further information at:

http://dti.ua.es/es/monti-english/monti-contact.html

MonTI es una revista académica con vocación internacional promovida por las universidades públicas valencianas con docencia en traducción e interpretación (Universidad de Alicante, Universidad Jaume I de Castellón y Universidad de Valencia).

Nuestra revista pretende ante todo centrarse en el análisis en profundidad de los asuntos relacionados con nuestra interdisciplina a través de monográficos caracterizados por el rigor científico, el debate y la pluralidad. Por consiguiente, la revista está dirigida a investigadores, docentes y especialistas en estudios de traducción.

MonTI publicará un número monográfico anual, primero en papel y a continuación en edición electrónica. Igualmente y con el fin de alcanzar un equilibrio entre la máxima pluralidad lingüística y su óptima difusión, la versión en papel admitirá artículos en alemán, castellano, catalán, francés, italiano o inglés, mientras que la edición en Internet aceptará traducciones a cualquier otro idioma adicional y tratará de ofrecer una versión en inglés de todos los artículos.

Más información en:

http://dti.ua.es/es/monti/monti.html

MonTI és una revista acadèmica amb vocació internacional promoguda per les universitats públiques valencianes amb docència en traducció i interpretació (Universitat d'Alacant, Universitat Jaume I de Castelló i Universitat de València).

La nostra revista pretén sobretot centrar-se en l'anàlisi en profunditat dels assumptes relacionats amb la nostra interdisciplina a través de monogràfics caracteritzats pel rigor científic, el debat i la pluralitat. Per tant, la revista va dirigida a investigadors, docents i especialistes en estudis de traducció.

MonTI publicarà un número monogràfic anual, primer en paper i a continuació en edició electrònica. Igualment, i a fi d'aconseguir un equilibri entre la màxima pluralitat lingüística i la seua difusió òptima, la versió en paper admetrà articles en alemany, castellà, català, francès, italià o anglès, mentre que l'edició en Internet acceptarà traduccions a qualsevol altre idioma addicional i tractarà d'oferir una versió en anglès de tots els articles.

Més informació a:

http://dti.ua.es/es/monti-catalan/monti-contacte.html

MAIN INSTRUCTIONS FOR AUTHORS

MonTI publishes one yearly issue. The contributions, which should be original and unpublished, will be strictly subjected to the following norms:

1. Maximum length: 10,000 words, including works cited.

2. Font and spacing: Font: Times New Roman; Size: 11 pt.; Line spacing: single.

3. Language options: Catalan, English, French, German, Italian or Spanish.

4. The title page should include the following information in this same order:

Title of the article, followed by a blank line. An English translation of the title should be included if this is not in English. Author(s). <e-mail>. Affiliation, followed by a blank line. Abstracts in English and in any of the other four languages. (Maximum length: 150 words each). Keywords: up to five subject headings in each of the same two languages.

5. Parenthetical citations: *MonTI* follows one of the main variants of the Chicago-style citation: Surname(s) (year: pages) or (Surname(s) year: pages).

6. Works cited / references: This section will only include works really cited in the text and will begin after the article has come to an end. The list will be arranged in alphabetical order by author and year of the first edition, and according to the following pattern: Monographs: Author (Surname(s), complete first name). Year (in brackets) *Title* (in italics). City: Publisher. Journal article: Author (Surname(s), complete first name). Year (in brackets). "Title of the article" (with quotation marks). *Name of the journal* (in italics). Volume: Issue, first page-last page (preceded by the abbreviation pp.)

At *MonTI*'s website (http://dti.ua.es/es/monti-english/monti-authors.html) numerous examples of each of these variants are available.

7. Deadline: The deadline will be May 31. The contribution and a short CV (a maximum of 150 words) for each of the authors in a separate file will be sent as an attachment (in Word or any other word processor compatible with Word) addressed to the Managing Editor of *MonTI*: <email: monti.secretaria@ua.es>

8. After requesting the editors' approval and receiving the reports from the referees, the journal will provide the authors with a reasoned statement regarding the acceptance of their contributions.

PRINCIPALES NORMAS DE REDACCIÓN

MonTI edita un número anual. Los trabajos originales e inéditos que se propongan para su publicación en la revista se someterán estrictamente a las siguientes normas:

1. Extensión máxima: 10.000 palabras, incluida la bibliografía.

2. Tipo de letra: Times New Roman. Tamaño de letra: 11 pt. Interlineado: sencillo.

3. Lenguas vehiculares: alemán, castellano, catalán, francés, inglés o italiano.

4. La primera página incluirá, por este orden y en líneas sucesivas, lo siguiente:

Título del trabajo, seguido de una línea en blanco de separación. Deberá aportarse, además, la traducción del título al inglés, si el artículo no está escrito en esta lengua. Autor(es). <Correo electrónico>. Centro de procedencia, seguido de una línea en blanco de separación. Resúmenes en inglés y en otra de las lenguas vehiculares (extensión máxima de 150 palabras cada uno). Palabras clave: se aportarán cinco términos en los dos idiomas de los resúmenes.

5. Remisión a la Bibliografía. Se seguirá una de las principales variantes del estilo Chicago de citas: Apellido(s) del autor (año: páginas) o (Apellido(s) del autor año: páginas).

6. Bibliografía: este epígrafe sólo recogerá los trabajos citados en el artículo, y aparecerá después del final del texto. Se ordenará alfabéticamente por autor y año del siguiente modo. Monografías: Autor (apellido(s), nombre completo). Año (entre paréntesis) *Título* (en cursiva). Ciudad: Editorial. Artículo de revista: Autor (apellido(s), nombre completo). Año (entre paréntesis). Título del artículo (entre comillas). *Nombre de la revista* (en cursiva). Volumen: fascículo, páginas de comienzo y fin del artículo (antecedidas por la abreviatura pp.).

En la página web de *MonTI* (http://dti.ua.es/es/monti/normas-de-redaccion.html) se puede acceder a numerosos ejemplos de cada una de las variantes de referencia bibliográfica.

7. Envío de originales: el plazo de recepción finalizará el 30 de junio. Los textos –y un breve currículo (150 palabras máximo) de los autores en otro documento– se remitirán en soporte informático (Word o cualquier programa de tratamiento de textos compatible con Word) dirigidos al Secretario de la revista: <e-mail: monti.secretaria@ua.es>

8. La Dirección de la revista, vistos los informes de los asesores y el parecer de los editores, comunicará a los autores la decisión razonada sobre la aceptación o no de los trabajos.

MONTI