BENJAMINS TRANSLATION LIBRARY

Corpora in Translation and Contrastive Research in the Digital Age

edited by

Julia Lavid-López
Carmen Maíz-Arévalo
Juan Rafael Zamorano-Mansilla

Corpora in Translation and Contrastive Research in the Digital Age

Benjamins Translation Library (BTL)

ISSN 0929-7316

The Benjamins Translation Library (BTL) aims to stimulate research and training in Translation & Interpreting Studies – taken very broadly to encompass the many different forms and manifestations of translational phenomena, among them cultural translation, localization, adaptation, literary translation, specialized translation, audiovisual translation, audio-description, transcreation, transediting, conference interpreting, and interpreting in community settings in the spoken and signed modalities.

For an overview of all books published in this series, please see
benjamins.com/catalog/btl

Volume 158

Corpora in Translation and Contrastive Research in the Digital Age
Recent advances and explorations
Edited by Julia Lavid-López, Carmen Maíz-Arévalo
and Juan Rafael Zamorano-Mansilla

Corpora in Translation and Contrastive Research in the Digital Age

Recent advances and explorations

Edited by

Julia Lavid-López
Carmen Maíz-Arévalo
Juan Rafael Zamorano-Mansilla
Universidad Complutense de Madrid

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/btl.158

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2021020974 (PRINT) / 2021020975 (E-BOOK)

ISBN 978 90 272 0918 4 (HB)
ISBN 978 90 272 5968 4 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

Table of contents

Corpus resources and tools
Looking back and going ahead

Julia Lavid-López
Universidad Complutense de Madrid

In November 2018, the research group FUNCAP in collaboration with the Instituto de Lenguas Modernas y Traductores and members of the Department of English Studies at Universidad Complutense of Madrid (UCM) hosted the *International Symposium PaCor 2018 (Parallel Corpora: Creation and Applications)*, as the second international event in the biennial series inaugurated in 2016 at the University of Santiago de Compostela. The event was held in conjunction with the *XVI Encuentros Complutenses en torno a la traducción*, a traditional forum for the dissemination of translation research, organised by the Instituto de Lenguas Modernas y Traductores at the Faculty of Philology (Universidad Complutense de Madrid).

The main goal of the PaCor2018 symposium was to encourage dialogue and contact among researchers and practitioners working on the creation, annotation and processing of parallel and comparable corpora and those exploiting such resources for a wide range of theoretical and applied purposes. In addition, we wanted to bring together the more technology-oriented translation and interpreting researchers and practitioners with others using more traditional methodologies, with the aim of fostering interaction among them and opening up new avenues of collaboration through the use of corpora.

Drawing on these two successful events, the current book contains a selection of papers which focus on corpora and translation research in the digital age, outlining some recent advances and explorations.[1] The chapters, which encompass a variety of research aims and methodologies, can be grouped into two main themes: one devoted to current advances in the creation of corpus resources and tools from research-oriented to applied perspectives; the other showcasing a number of corpus-based studies which illustrate the impact of corpus resources and tools in contrastive and translation research.

These two themes are described in detail in the following sections, looking back at some of the most important achievements and exploring ways ahead.

[1] All papers included in this volume have undergone a rigorous double blind peer reviewing process, each being assessed by three reviewers.

https://doi.org/10.1075/btl.158.int

1. Corpus resources

More than two decades ago, Mona Baker announced what she considered to be a turning point in the history of Translation Studies, coming “as a direct consequence of access to large corpora of both original and translated texts, and of the development of specific methods and tools for interrogating such corpora in ways which are appropriate to the needs of translation scholars” (Baker 1993: 235). Similarly, but from a linguistic perspective, John Sinclair anticipated that the new corpus resources would have a profound effect on the translations of the future. (Sinclair 1992: 395)

A few years later, Maria Tymoczko underlined the centrality of the corpus approach within the discipline of Translation Studies. In her view, its strength lay in the flexibility and adaptability of corpora, as well as in the openendedness of their construction, which allows researchers “to move from text-based questions to context-based questions, thus marking “a turn away from prescriptive approaches to translation toward descriptive approaches” (Tymoczko 1998: 652).[2]

This new vision of Translation Studies was facilitated by the increasing availability of electronic corpora from the 1980s onwards and the emergence of Corpus Linguistics as a discipline in subsequent years. However, most corpora at that time remained monolingual and, for the first time, used for the compilation of English grammars (see for example Quirk, Greenbaum, Leech and Svartvik’s *A Comprehensive Grammar of the English Language*, 1985) and dictionaries (for example, Sinclair’s 1987 *Collins Cobuild English Language Dictionary*). It was not until the late 1980s and the 1990s that scholars, particularly in the Scandinavian countries, showed interest in the potential role of corpora for Contrastive Linguistics and Translation Studies. Collaborative efforts among researchers from the Nordic countries led to the compilation of the English-Norwegian Parallel Corpus (ENPC), the Swedish-English Parallel Corpus (ESPC), and the Finnish-English Contrastive Corpus Studies Project (FECCS). They were all active in the field of Contrastive Linguistics and shared an interest in “linking cross-linguistic studies with corpus linguistics and the use of the computer as a tool in linguistic research” (Aijmer et al. (eds) 1996, Acknowledgments).

The ENPC was created by Stig Johansson and his team at the University of Oslo using a model which would be used later for the creation of other parallel corpora (Johansson & Hofland 1994). The model, which can be characterized as a *bidirectional translation corpus*, consisted of a combination of multilingual corpora of original texts and their translations (for Contrastive and Translation Studies), multilingual corpora of original texts which are matched by criteria such as genre, time of composition, etc. (for Contrastive Studies), and monolingual corpora consisting

[2] The different modes of interrogating corpora can serve to address not only linguistic questions but also questions of culture, ideology and literary criticism.

of original and translated texts (for translation studies). By combining all three types under the same framework it would be possible to use the same corpus both for Contrastive and Translation Studies (Johansson 1998). Following the ENPC model, a number of corpora were assembled in the following years for different language pairs, such as the PLECI corpus of English and French, the Chemnitz Corpus of English and German, and the ACTRES Corpus, among others. More recent developments include the CroCo Corpus (Hansen-Shirra et al. 2013), the MULTINOT (Lavid et al 2015), and the P-ACTRES 2.0 corpus (Sanjurjo-González & Izquierdo 2019) for English and Spanish, among others.[3]

The increasing availability of multilingual corpora in the following decades boosted both Contrastive and Translation Studies, leading to a convergence between both scientific camps, in spite of their different objectives. Researchers in Contrastive Linguistics and Translation Studies could now rely on corpora to "verify, refine or clarify theories that hitherto had had little or no empirical support and to achieve a higher degree of descriptive adequacy". (Granger 2003: 19).

At the same time, researchers in the field of Natural Language Processing (NLP) had begun to use large collections of texts for various NLP applications since the early 1990s, and bilingual and multilingual corpora were essential resources for the needs of Machine Translation (MT), the original goal of Computational Linguistics (Kay 1997).

At this point, it is important to clarify some terminological issues regarding the types of corpora available for linguistic and translation research. If we take the number of languages involved as the criterion for definition, a general distinction can be made between monolingual and multilingual (including bilingual) corpora. Within multilingual (and bilingual) corpora a further distinction can be made between comparable corpora and parallel corpora. While comparable corpora are those compiled using similar design criteria –the same sampling frame and similar balance and representativeness-, parallel corpora are generally understood to be those consisting of source texts aligned with their target texts. However, the terminology is somewhat unstable and the distinction between the two types of corpora is not always clear cut (see Fantinuoli & Zanettin 2015: 3).[4] According to

[3] More information about these corpora can be found at their respective websites:
PLECI at https://uclouvain.be/en/research-institutes/ilc/cecl/pleci.html; CROCO at http://fedora.clarin-d.uni-saarland.de/croco-gecco/croco/index_en.html; MULTINOT at https://www.ucm.es/funcap/multinot; and P-ACTRES 2.0 at https://actres.wesped.es/es/corpus-paralelo/

[4] These authors rightly point out that parallel corpora do not necessarily contain translations, as it is the case with the large multilingual parallel corpora Europarl and Acquis Communautaire, which contain original texts as well as translations. Comparable corpora may contain not only original texts but also translations; and there exist various "hybrid texts" such as news translations and text crowdsourcing which are partly original texts and partly translations.

these authors, it is more useful to use the terms "parallel" or "comparable" to refer to the type of corpus architecture, rather than to the status of the texts as originals or translations. Parallel corpora would be those whose components are aligned, while comparable corpora would be those "which are compared on the whole on the basis of assumed similarity" (Fantinuoli & Zanettin 2015: 4).

When parallel corpora contain translations they are usually classified as bilingual or multilingual, unidirectional (e.g. from English into Spanish or from Spanish into English alone), bidirectional (e.g. containing both English source text with their Spanish translations and Spanish source texts with their English translations) or multidirectional (e.g. the same text with English, French, German and Spanish versions). Their development has experienced an exponential growth over recent years, with workshops, conferences and publications dedicated to the creation, annotation and exploitation of parallel corpora. The most important multilingual parallel resources available for the research community are the ones released by the European Commission's Joint Research Centre (JRC) and other European Union organisations (Steinberger et al. 2014), covering between 22 and 26 languages.[5]

Further developments include the Europarl Corpus (Koehn 2005), the new Digital Corpus of the European Parliament (DCEP; Hajlaoui et al. 2014), the Multext project, and the OPUS open parallel corpus collection (Tiedemann & Nygaard 2004; Tiedemann 2009, 2012). The most widely used in the Natural Language Processing (NLP) community is the Europarl Corpus, which was released by Philip Koehn in 2005 and consists of the verbatim reports of the speeches made in the European Parliament's Plenary. Although initially it covered 11 languages, in its latest release in 2012 it includes 21 European languages with a final size around 60 million words per language.

The OPUS project deserves special mention in this respect (Tiedemann 2009, 2012). It started as a collection of translated texts from the web compiled with the goal of making parallel resources freely available, especially emphasizing the support of low density languages (Tiedemann & Nygaard 2004). Today it is the largest collection of freely available multilingual parallel corpora and a growing language resource covering over 90 languages and including data from several domains, mainly legislative and administrative texts, but also newspaper texts and smaller collections of subtitles and technical documentation.

The availability of these parallel resources has been enabled through research infrastructures such as CLARIN ERIC, which provides access to 86 parallel corpora,

[5] These include the full text corpora JRC-Acquis (Steinberger et al. 2006), DGT-Acquis and DCEP (Digital Corpus of the European Parliament), the translation memories (TMs) DGT-TM (Steinberger et al. 2012a), ECDC-TM and EAC-TM, and the document collection accompanying the multi-label categorisation software JRC EuroVoc Indexer (JEX; Steinberger et al. 2012b).

downloadable from national repositories as well as through concordancers such as Korp, Corpuscle, and KonText.[6]

Translators and other language practitioners have now at their disposal an unprecedented number of corpus resources and tools for a wide range of explorations and applications in the digital age. However, not all these resources cater for the needs of specific tasks or users. For example, if the aim is to compile a bidirectional parallel corpus which is balanced in terms of variables such as text type, genre, domain and directionality of the translation, researchers face a number of difficulties. First, most parallel data available are opportunistic, that is, limited by the availability of translated texts that are at one's disposal, which is usually imbalanced with respect to variables such as domain, mode, genre, number of texts and time spans. For example, large parallel corpora such as the ones compiled by the European Union organisations are restricted to the legislative or administrative domains, and although projects such as OPUS have extended the number of parallel data belonging to different domains, the imbalance still prevails.

Second, there is also imbalance in the direction of the translations available from one language to another. For example, the amount of scientific texts translated from English into Spanish is much greater than the number of scientific texts from Spanish into English (Lavid et al. 2015), and the same applies to other text types and language pairs, where English dominates as the main source language for many language pairs in genres such as movie subtitles or literary translations (see Frankenberg-García 2009). The different subcorpora available within the OPUS corpus reflect this asymmetry for a number of language pairs. For example, the number of aligned texts from English into Spanish doubles the amount of those from Spanish into English.

In the world of interpreters where the use of corpora is very recent (Corpas & Sánchez Rodas, this volume) and basically focused in the preparation phase, large parallel or comparable corpora are not tailored to the specific needs of interpreters, who require ad-hoc specialized corpora each time they engage with a new assignment, a new subject, or a new client (Fantinuoli 2018: 141).

In view of these difficulties, the last decade has witnessed a trend towards the creation of new corpus resources, which cater for the needs of different users. Thus, new corpora have been created for under-resourced language pairs, language varieties, specialized genres, historical periods or translation direction, as well as of tools which enable their compilation and processing.

[6] The bilingual corpora in the CLARIN infrastructure mostly contain European language pairs but also non-European languages such as Hindi, Tamil, and Vietnamese (https://www.clarin.eu/resource-families/parallel-corpora).

Two chapters in this volume address the need for new corpus resources in an under-resourced translation direction and in different historical periods of a given language. Thus, the chapter by Yi Gu and Ana Frankenberg-García in the first part of this volume is an example of how the need for corpora representative of an under-resourced translation direction (Chinese into English) motivated the creation of the ZHEN corpus, a unidirectional parallel corpus of circa one-million characters of contemporary simplified Chinese (ZH) source texts aligned with authentic translations into English (EN). The authors explain that the corpus resources available for the Chinese to English translation direction are scattered and not representative of current translation practices in China, which prompted them to compile this new resource with the aim of contributing to the understanding of Chinese to English translation features.

The chapter by Martín Arista points to the need for parallel corpora comprising texts from a historical period of a language and their modernised versions. The development of ParCorOE tries to fill this gap in this area, by compiling an aligned parallel corpus of Old English prose, aligned with Present Day English versions. The study focuses on areas of syntactic divergence between the aligned texts and describes four areas of alignment asymmetry between the source and the target texts.

2. Corpus-related tools

Together with advances in the compilation of bilingual and multilingual corpora, the last decades have also witnessed the development of different types of corpus-related tools and technologies for different tasks. These include translation memory (TM) systems and corpus management tools.

2.1 Translation memory systems

Translation Memory (TM) systems emerged as one of the earliest translation technologies as a result of advances in computing and computational linguistics in the late 1970s and early 1980s. The original idea of a translation memory was attributed to Martin Kay when he proposed a device which he called "the translator's amanuensis" which would be based on storing records of the translator's decision during the translation process which could be later examined when having to translate texts that contain similar material (Kay 1997: 19). Without explicitly calling this device a translation memory, in 1980 Kay was laying the foundations for the very popular idea that was implemented later on by commercial system in the mid-1990s. TM systems are software programs which enable translators to store previously translated texts and consult them for potential reuse in a new translation project (Bowker & Des Fisher 2010: 61). The source and the target texts are stored in a database as aligned bitexts, which are usually sentences, paragraphs or sentence-like units that have been previously translated individually or collectively around specific translation projects. A translation memory can thus be considered as "a parallel corpus which translators manually query for parallel concordances of (already translated) specific terms or patterns" (Zanettin 2002: 10).

TM systems are the main module within what are called Translation Environment Tools (TEnT), which may also contain additional tools such as terminological databases or termbases, automatic term extractors, bilingual concordancers, text analysers, spell-checkers, quality assurance checkers, and project management and translation workflow tools (Bowker & Corpas Pastor 2018). The integration of different tools into a common TEnT or translator's workbench facilitates streamlined workflow, although not all TEnT tools contain all possible components. The most widely used combination is that of a TM system in close association with a terminology database or termbase (i.e. a collection of systematically organized term records). Well-established segment-based systems (e.g. SDL Trados) and others taking a bi-text-based approach (e.g. MultiTrans) include both components (Bowker & Corpas 2018).

However, in spite of their success as the most widely used technological product for professional translators, TMs have a number of limitations: they are limited to specific and highly specialized text types from a narrow domain (e.g. online help files, manuals); they are "proprietory", i.e. created individually or collectively around specific translation projects and not much useful for new translation projects; the texts translated using TMs can be very repetitive since they standardize and restrict the range of linguistic options (Zanettin 2002: 10), and can become outdated for dynamically developing domains where new terminology is created daily (Corpas Pastor 2007). In addition, translators working with TMs have raised quality-related issues as well, the most significant being the "sentence salad" problem, due to the fact that a TM stores isolated segment pairs, rather than complete texts (Bédard 2000). This may result in a text that lacks cohesion and readability (Heyn 1998: 135).

As to the technology behind the development of TMs, they are based on pattern-matching techniques, i.e., comparing the segment of the new text to be translated with the contents of the TM database. If the TM system finds a match for a given segment, it offers it for the translator to accept it, modify it or reject it.

The first-generation TM systems were only able to search exact matches on the level of the sentence, but the second generation worked with fuzzy matches, i.e., segments which had some degree of similarity to segments stored in the TM database. A third generation of TM technology emerged in the 2000s based on the exploitation of subsentential matches (Gotti et al. 2005), and some scholars proposed to use 'semantic matching' by integrating WordNet into the TM system in order to identify synonyms (Mitkov 2005). Further developments and explorations have continued in the last years incorporating semantic textual similarity (STS) metrics and using machine learning techniques (Gupta et al., 2014). More recently, experts in the field have begun to explore innovative Natural Language Processing (NLP) and Deep Learning (DL) methodologies as the operational basis for a new generation of TM systems. These advances are a promising avenue of research for the improvement of the current state of TM systems, as shown by Ranshinghe et al. in their chapter on semantic textual similarity based on deep learning in the first part of this volume.

2.2 Corpus management tools

Despite the success and popularity of TM systems in the translation profession, corpora offer complementary advantages to TM systems in terms of ease of access, search flexibility and quality related issues (see Bowker & Barlow 2004, Corpas & Sánchez, this volume). But, as rightly pointed out by McEnery & Hardie: "the corpus alone solves few problems for a linguist. Its potential is unlocked by tools that allow linguists to manipulate and interrogate the corpus data in linguistically meaningful ways" (McEnery & Hardie 2012: 37).

The most important corpus management tools are undoubtedly concordancers or concordancing systems – also called corpus query systems (CQS) – that search through a corpus for each instance of a given word, phrase or other element and the immediate context in which each instance occurs, to create a concordance index. They typically display the search item in one-example-per-line format with a bit of context to the right and to the left of each example. The procedure is called key word in context (KWIC) concordancing and it enables users to search for single and multi-word units, suffixes, as well as tags, if the corpus is annotated. First generation concordancers ran on mainframe computers and only provided a KWIC concordance, but second-generation concordancers added some more functionalities, such as sorting alphabetically the left and right context of the search item, producing lists of words in alphabetical order, and calculating some basic descriptive statistics (McEnery & Hardie 2012: 39). Third-generation concordancers included a wider range of tools and were able to process large corpora on personal computers. Desktop concordancers such as AntConc (Anthony 2005), MonoConc (Barlow 2000) or Wordsmith (Scott 2008) became popular for working with monolingual corpora, while Multiconcord (Wools 1998) and ParaConc (Barlow 1995) were especially conceived for parallel corpora.

Later, a new generation of corpus analysis tools was developed to address issues such as the limited power of desktop computers, compatibility problems between operating systems, and legal restrictions on the distribution of corpora (McEnery & Hardy 2012: 43). These "fourth generation systems" began as websites for searching items in specific corpora but were later extended into generalizable systems, some of them allowing users to query large monolingual and multilingual parallel corpora through a web interface. For example, the original website developed by Mark Davies to access the British National Corpus (BNC) has been now extended to allow corpus queries on a wide range of very large corpora not only of the English language and its different varieties, but also of other European languages such as Spanish and Portuguese (Davies in press).[7] The original BNCweb (Hoffmann et al. 2008) was a web interface for just one corpus (the BNC), but it was later re-engineered as CQPweb (Hardie 2012), giving access to corpora of other languages and also to translated texts from different language pairs.

These "fourth-generation" web-corpus query systems satisfy simultaneously the need for power and the need for usability by using a client/server software model, where the query composed by the user does not run locally but on a powerful web server program. The advantages of these systems are manifold: they allow users to query large corpora across the web without infringing copyright laws because they only give access to text fragments; corpus searches are not limited to the memory and processing power of the user's desktop computer, and they run on every operating system. Another advantage is that they don't require specific technical competence on the part of the user, since they don't require installation (Hardie 2012: 384).

The technology behind these websites can be of two types: one is an SQL database system, as illustrated by the textual databases developed by Prof. Mark Davies at Brigham Young University and made available through his corpus.byu.edu interface (Davies 2005, 2019). The other is a dedicated corpus indexing and querying system, such as the Corpus Query Processor, (CQP) which is a part of the Open Corpus Workbench (CWB) (Christ 1994). BNCweb (Hoffmann et al. 2008) and its clone CQPweb (Hardie 2012) combine an SQL database with a CQP back-end.

The most popular web-based corpus-query system nowadays is SketchEngine, which uses a uses a CWB/CQP-compatible program, and is introduced by their creators as "a leading corpus tool" and "widely used in lexicography" (Kilgarriff et al. 2014). It allows users to query large monolingual and multilingual parallel corpora as well as to upload corpora following certain requirements.[8] As stated by Ana Frankenberg-García in the Sketch Engine web page:

[7] See Mark Davies' corpora page at http://www.corpus.byu.edu.

[8] Sketch Engine is one of the first and most widely used online systems that houses 500 ready-to-use corpora in 90 languages, many of them having a size of up to 30 billion words. It can be accessed at www.sketchengine.eu.

> The beauty of Sketch Engine is that with one single tool you can look up answers to all sorts of questions in many different languages. In the same way as CAT tools are like an enhanced text editor that has been adapted to the needs of translators, think of Sketch Engine as a kind of tweaked search engine that has been customized for linguistic analysis, helping you get unprecedented access to how language is really used. (www.sketchengine.eu)

The development of fourth generation concordancers marked a major milestone in the history of corpus analysis tools by providing immediate and easy access to large corpora in many languages. However, even though they are more powerful and user-friendly than third-generation corpus analysis tools, they do not cater for the needs of all possible users. For example, they are not as useful for the study of small corpora, or individual texts (McEnery & Hardie 2012: 46), nor for very specific domains or topics which may be necessary in the field of specialized translation.

Thus, research groups and scholars with specific research requirements have begun to design and build their own tools, either by programming the software themselves or collaborating with computer scientists. Two chapters in the first part of this book illustrate this trend: the chapter by Sánz-Villar and Andalus describes the design of the tool *TAligner 3.0* to create small parallel and multilingual corpora (around 3.5 million words), such as a corpus of narrative German-into-Basque literary translations for investigating phraseological units, or a small corpus of theatre texts from English into Spanish to investigate the translation of orality markers. One of the main advantages of this tool is that it integrates corpus alignment and query functions within the same program.

Pérez Blanco and Izquierdo present *Promociona Té*, a writing tool for Spanish professionals who need to write herbal tea promotional texts (HTPTs) in English. In view of the dearth of parallel data, researchers from the ACTRES research group developed this tool as an alternative writing aid, using the ACTEaS_Promo comparable corpus of online tea descriptions written in British English (EN) and European Spanish (ES) to conduct a contrastive rhetorical analysis to identify and tag functional chunks which could be used by the writing tool.

3. Impact on corpus-based translation and contrastive studies

Corpus technologies in the last decades have not only revitalized and given new impetus to traditional research approaches to Translation and Contrastive Linguistics, but have also contributed to significant advances and innovations in numerous contexts, such as translation practice, the teaching of foreign languages and translation, lexicography, and the development of multilingual NLP applications such as Statistical Machine Translation or bilingual lexicon extraction, to mention some of the most important ones.

In the field of Translation Studies, Tymoczko's prediction about the role of the corpus in translation has been confirmed with the establishment of Corpus-Based Translation Studies (CBTS) as a growing and fruitful area of translation research. The corpus-based approach has become a new paradigm in Translation Studies since the late 1990s, when it began to evolve into a "coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description and the practice of translation" (Laviosa 1998). Fifteen years later, this trend was consolidated with one out of ten publications in the field being concerned or informed by corpus linguistics methods (Zanettin, Saldanha & Harding 2015).

In the field of Contrastive Linguistics, a new corpus-based approach emerged as a thriving field in the 1990s, as a result of the rapid development of Corpus Linguistics and Natural Language Processing which began to focus on cross-linguistic issues in that decade. The availability of large bilingual corpora provided the empirical basis to address a wide range of cross-linguistic hypotheses which could not be confirmed by previous 'old style' contrastive analysis, which was intuition-based. A 'new era' in Contrastive Linguistics began to take shape based on the comparison of different languages on the basis of computer corpora and an extension of linguistic interest beyond syntax to pragmatics and discourse studies.

The contributions in the second part of this volume illustrate the vitality of research in these two converging disciplines –Translation Studies and Contrastive Linguistics- and their interconnection through the shared type of data (e.g. comparable and parallel corpora). The studies use these two types of corpora to address a number of linguistic phenomena in several languages such as English, German, Swedish, French, Italian, Spanish, Portuguese and Turkish, thus showcasing the richness of the linguistic diversity carried out in these recent investigations.

As to the themes addressed in these studies, they can be grouped into two major approaches: one based on the manual analysis and comparison of discourse and generic features of original and translated texts; the other based on the use of automatic statistical techniques for the identification of linguistic features.

The first approach is illustrated by several chapters which focus on the comparative analysis of discourse elements such as discourse markers (Lavid-López), pragmatic markers (Mendes & Zeyrek), and evidential expressions (Marín Arrese), highlighting their multifuncionality in different genres and comparing their behaviour in original and translated texts in different languages, such as English, Spanish, Portuguese and Turkish. Two chapters compare characteristic features of specific genres in English and Spanish, such as Westerns (Sanderson), and mobile application reviews (Mora López).

The second approach is illustrated by two chapters which describe the use of automatic statistical techniques to detect linguistic features of translations, such as relative entropy to measure the similarity/dissimilarity between probability distributions of words in original vs. translated texts (Karakanta et al.), and other association measures such as mutual information, log-likelihood ratio and reversibility score to identify multiword units such as binomials in parallel corpora (Graën & Volk).

4. Organization of the volume and chapter overview

The book is divided into two parts. The first part is dedicated to new corpus resources and tools and groups together six chapters which describe recent research efforts devoted to the creation of new parallel corpora for under-researched areas, the development of tools to manage parallel corpora or as an alternative to parallel corpora, and new methodologies to improve existing translation memory systems.

G. Corpas and F. Sánchez Rodas' chapter serves as an introduction to the whole volume by showcasing the range of language technologies that are currently applied to translation and interpreting. Such vast field is effectively organized into three strategic areas: (a) the automation of processes and the integration of language technologies in translators and interpreters' workflows and industry demands; (b) technology-based resources for oral and written mediation, ranging from new types of corpora for interpreting to computer-assisted interpreting tools (CAI) and cloud interpreting; (c) the notion of 'tech-savviness' and adoption of technology among language service providers. The chapter offers an essential overview of the role technology plays in translation and interpreting, identifying challenges and research opportunities for the future, but it also pinpoints the crucial contribution of corpora in the latest advances.

The second chapter by Y. Gu and A. Frankenberg-García discusses the need for corpora representative of the under-resourced Chinese-into-English translation direction. The authors provide a valuable overview of the current Chinese-English translation scenario before presenting the ZHEN corpus, a unidirectional parallel corpus of contemporary Chinese (ZH) source texts aligned with authentic translations into English (EN), which the authors have compiled from a wide range of sources aligned with authentic English translations.

In chapter three J. Martín Arista addresses a relatively unexplored domain within parallel corpora: the creation of aligned corpora for different diachronic versions of texts in one language, which may be regarded as a special subtype of parallel corpus. In the chapter, the author focuses on the requirements of syntactic annotation imposed by such corpus, which are derived from experience with a historical corpus of English. Syntactic divergence between previous periods of English and modernized versions is described in terms of alignment asymmetry types (markedness, constituency, order, and configuration), allowing to identify areas of stability and change in the grammar of English.

The next three chapters illustrate advances in the development of different types of corpus-related tools and technologies for different tasks, with one study devoted to translation memory (TM) systems and two contributions describing two new corpus tools.

In their contribution, T. Ranasinghe, R. Mitkov, C. Orăsan and R. Caro Quintana propose a new method that could improve the effectiveness of Translation Memory systems in retrieving a match for translation. While current TM systems rely on the Levenshtein distance metric -which is based on strict character comparison-, the authors explore the feasibility of applying metrics based on semantic textual similarity, which relies on deep learning and various vector-based representations of sentences. They carry out experiments which compare the performance of three sentence encoders to retrieve translation memory matches in English-Spanish sentence pairs with the results from Okapi, which uses edit distance to acquire the best match from the translation memory. The results obtained show the advantages of their approach based on deep learning techniques over traditional ones which use only edit distance.

The last two chapters in this first part of the book are devoted to the presentation of new tools which address specific research requirements which cannot be fully satisfied by general-purpose corpus resources and tools. In chapter five Z. Sánz-Villar and O. Andaluz-Pinedo present *TAligner* 3.0, a tool developed by the TRALIMA/ITZULIK research group to enable users to create small corpora simultaneously by aligning any number of texts (as required by the research), as well as to compile corpora consisting of different text types. *TAligner*'s main advantage is that it allows both corpus alignment and analysis within one tool, as well as to make queries within the same program and to align several texts simultaneously, which, as the authors acknowledge, is especially useful for the study of indirect translations or retranslations of the same source text. The tool has mainly been used with English, Basque, Spanish and German texts, but it is language independent. The authors also describe their experiences creating a small corpus of narrative texts –the *AleuskaPhraseo* parallel and multilingual corpus- consisting of German-into-Basque literary translations, and the *TEATRAD* corpus, consisting of original plays in English from the 1950s and 1960s and available (re)translations generated in Spanish in the 20th and 21st centuries.

In their turn, M Pérez Blanco and M. Izquierdo present *Promociona Té* in chapter six, a corpus-informed writing tool customised for small businesses as an alternative to translation, given the lack of Spanish-English translations in the domain of herbal tea promotional texts. The tool was developed by members of the ACTRES research group to assist Spanish professionals in writing promotional texts of their products in English, using a similar methodology to other text generators developed by this research group, based on the implementation of rhetorical and lexicogrammatical data extracted from English-Spanish comparable corpora. The main advantage of this tool is that it allows potential users, such as small local businesses, to dispense with the costs of buying in translation services or the costly post-editing of machine translation output. From the linguistic perspective, it offers an interesting contrastive analysis of the rhetorical and lexicogrammatical patterns of this specific knowledge domain.

The second part of the book consists of presentations of a number of corpus-based studies and explorations which address cutting-edge issues in the area of contrastive discourse studies and translation analysis. The first three chapters provide contrastive and translation analyses of discourse elements such as pragmatic markers, discourse markers and evidential expressions in different European languages, with a focus on their multifunctional and polysemic nature, and the challenges posed by these inherent features for corpus annotation purposes.

Thus, in chapter seven, J. Lavid-López scrutinizes discourse markers in the context of translation, by focusing on three highly frequent discourse markers in English -*in fact, actually* and *really*- and their Spanish counterparts. As the author rightly argues, the analysis of discourse markers is a much needed and challenging task not only for descriptive translation studies, but also for NLP applications. The reason is that discourse markers are multifunctional and often have a wide range of meanings, which are difficult to identify and annotate, even for trained human experts. The author uses translations and back translations of the three English discourse markers to capture a wide range of meanings and pragmatic functions, some of which had not been identified in previous monolingual analysis. She also puts forward a detailed annotation model for discourse markers so as to ensure the reliability and consistency of human-coded annotations.

Along similar lines, A. Mendes and D. Zeyrek examine how the discourse markers *well* and *so* are annotated in the English part of the TED-Multilingual Discourse Bank in comparison with their corresponding translations in Turkish and Portuguese. While the DM *well* is frequently omitted in the Turkish translations, the translations of *so* showed that in Turkish, this marker tends to be kept in the translation rather than omitted. The authors conclude that some languages are more prone to implicitation but only regarding DMs that do not have a discourse connective function.

Focusing on core evidential expressions (verb and sentence adverbs) derived from the domains of perception, cognition and communication in English and Spanish, J. Marín Arrese studies their indirect inferential and reportative values and their multifunctionality in two discourse genres: oral conversation and written journalistic discourse. The author examines cross-linguistic correspondences across the two languages (English to Spanish and Spanish to English) in order to build the core cross-linguistic paradigm of corresponding potential evidential expressions in both languages, using both parallel and comparable samples in both languages, from both discourse domains. On the basis of her corpus-based contrastive analysis, the author argues that variation in the use of particular values of evidentiality is sensitive to discourse domains and genres, and that multifunctionality seems to favour expressions derived from the perceptual domain. She also points out that the results of the study may be limited due to the reliability and validity of the Spanish corpus as truly representative of unscripted spoken discourse in European Spanish.

The next two chapters investigate linguistic features of two new emerging genres through the corpus analysis of English and Spanish samples. Thus, J. D. Sanderson analyzes the process by which the audiovisual translation of American films into Spanish has contributed to develop a sociolect with distinctive lexis. To this purpose, Sanderson compiles a parallel corpus of source texts in English and their translations into Spanish and carries out a diachronic analysis. He shows how the Western, the most distinctive American film genre, and culturally alien in origin to the Spanish target context has developed its own a sociolect and word usage rules, partially as a result of censorship during the Spanish dictatorship.

Using a comparable (English-Spanish) sample of mobile application reviews extracted from a larger corpus from Google Play Store, N. Mora López focuses on the study of what can be regarded as another new genre. Her detailed analysis, based on the Systemic-Functional approach, allows her to define this genre more specifically, distinguishing two optional stages (Evaluation and Description). Her results also show that the combination of the polarity of the contents of each stage together with the number of stars given to the item reviewed creates six patterns in these reviews. Interestingly, no noticeable differences between English and Spanish were observed, which seems to point to a globalization of the genre under scrutiny.

While the previous chapters are all manual analyses comparing discourse and generic features of original and translated texts, the last two chapters illustrate the use of automatic statistical techniques for the identification of linguistic features in parallel corpora in several Romance and Germanic languages. In Chapter 11, A. Karakanta, H. Przybyl and E. Teich use techniques such as relative entropy (Kullback-Leibler Divergence) and visualization with word clouds to identify features of translationese and interpretese that distinguish translated and interpreted texts from original ones. The method proposed allows them to show variation in translation according to two parameters: mode (translation vs. interpreting) and language pair/target language (German and Spanish as target languages, English as source) both at the lexical and grammatical levels. In addition, their approach offers significant benefits over standard corpus-based accounts based on relative frequencies: the identification is not biased towards particular linguistic features; it makes comparison easier by being able to incorporate a large number of variables, and it is also sensitive to low-frequency phenomena.

In the final chapter J. Graën and M. Volk describe their method to automatically identify binomial adverbs and their fixedness in the Sparcling corpus, consisting of parallel texts in sixteen languages from a cleaned version of the Europarl corpus. Adverbial binomials are constructions in which two words linked by a conjunction are used as a single unit that functions as an adverb in the sentence, whose identification is an important step for many language technology applications. While adverbial binomials are not problematic in machine translation, parsing and automatic annotation often produce inadequate results. By employing a combination of previously known measures (e.g. reversibility and statistical association), the authors manage to increase the list of adverbial binomials candidates for identification in a corpus.

As shown by the contributions included in this volume, corpus-based contrastive and translation research are areas that keep evolving in the digital age, as the range of new corpus resources and tools expands, opening up to different approaches and application contexts. We hope that this book contributes to this expansion and paves the way for further developments and explorations in the coming years.

Acknowledgements

We are very grateful to several UCM institutions for their support in the organization of the International Symposium PaCor 2018 (Parallel Corpora: Creation and Applications), as the second international event in the biennial series inaugurated in 2016 at the University of Santiago de Compostela. The event was held in conjunction with the XVI Encuentros Complutenses en torno a la traducción.

References

Aijmer, Karin, Bengt Altenberg and Mats Johansson (eds). 1996. *Papers from a Symposium on Text-based Cross-linguistic Studies*. Lund 4–5 March 1994 [Lund Studies in English 88]. Lund: Lund University Press.

Anthony, Laurence. 2005. "AntConc: a learner and classroom friendly, multi-platform corpus analysis toolkit," in *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, 7–13. Tokyo: Waseda University.

Baker, Mona. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In *Text and Technology: In Honour of John Sinclair*, ed. by Mona Baker, Gill Francis and Elena Tognini-Bonelli, 233–250. Amsterdam: John Benjamins. https://doi.org/10.1075/z.64.15bak

Barlow, Michael. 2000. *MonoConc Pro*. Houston, TX: Athelstan.

Barlow, M. 1995. *A Guide to ParaConc*. Houston, TX: Athelstan.

Bédard, Claude. 2000. "Mémoire de traduction cherche traducteur de phrases." *Traduire* 186, 41–49.

Bowker, Lynn and Des Fisher. 2010. "Computer-aided Translation." In *Handbook of Translation Studies*, ed. by Ives Gambier and Luc Van Doorslaer, 60–65. Amsterdam: John Benjamins. https://doi.org/10.1075/hts.1.comp2

Bowker, Lynn and Michael Barlow. 2004. "Bilingual concordances and translation memories: a comparative evaluation." In *Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training*, Geneva, Switzerland, 52–61.

Bowker, Lynn and Gloria Corpas Pastor. 2018. "Translation Technology." In *The Oxford Handbook of Computational Linguistics* (2nd edition), ed. by Ruslan Mitkov, 1–43. Oxford Handbooks Online. Mar 2015.

Corpas Pastor, Gloria. 2007. "Lost in Specialised Translation: The Corpus as an Inexpensive and Under-exploited Aid for Language Service Providers." In *Translating and the Computer 29: Proceedings of the Twenty-ninth International Conference on Translating and the Computer*, 29–30 November, London.

Christ, Oliver. 1994. "A modular and flexible architecture for an integrated corpus query system," in *Proceedings of COMPLEX'94*, 23–32. Budapest.

Davies, Mark. 2005. "The advantage of using relational databases for large corpora: speed, advanced queries and unlimited annotation," *International Journal of Corpus Linguistics* 10 (3): 307–34. https://doi.org/10.1075/ijcl.10.3.02dav

Davies, Mark. 2019. "The best of both worlds: Multi-billion word 'dynamic' corpora". In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019*, ed. by Piotr Bański. Mannheim: Leibniz-Institut fur Deutsche Sprache.

Davies, Mark. In press. "Constitución de corpus crecientes del español." In *The Routledge Handbook of Spanish Corpus Linguistics*, ed. by Giovanni Parodi, Pascual Cantos, Chad Howe.

Fantinuoli, Claudio. 2018. "The Use of Comparable Corpora in Interpreting Practice and Training". *The Interpreters' Newsletter* 23.

Fantinuoli, Claudio and Federico Zanettin (eds.) 2015. *New directions in corpus-based translation studies (Translation and Multilingual Natural Language Processing 1)*. Berlin: Language Science Press. https://doi.org/10.26530/OAPEN_559833

Frankenberg-García, Ana. 2009. "Compiling and Using a Parallel Corpus for Research in Translation." *International Journal of Translation* 21 (1–2): 57–71.

Gotti, Fabrizio, Langlais, Phillip, Macklovitch, Elliott, Bourigault, Didier, Robichaud, Benoit and Claude Coulombe. 2005. "3GTM: A Third-Generation Translation Memory." In *Proceedings of the Third Computational Linguistics in the North East (Cline) Workshop*. Gatineau, Québec. 26 August 2005, 8–16.

Granger, Sylvianne. 2003. "The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies?" In *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, ed. by S. Granger S, J. Lerot and S. Petch-Tyson, 17–29. Amsterdam: Rodopi.

Gupta, Rohit, Hanna Bechara, and Constantin Orasan. 2014. "Intelligent Translation Memory Matching and Retrieval Metric Exploiting Linguistic Technology." In *Proceedings of the Translating and Computer* 36, 86–89.

Hajlaoui, Najeh, David Kolovratnik, Jaakko Väyrynen, Dániel Varga and Ralf Steinberger. 2014. "DCEP –Digital Corpus of the European Parliament." In *Proceedings of the 9th Edition of its Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland. 3164–3171.

Hansen-Schirra, Silvia, Stella Neumann and Erich Steiner. 2013. *Cross-linguistic corpora for the study of translations. Insights from the language pair English-German*. Berlin: de Gruyter.

Hardie, A. 2012. "CQPweb – combining power, flexibility and usability in a corpus analysis tool." *International Journal of Corpus Linguistics* 17(3). 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Heyn. Matthias. 1998. "Translation Memories: Insights and Prospects." In *Unity in Diversity? Current Trends in Translation Studies*, ed. by L. Bowker, M. Cronin, D. Kenny & J. Pearson, 123–136, Manchester: St. Jerome Publishing.

Hoffmann, Sebastian, Stephan Evert, Nicolas Smith, David Y. W. Lee and Ylva Berglund Prytz. 2008. *Corpus Linguistics with BNCweb: A Practical Guide*. Frankfurt am Main: Peter Lang.

Hardie, Andrew. 2012. "CQPweb – combining power, flexibility and usability in a corpus analysis tool". *International Journal of Corpus Linguistics* 17:3 (2012), 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Johansson, Stig and Knut Hofland. 1994. "Towards an English-Norwegian parallel corpus." In *Creating and using English language corpora*, ed. by U. Fries, G. Tottie, and P. Schneider, 25–37. Amsterdam and Atlanta, GA: Rodopi.

Johansson, Stig. 1998. "On the role of corpora in cross-linguistic research." In *Corpora and cross-linguistic research: Theory, method, and case studies*, ed. by S. Johansson and S. Oksefjell, 3–24. Amsterdam and Atlanta, GA: Rodopi.

Kay, Martin. 1997. "The proper place of man and machine in language translation." *Machine Translation*, volume 12, Nos. 1–2, 1997, 3–23 (reprint from 1980) https://doi.org/10.1023/A:1007911416676

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Milos Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine: Ten years On." *Lexicography*, Springer Berlin Heidelberg, 2014, vol. 1, no 1, p. 7–36.

Koehn, Philip. 2005. "Europarl: A Parallel Corpus for Statistical Machine Translation." In *Conference Proceedings: the tenth Machine Translation Summit*, 79–86. AAMT, Phuket, Thailand.

Quirk, Randolph, Sydney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Lavid, Julia, Jorge Arús, Bernard DeClerck and Veronique Hoste. 2015. "Creation of a high-quality, register-diversified parallel corpus for linguistic and computational investigations." In *Current work in Corpus Linguistics: working with traditionally-conceived corpora and beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015)*, Procedia – Social and Behavioral Sciences 198, (2015), 249–256. https://doi.org/10.1016/j.sbspro.2015.07.443

Laviosa, Sara. 1998. "The Corpus-based Approach: A New Paradigm in Translation Studies." *Meta*, 43 (4), 474–479. https://doi.org/10.7202/003424ar

McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Mitkov, Ruslan. 2005. "New Generation Translation Memory systems." Panel discussion at the *27th international Aslib conference 'Translating and the Computer'*. London.

Sanjurjo-González, Hugo and Marlén Izquierdo. 2019. "P-ACTRES 2.0: A Parallel Corpus for Cross-linguistic Research." In *Parallel Corpora for Contrastive and Translation Studies. New* *resources and applications*, ed. by Irene Doval and Maria T. Sánchez Nieto, 215–232. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.90.13san

Scott, Mike. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software. Available online at: [http://www.lexically.net/wordsmith/] (accessed: 3 June 2011).

Sinclair, J. 1992. "The automatic analysis of corpora." In *Directions in corpus linguistics: proceedings of the Nobel Symposium 82*, Stockholm 4–8 August 1991, ed. by J. Svartvik, 379–97. Berlin and New York: Mouton de Gruyter. https://doi.org/10.1515/9783110867275.379

Steinberger, Ralf, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski and Signe Gilbro. 2014. "An overview of the European Unions highly multilingual parallel corpora." *Language Resources and Evaluation* 48(4): 679–707. https://doi.org/10.1007/s10579-014-9277-0

Steinberger, Ralf, Andreas Eisele, Szymon Klocek, Spyridon Pilos and Patrick Schlüter. 2012a. "DGT-TM: A freely Available Translation Memory in 22 Languages." In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*, 454–459, Istanbul, 21–27 May 2012.

Steinberger, Ralf, Mohamed Ebrahim and Marco Turchi. 2012b. "JRC EuroVoc Indexer JEX – A freely available multi-label categorisation tool." In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*, Istanbul, 21–27 May 2012.

Tiedemann, Jörg. 2012. "Parallel Data, Tools and Interfaces in OPUS." In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, ed. by N. Calzolari, J. Choukri, T. Declerck, M. Uğur Doğan, B.

Tiedemann, Jörg. 2009. "News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces." In *Recent Advances in Natural Language Processing* (Vol. V), ed. by N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, 237–248. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/cilt.309.19tie

Tiedemann, Jörg and Lars Nygaard. 2004. "The OPUS corpus – parallel and free." In *Proceedings of the 4th International Conference on Language Resources an Evaluation (LREC)*, 1183–1186, Lisbon, Portugal. May, 26–28 2004.

Tymoczko, M. 1998. "Computerized Corpora and the Future of Translation Studies." *Meta*, 43 (4): 652–660. https://doi.org/10.7202/004515ar

Wools, D. 1998. *Multiconcord*. Birmingham: CFL Software Development.

Zanettin, Federico. 2002. "Corpora in translation practice." In *Language Resources for Translation Work and Research, LREC 2002. Workshop Proceedings*, ed. by E. Yuste Rodriguez, 10–14.

Zanettin, F., Saldanha, G. and Sue-Ann Harding. 2015. "Sketching landscapes in translation studies: A bibliographic study." *Perspectives: Studies in Translatology*, Volume 23, 2015, 161–182. https://doi.org/10.1080/0907676X.2015.1010551

PART I

Corpus resources and tools

Now what?

A fresh look at language technologies and resources for translators and interpreters

Gloria Corpas Pastor[1,2] and Fernando Sánchez Rodas[1]
[1]Universidad de Málaga / [2]University of Wolverhampton

This chapter provides a brief outline of language technologies applied to translation and interpreting with a view to identifying challenges and research opportunities. Section 1 covers new trends within the automation of processes, the integration of language technologies in translators and interpreters' workflows and industry demands. Section 2 moves on to other relevant technology-based resources for oral and written mediation, including new types of corpora for interpreting, computer-assisted interpreting tools (CAI) or cloud interpreting, among others. Section 3 delves into the concept of 'tech-savviness' and adoption of technology among language service providers (LSPs). Final thoughts are presented as a conclusion in Section 4.

Keywords: translation technologies, technology-based training, remote interpreting, corpora, CAI tools, professional uptake, tech-savviness

1. Introduction

From ancient times, words and the messages they convey have played a very important role in mediated discourse. Just equipped with the power of words, over the years translators and interpreters have practiced their work on a daily basis. Both have relied heavily on dictionaries, glossaries, term spreadsheets and the like. Later on, e-resources and language technologies became translators' best friends.

Nowadays, the Digital Age has shaped the world of translators and interpreters to such an extent that 'tech-savviness' is not just a desirable feature, but a pressing need. However, there is a gap in the literature with regard to an updated account of technologies currently available or under development in the field of interpreting, as well as to trends that are rapidly shaping the future of the translation and interpreting industry. So far, most research has focused on computer-assisted translation

https://doi.org/10.1075/btl.158.01cor

(CAT) and machine translation (MT) tools, web-based resources and applications, such as glossaries, dictionaries, corpora, concordancers, terminology management systems, knowledge based tools, cross-linguistic information retrieval (CLIR) systems, etc., and their degree of adoption by translators (cf. Bowker and Corpas Pastor 2015).

However, it took some time until interpreters began to show some interest in technology, mainly digital resources and corpora. With the exception of some publications, much less attention has been devoted specifically to interpreting tools, which include (but are not limited to) Over-the-Phone, Remote-Video and Web-based interpreting.

This chapter provides a brief outline of language technologies in translation and interpreting, with a focus on the last decade. Our main goal is to contribute to identify expected trends and new challenges for the profession and the curriculum. The paper is organised as follows. Section 1 covers new trends with a focus on the automation of processes, the integration of language technologies in translators and interpreters' workflows and industry demands. Section 2 moves on to other relevant technology-based resources for oral and written mediation, including new types of corpora for interpreting, computer-assisted interpreting tools (CAI) or cloud interpreting, among others. Section 3 delves into the concept of 'tech-savviness' and adoption of technology. Finally, potentials for research and final thoughts are presented as a conclusion in Section 4.

2. Current and emerging trends

This section focuses on the observed trend of an increasing "detachment" of language service providers (LSPs) from the process and products of translation and interpreting. In the past, translation, and for that matter, interpreting as well were considered an 'art' or an 'aptitude', in much the same way as literary translation still is. Nowadays, translators and interpreters are becoming a sort of technology-mediated agents for multicultural and multilingual communication. The immediacy, art craft, and solitude of the past are being rapidly replaced today by a significant displacement via automation, real-time collaboration in crowd-powered systems and technology-orientated workforce models.

2.1 Automating translation and crowdsourcing

The celebration of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT) in Bulgaria (Temnikova et al. 2019), together with the 40th and 41st editions of the Translating and the Computer Conference

(TC40 and TC41) in London, have recently identified relevant issues that set the guidelines for present and future research on translation technologies. Some of the hottest topics (in alphabetical order) were collaborative translation, crowdsourcing, Neural Machine Translation (NMT), post-editing, and technological training, among others.[1]

Collaborative translation is closely linked to the use of cloud-based solutions, which Steurs (2016) lists as one of the main skills desirable for those working on the language industry of the 21st century. Cloud-based applications, such as Lilt,[2] Memsource,[3] and XTM Cloud,[4] eliminate the need for installation on a computer, but bring around new issues in relation to workflow management and sharing, safety, and translation quality (ibid.). Closely-related concurrent translation occurs when multiple individuals work on a text collaboratively and simultaneously in a cloud-based environment. Although some preliminary studies with surveys from users have been carried out, the lack of information about this modality still calls for a cautious approach (Gough and Perdikaki 2018).

Crowdsourcing is another type of collaboration which nourishes from innovative workflows such as those of cloud computing, and which has made a considerable impact in terms of how the industry and translation studies conceptualise and implement quality (Jiménez-Crespo 2017, 2018). Collaborative translation platforms, such as the cloud subtitling platform Amara,[5] the TED Open Translation initiative,[6] Youtube Studio,[7] and the Translate Facebook initiatives,[8] allow for a better integration and recognition of Non-Professional Translation (NPT). This kind of activity enjoys increased focus in Translation Studies, which would not have been possible without the impact of digital technologies and the emergence and consolidation of the World Wide Web, fostering the intersection of technology and sociological approaches in TS (Jiménez-Crespo 2019).

[1] Other well-mentioned topics not covered in this chapter were terminology (e.g. Wladyka-Leittretter, 2018) and MT evaluation (e.g. Esperança-Rodier and Rossi, 2019).

[2] https://lilt.com/.

[3] https://www.memsource.com/.

[4] https://xtm.cloud/.

[5] https://amara.org/en/.

[6] https://www.ted.com/about/programs-initiatives/ted-translators.

[7] https://studio.youtube.com/.

[8] https://www.facebook.com/TranslateFacebookTeam/.

When compared to its predecessor, Statistical Machine Translation (SMT), generic, freely accessible NMT systems such as DeepL[9] have proven to be overall better in terms of post-editing effort and quality of the final translation than customised SMT systems employed by in-house services (Volkart, Bouillon, and Girletti 2018). Further studies carried out in similar environments have described the comparison between the output of customised SMT and NMT systems from the point of view of professional human translators, suggesting that errors of deletion, substitution, insertion and word order are easier to identify in SMT, and confirming the ability of NMT to produce correct paraphrases (Mutal et al. 2019). End-users of translations, however, seem to prefer post-edited NMT to the raw output, which suggest that customers still value human intervention, even when translation production metadata, such as method and cost, are revealed (Girletti et al. 2019).

Post-editing (PE) can be defined as "the correction of machine translation output by human linguists/editors" (Veale and Way 1997). As Jeffrey Allen (2003: 297) details, "the task of the post-editor is to edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)." From the point of view of translators, MT in general and post-editing in particular have suffered historically from an unfavourable reputation.[10] Nevertheless, a survey of Guerberof Arenas (2013) showed that post-editing is a task for which professional translators have mixed feelings, not necessarily because of reluctance to the activity itself, but because of the different output qualities and the payment they receive. In the meanwhile, post-editing keeps growing as a service, and as proof, it now counts with its own international standard (ISO 18587: 2017). Vieira (2019: 320) states that the continuous integration of MT into CAT tools have now caused post-editing to be "in a state of terminological flux where it can be seen to comprise different tasks and procedures." Increased post-editing use has also led to a more accurate, research-based definition of post-editese as the simpler, more normalised, and more interfered language found in post-edited texts, in comparison to human-translated texts (Dames, De Clerq, and Macken 2017; Toral 2019; Castilho Resende, and Mitkov 2019). Because of all these findings, research on post-editing training is also a trend, especially at postgraduate level. In this sense, Plaza Lara (2019) recently conducted a SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis of the inclusion of MT and PE in the Master's Degrees of the European Master's in Translation (EMT) Network. The study concluded that the future perspectives in relation to MT demand a deeper focus on the

[9] https://www.deepl.com/translator.

[10] Vieira (2019: 319) locates the origin of these negative attitudes at the beginning of the MT discipline, as post-editing "came about as part of a paradigm where human editors assisted the machine rather than one where the machine assisted them."

technological competences of the students, although it highlights a clearly defined competence model and a high context adaptation capacity as positive strengths and opportunities of the analysed programmes.[11]

2.2 Displacing traditional forms of interpreting

Cloud computing and human-centered computing also lie at the heart of new trends in interpreting. A direct consequence of this is the virtualisation of the interpreter's workplace which involves distant participants and different types of virtual communication (Česonis, 2020). The notion of distance interpreting is linked to both 'remote interpreting' (RI) and 'simultaneous interpreting delivery platforms' (SIDPs). RI implies the provision of interpreting services from a distant site (usually with traditional interpretation equipment), while SIDPs refer to platforms that allow distance communication between participants via interpreters through cloud interpreting or in-house networks (usually with computer hardware and software). This technology-based distinction is becoming increasingly blurred due to the rapidly evolving RI modalities and the hybrid modes of cloud-based interpreting.

In a seminal work, Braun (2015: 346) defines remote interpreting as "the use of communication technology for gaining access to an interpreter who is in another room, building, city or country and who is linked to the primary participants by telephone or videoconference." Nowadays, RI comprises over-the phone interpreting, video interpreting and also cloud interpreting, although the terminology referred to interpreting modalities and configurations is far from being unified (cf. Braun 2015 and 2018).

Over-the-phone interpreting, often called *telephone-mediated interpreting, telephone interpreting* or *telephonic interpreting* in the literature (Wadensjö 1999; Gracia-García 2002; Mikkelson 2003; Locatis et al. 2010; Ozolins 2012; Price et al. 2012), is "the most common form of remote interpreting, [...] predominantly used for dialogue interpreting in the consecutive mode in community settings" (Wang 2018b). Telephone interpreting is employed in highly demanding medical and legal settings, in which some conditions are required for a high quality service, such as specific training for the interpreter or proper equipment to ensure audibility and accuracy (Kelly 2008; NAJIT 2009). Consequently, the main concerns are the lack of specialisation and visual cues, costs, and the connection quality (ibid.). Wang (2018b) describes this last point as a "prevalent and severe problem [...] due to

[11] Cf. also Guerberof Arenas and Moorkens (2019), about the integration in a Localisation Master's of two MT and post-editing courses and one MT project management module following the principles of project-based learning.

hospitals and police stations' use of equipment inappropriate for telephone interpreting services (phone passing, speaker phones, etc.)," and proposes the investment in "dual handset phones and videoconference interpreting facilities" from the side of the public service organisations. As this author points out, training is fundamental for overcoming these obstacles and must be included in future protocols for telephone interpreting at national levels, which could "differentiate pay rates according to practitioners' formal interpreting training" (ibid.). New contributions to research in this area include the implementation of online training programmes, proposals for in-company training, enhancement of the trainee's Intercultural Communicative Competence (ICC), or the use of role-plays for terminology practice (Ruiz Mezcua 2018). Results of a survey answered by 465 interpreters who work over-the-phone have led to conclude that compulsory training should also reach clients of this modality (Wang 2018a).

Remote interpreting by videoconference is termed *video interpreting*, but also *video-mediated interpreting* (VMI), *videoconference-based remote interpreting* (VRI), *videoconference interpreting* (VCI), *interpreting via video link*, or simply *remote interpreting* (Mouzourakis 1996; Mikkelson 2003; Moser-Mercer 2003; Locatis et al. 2010; Braun and Taylor 2012; Price et al. 2012; Braun 2013). The first experiments with interpreting via video link were conducted in the 1970s under the direction of supra-national institutions (Mouzourakis 1996), and some of them maintained the interest on this type of service because of the arising of physical limitations, e.g. the EU with its 2003 enlargement (Mouzourakis 2003 and 2006). In a parallel way, telephone interpreting in medical and legal settings have been gradually replaced by video interpreting from the 1990s onwards (cf. Braun 2015). This rapid implementation has led to mixed feelings towards video interpreting in the professional community, both in sign and spoken language (Wessling and Shaw 2014; Napier, Skinner, and Turner 2017; Seeber et al. 2019), but reports on burnout and stress seem subjective, as their cause is not clearly defined (Alley 2014; Bower 2015).

Training is fundamental to cope with the emotional and technical difficulties involved in RI. In this line, it is worth mentioning the AVIDICUS project, which has resulted in the publication of training material online, alongside with a new videoconference training service for which anyone interested can apply (cf. AVIDICUS 2019).[12] Overall, video interpreting is nowadays generally regarded as a more effective modality than telephone interpreting, mainly because of the inclusion of non-verbal communication (eye gaze, gestures, etc.) and the ease of

[12] AVIDICUS (Assessment of Video-Mediated Interpreting in the Criminal Justice System) comprises three European collaborative projects which were carried out from 2008 to 2016 with financial support from the European Commission's Directorate-General for Justice and focused on video interpreting in legal proceedings (AVIDICUS 2019).

interaction which the widespread video technologies provide (Skinner, Napier, and Braun 2018: 1).

As Corpas Pastor (2018: 153) states, "the basic distinction between onsite and offsite technology marks the shift from over-the-phone and video remote interpreting to cloud interpreting, i.e. video remote interpreting where the videoconferencing is also online." Cloud-based interpreting, also called *Virtual Interpreting Technology* (VIT), is experiencing a very fast growth (especially in the field of conference services) because of the obvious connectivity advantages which offsite dynamics offer both for interpreters and for the rest of participants (Amato, Spinolo, and González 2018; Naimushin 2019). Web Real-Time Communication (WebRTC) technology has also played an important role in this expansion, making it possible to use these online platforms (SIDPs) from free browsers such as Google Chrome or Mozilla Firefox (ibid.). Cloud-based interpreting solutions, such as Akorbi, InDemand Interpreting, Interprefy, Speakus, TikkTalk or Veasyt Live lead the market up to date (Corpas Pastor 2018; Naimushin 2019). SIDPs may not yet be good for large multilingual meetings or conferences, but they have proved to be very suitable in certain situations (conflict zones and war, emergencies, healthcare, etc), during short bilateral meetings in remote locations, or even as portable interpreting systems (cf. Porlán Moreno 2019).

Some examples of successful SIDPs are TikkTalk or InDemand. TikkTalk has enjoyed special attention in the field of interpreting in asylum settings, where it has received positive reviews by stakeholders when compared to the traditional interpreting agencies (UNHCR 2016; Wasik 2017). InDemand, that integrates video remote interpreting services with telehealth, has been proven to be more efficient than telephone interpretation for improving patient care of families with limited English proficiency in the paediatric field (Lion et al. 2015), as well as preferred by healthcare providers in what refers to nursing time and communication (Marcus 2017). These video RI systems are clear examples of novel, speedy, low-cost solutions to communications needs in hospitals and healthcare centres.[13] As a direct consequence of the use of this technology, these systems are rapidly displacing traditional forms of interpreting.

[13] Medical interpreting in general is giving rise to other profound changes in the sector, although they are not necessarily related to technology. For instance, in the United States, dedicated, professional interpreters and *ad hoc* interpreters are being increasingly replaced by language concordant providers (LCPs) and dual-role interpreters (Nash Fernandez 2017).

3. Some data tools and resources

The previous section has presented some trends that are already shaping translators and interpreters' work environments nowadays. Core to both is a common thread of 'virtualisation' and 'technologisation'. However, technology growth and digitalisation in interpreting still appears rather limited and slow-paced, as compared to translation, despite some evidence that the profession is heading towards a technological turn (Fantinuoli 2018a and 2018b). While language technologies have already had a profound transformative effect in translation, they have not yet led to a paradigm shift to the interpreters' "digital workplace". Another difference is the degree of automation achieved. While MT (especially neural MT) is gaining ground among translators, machine interpretation is not a reality yet.

In what follows, other relevant technology-based tools and resources for oral and written mediation will be discussed. We do not intend to provide a comprehensive account or deal with the existing plethora (see, for instance, the papers in Corpas Pastor and Durán, 2018), but just focus on two which appear to be particularly relevant in the present volume: corpora and computer-assisted interpreting tools. Both can be taken as a sort of simplified benchmark to compare translators and interpreters in terms of tech-savvines.

3.1 Corpora

Translators tend to use a myriad of tools, integrated applications, virtual platforms and resources in their daily work. The most popular CAT tools are translation memory systems, terminology tools, localisation tools, quality checkers and the like. Machine translation is also increasingly being used, thanks to the dramatic improvements brought about by NMT. As a result, post-editing seems to have entered translators' workflow to stay, although little is known about how translators and other end-users engage with NMT output (see the papers in the special issue edited by Castilho et al., 2019).

Regarding CAT tools, they still present shortcomings despite their popularity and usefulness. They present a high learning curve for translators and trainees, and licences can be equally expensive for these private users.[14] Besides, the use of translation memories and predefined terminology banks does not update accordingly to the development of specialised domains and the creation of neologisms. This recurrent use of previous translations also fosters the perseverance of translationese,

[14] However, nowadays many TM producers, SDL and Wordfast, among others, provide free licences for research and education.

that is, translations' own lexico-grammatical and syntactic fingerprints (Gellerstam 1986; Ilisei et al. 2010).

Corpora and corpus management tools offer complementary advantages to CAT tools.[15] A corpus is "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" (EAGLES 1996: 4). Traditional types of corpora include parallel, comparable, and ad hoc/DIY collections (Corpas Pastor 2001: 158), which are now complemented by web and gigatoken web corpora (cf. Bowker and Corpas Pastor 2015). Corpora and corpus management tools are generally easier to deal with and count with several licence-free options. Moreover, they are able to process larger amounts of data, and account for reusability, modularity, and flexibility. The possible applications of corpora in translation are numerous and well-documented in the literature, and include: problem solving related to documentation, terminology and phraseology, the application of register and genre-specific conventions, the description of translation strategies and decision making processes related to textual equivalents, revision, evaluation and Quality Assurance (QA) of translation, and the study of naturalness and biculturalism (cf. Arire 2014, Fantinuoli and Zanettin 2015, Corpas Pastor and Seghiri 2016; Hu 2016; Mikhailov and Cooper 2016; Malamatidou 2019). In the field of Natural Language Processing (NLP), corpus-based machine-learning studies have made possible a deeper understanding of translationese. They sought evidence for the existence of the convergence and simplification universals by providing feature vectors for the machine learning process (Corpas Pastor et al. 2008, Ilisei et al. 2010, Ippolito 2013, Rabinovich and Wintner 2015, Zhang and Toral 2019).

However, corpora research and corpora-related practice also face their own shortcomings. Above-mentioned translationese is still a present feature in parallel corpora. In addition, ad hoc corpora compilation usually results in small-sized products with a narrow scope. Ideally, corpora should strive for a right balance between size and representativeness of the text type or domain in question, but compilation must not breach data capture and copyright rules. This balanced view defies more traditional positions in corpora studies which followed the principle "the more data, the better". From the side of NLP, different types of errors (language identification and segmentation, deduplication, misrecognition of character, spelling, etc.) must be corrected in the pre-processing, processing, and alignment of corpora, as well as in the mark-up, encoding, and annotation.

[15] In fact, NMT owes its great success to the use of huge training corpora, corpus augmentation and corpus-enrichment of their outputs (see, for instance, Karanta, Dehdari and van Genabith 2018, or Zang and Matsumoto, 2019).

By contrast to the widespread use of corpora among translators (both comparable and parallel), interpreters have timidly started to use corpora (cf. Russo, Bendazzoli and Defrancq 2018). However, corpora and corpus management tools can provide interpreters with valuable lexical, terminological, and phraseological information in real contexts (cf. Corpas Pastor and Seghiri 2016). These types of tools also reduce the cognitive load of the process and are expected to improve quality (cf. Straniero and Falbo 2012; Aston 2015). Interpreters are starting to use corpora mainly in the preparation phase (Fantinuoli 2017 and 2018c, Pérez Pérez, 2018, Xu 2018), although their popularity remains rather low.

In an attempt to bridge the gap between corpus-based translation and interpreting studies, which could foster mutual feedback by comparison (cf. Gile 2004), intermodal corpora are one of the growing trends today. First described by Shlesinger (1998), the idea was to add translations to monolingual comparable corpora of interpreted and non-interpreted speeches. By doing so, users can observe the "differences between the oral and written modalities of translation, [and] to observe the effects of the ontology variable (original vs. translated) as well" (Shlesinger and Ordan 2012: 47). Some intermodal corpora are also Multiple Translation Corpora (MTC). As Ferraresi (2016: 27) puts it:

> By bringing together multiple target texts (TTs) of a single source text (ST), MTC make it possible to observe and compare the strategies adopted by different translators, be they professional or trainees, when faced with the same input.

The first intermodal MTC was mentioned in Shlesinger (2008). The European Parliament Interpreting and Translation Corpus, or EPTIC (Bernardini, Ferraresi and Miličević 2016), is another remarkable contribution to this category, with further follow-up studies (e.g., Bernardini 2016 and Bernardini, Ferraresi, Russo, Collard and Defrancq 2018). These types of corpora have also been used to study interpretese, i.e. the extrapolation of translationese and translation universals to interpreted speeches, whose first mention was in Fumagalli (1999).

Interpreting and intermodal corpora still face severe limitations. They do not align interpreting speeches authentically, but transcriptions of interpretations, with no audiovisual data. Therefore, they do not have an aligned oral component, nor aligned spoken or paralinguistic features. Other negative points are data sparseness and unrepresentativeness. For NLP, the challenges of creating such a tool are considerable. Compilation of interpreting corpora is complex and time-consuming, and the transcription process is lengthy (Thompson 2005). Segmentation must deal with a continuous speech, and processing must take into account the existence of language varieties, diverse disfluencies, repetitions, hesitations, and mistakes, as well as emotional and pragmatic issues (Corpas Pastor 2018).

3.2 CAI tools

A similar pattern of interpreters' 'techno-shyness' (for lack of a better word) can be observed with regard to interpreting-related technologies. In general, the number and types of computer-assisted interpreting (CAI) tools appear rather limited: mainly, terminology management tools and note-taking applications (e.g., digital pens, tablets, iPads). Voice-text devices convert speech into text automatically. Some examples are Voice Dictation for Pages (iOS), Voice Pro (Android), Voice Dictation (iOS, Android, and Linux) and AudioNote LITE (Windows, MacOS and Android). Although they have been sometimes included as a type of CAI tool (cf. Corpas Pastor 2018: 142), they have not been designed with interpreters' in mind and could be considered as a multi-purpose, general tool targeting a very wide audience.

Terminological tools[16] are crucial in interpreting, since "the acquisition of terminology and specialized knowledge prior to a technical conference represents a fundamental phase in the interpreter's workflow" (Prandi 2018: 29). Even though the existing tools are user-friendly, they are usually restricted to one platform, e.g. Intragloss for MacOS, LookUp and Terminus for Windows, or Glossary Assistant and InterpretBank for both Android and Windows. Others are cross-platform (e.g. Interpreters' Help and Flashterm) or web-based (EU-Bridge). Most of the tools can process only glossaries, which must be manually elaborated, except in the case of EU-Bridge, which includes a term extraction and named-entity recognition module. Import options are normally included, but they are limited to Word, Excel, or exclusive formats for each tool. To conclude, most of them are oriented to preparation assistance only, except two booth-friendly choices: BoothMate (the offline companion app of InterpretHelp) and the latest version of InterpretBank.[17]

Note-taking applications refer to assistants for consecutive interpreting, mostly digital pens. Digital pen technology was first introduced around 2010, and it has progressively received academic attention (Hiebl 2011; Kostal 2011; Orlando 2010, 2011, 2013 and 2014). Corpas Pastor (2018: 145) describes the tool as follows:

> A digital pen is a writing or scanning tool capable of capturing and storing notes, text or drawings to upload to a computer. This type of smart pen is often used in conjunction with digital paper to create digital handwritten documents that can be edited at a later time. Some of them also feature Bluetooth antennas that transmit stored data wirelessly.

[16] On terminology tools for interpreters, see Costa, Corpas Pastor and Durán Muñoz (2017) and Rütten (2017).

[17] http://interpretbank.com/.

Corpas Pastor (ibid.) mentions two types of digital pens. A first group comprises those whose main functionalities are to take notes electronically, make sketches, and share them by e-mail (Inkeness, LectureNotes, PenSupremacy, My BIC Notes, Smarssen Bluetooth, Neo N2, or the Wacom smartpads). The second group has expanded functions, as they are capable of recording spoken words and synchronising them with notes that users manually write on special paper (Sky Wifi, Echo, Livescribe, or Equil Note). This last group is often referred to simply as *smartpens* (Orlando 2016: 102).

Notwithstanding the fact that digital pens were not originally designed as a tool for interpreters, their use has paved the way for the implementation of a new interpreting modality called simultaneous-consecutive interpreting or hybrid interpreting, "an exciting combination of two interpreting skills + portable technology, which is quickly becoming the technique of choice for today's interpreter" (Navarro-Hall 2012, quoted in Orlando 2015b: 12). Regarding traditional interpreting modes, the interest on these new, multi-faceted devices by professional interpreters has also been documented through several means: blogs, Q&A sites like *Interpreting.info*,[18] or tweetchats (Orlando 2015b). They have also been successfully tested in consecutive interpreting training for improved note taking and feedback based on metacognitive and cross-student strategies. In these surveyed sessions, students and educators praised the digital pens for helping them to better analyse and diagnose problems, thanks to their recording and playback functions (Orlando 2015a). A later study by Kellett, Bidoli, and Vardè (2016) suggest that digital pens could be more useful in hybrid interpreting than in the traditional consecutive-only mode.[19]

Tablets are another type of note-taking tool. Although Hof (2012) early described tablets as "the ideal boothmate", and Drechsel (2013b) published a first tablet interpreting manual, academic research in this area still needs more time to thrive. Tablets can be integrated in the interpreter's workflow in several ways; for a paperless preparation phase, for taking advantage of sophisticated PDF reading and editing features and glossary management apps,[20] and for split-screen visualisation (Drechsel and Goldsmith 2016). Atabekova et al. (2018: 352) mention iPads and tablets at the same level of laptops, in the category of "standard equipment supporting and facilitating the interpreter's work as cross-cultural mediator." Additionally, a pilot survey by Goldsmith (2018: 19) with half a dozen interpreters

[18] http://interpreting.info/.

[19] The potential of using smartpens for studying consecutive interpreting from a cognitive perspective has been explored by Chen (2017, 2018a and 2018b).

[20] Cf. Costa, Corpas Pastor and Durán Muñoz (2014a and 2014b).

in two continents demonstrates that tablets enjoy a regular and effective use in consecutive interpreting, suggesting that "tablet interpreting can equal pen-and-paper interpreting" and is "generally well received in most settings", with the exception of some concerns.[21]

4. Translators and interpreters' technology uptake

It is commonly believed that translators have at their disposal a higher number of (better) computerised tools and resources than interpreters, and that technology has shaped the work of translators to a higher extent. In the words of Fantinuoli (2018b: 1): "When compared to written translation or other language professions, the advances in information and communication technology have had a modest impact on interpreting so far." In fact, a cursory search in Google shows 320,000,000 hits for "*translation technology*" versus 58,900,000 for "*interpreting technology*"; a similar picture emerges as regards the following search strings: [translation + technology] provides 743,000,000 results, but [interpreting + technology] returns only 71,000,000.[22]

In what follows we will offer a brief overview of technology adoption among translators and interpreters, with a view to compare them in terms of resistance/accommodation to automation and computerised tools.

Although translators have been said to be generally well disposed to technology (Koskinen and Ruokonen 2017; LeBlanc 2013 and 2017), the adoption has not been straightforward. Fears of being replaced by machines or losing their jobs have been looming from the very beginning. Nowadays translators have finally accepted those computer-assisted tools and learned how to use them to their maximum potential as a means to increased productivity and quality improvement, to such an extent that translation can be considered as a form of human-computer interaction (O'Brien 2012).

There is still some technology resistance, though, when it comes to full automation of the translation process. Main sources of concern seem to be technology's limitations and market consequences, rather than fears of being outperformed by MT systems (Vieira 2018: 1). This 'automation anxiety', as Vieria calls it, seems to cause mixed feelings within the profession. In a study on two groups of professional translators (users and non-users of MT), Cadwell, O'Brien and Teixeira (2018)

[21] See Goldsmith and Holley's (2015) summary of pros and cons of tablet interpreting in relation to technical, visual, physical, and client relation aspects.

[22] Google searches were carried out on 31st January 2020.

reported that participants provided an equally diverse set of reasons for using MT as for not using it (depending on text type, language pairs, output quality and trust).[23]

Some survey studies have provided more detailed insights into translators' uptake of MT and CAT tools. One of the latest survey studies by Zaretskaya, Corpas Pastor and Seghiri Domínguez (2015) is based on an online questionnaire that is used to draw conclusions regarding MT, Translation Memory (TM) software, and corpus usage. Over 700 respondents showed that MT popularity is low because of the poor quality and confidentiality requirements, but translators would benefit from it if they knew accuracy is high or very high. Among TM software features, professionals value high working speed, simplicity, and flexibility (e.g. compatibility and integration of additional features). As regards corpora, they seem to be less popular than other electronic resources and CAT tools. From those who admit using corpora, a majority preferred to use available ready-made corpora and online resources rather than compiling their own ad hoc corpora, and when doing so, they generally did not resort to corpus compilation and management tools. A second study (Zaretskaya, Corpas Pastor and Seghiri Domínguez 2018) confirmed the previous findings in terms of simplicity and preferred module integration for all CAT tools. Remarkably, terminology management was simultaneously one of the best-loved and most hated functionalities by respondents.

All respondents in the previous study were experienced translators working in specialised fields. This professional profile is more inclined to use technology, due to the very nature of the texts to be translated (more repetitive, less creative, more homogeneous, technical, etc.). But another recent research trend is beginning to study the interaction of literary translators with technology. This recently surveyed topic is also present in TC40 (see Section 1.1.). The interesting thing is that literary translation is regarded as "the last bastion of human translation" (Toral and Way 2014: 174). Thus, deepening in this field will enable researches and practitioners to understand reluctant positions at a more general level in translation.

A preliminary study for an ongoing doctoral research project by Ruffo (2018) found that professional literary translators showed mixed feelings, making a sharp distinction between general technology and translation-specific technology. In particular, translation-technology in the form of Computer-aided Translation (CAT) tools and Machine Translation (MT) is dismissed in all answers, among claims that it will never catch on in the field of literary translation as it did for non-literary translation. By contrast, participants praised the role of general technology, namely corpora, terminology tools and the Internet as research tool (Ruffo 2018: 130).

[23] In this same vein, some scholars have lately raised critical voices. They focus on ethical issues, making critical reviews on the transition from SMT to NMT or on the general history of MT (Kenny 2018 and 2019).

A later study by Ruffo (2019) reports the existence of a clear relationship between perceptions of role and attitudes towards technology and that the adoption of a proactive and collaborative approach between different social groups could benefit the process of technological innovation in literary translation (Ruffo 2019).

(Partial) resistance to technology in literary translation could be considered a mid-way scale point between two opposites in terms of acceptance within the profession: CAT versus CAI tools. In the case of interpreting, "there is still a scarcity of empirical studies about the extent to which interpreters have embraced technology" (Corpas Pastor 2018: 157). In fact, the resistance to technology has been always higher in this field (cf. Berber-Irabien 2010). Corpas Pastor and Fern (2016) conducted an online survey addressed at international interpreting associations, forums and freelance interpreters. Over 50% of respondents in this study did not use any technology tool or resource during their work, but their attitude towards these advances was positive. Three years after its completion, in 2017, the same questionnaire (with minor modifications) was distributed to students of the National Center for Interpretation of the University of Arizona (see Corpas Pastor 2018 for details). Respondents also showed a positive attitude towards technology, but they were concern about the robustness and pricing of these tools. Overall, both professional and students of this field seem largely unaware of interpreting technology, and those who show competence and admit using CAI tools rarely use corpora.

In a recent study, Pielmeier and O'Mara (2020) report results from a large-scale survey of over 7,000 experienced translators and interpreters from all corners of the world that either work as freelancers or in-house at language service providers (LSPs) or buy-side companies. All respondents have worked for more than 14 years, but just over half had some formal training. The report provides detailed data on languages and services offered, background and career, clientele, technology usage, income from language services, volunteer work, and the evolving future. We will focus on technology use, which included tech-savviness, vendor portal usage, and perspectives on translation and interpreting technologies.

As Pielmeier and O'Mara say (ibid. 38): "The translation industry is continuously adding new software to support linguist work, thereby making the profession very tech-oriented". When asked about their degree of tech-savvines, only 7% did not feel very comfortable with technology, over one-half (51%) reported feeling very comfortable and ready to try out new software, and the balance (42%) fell somewhere in between. Technology is also present when it comes to retrieving jobs, delivering files, and submitting invoices: only 11% of respondents reported using no portal at all, while more than two-thirds (69%) of portal users believe it streamlines their work, although they find those tools rather impersonal.

Regarding language tools and resources, translation memories appear in top positions (used on most projects by 66% of respondents), followed closely by quality checkers (60%) and terminology management tools (48%). Machine translation is used only by 22% of respondents, but only a third of them are satisfied with the overall quality of the output. Glossaries and terminology tools are highly valued for quality (91%), whereas TMs appear to enable fast delivery (86%).[24]

As to which interpreting technologies respondents use on client assignments, the authors focus on several items and their percentages of users: over-the-phone interpreting platform (43%), video remote interpreting platform (26%), interpreter portal (24%), interpreter console (22%), computer-assisted technology (18%), remote simultaneous interpreting platform (17%), interpreting management system (12%) and machine interpreting platform (11%). However, most of these interpreting-related technologies are modalities of distance/remote interpreting. It remains unclear what the authors mean by interpreter portal or machine interpreting platform. Besides, computer-assisted technology is a very vague concept that could encompass diverse interpreting-related tools and resources. In any case, a clear conclusion from this study is that most respondents prefer to interpret in person. An overwhelming percentage (74%) misses in-person interactions, among other reasons. However, remote interpreting has the advantage of increasing productivity and it is perceived as positively challenging.

Finally, respondents view the future in terms of translator productivity, market changes and perspectives on the profession. The three main conclusions are: (a) productivity is increasing thanks to experience and technology; (b) changes in the market evidence greater pressure to offer services at lower prices or with faster turnaround times; and (c) professionals are also concerned about fierce competition, the drop in market demand and the impact of artificial intelligence. All of them are intimately related to technology tools and the interaction with them.[25]

5. Food for thought

According to Pym (2011: 4), "resistance to technological change is usually a defense of old accrued power, dressed in the guise of quality." While we agree with that, it should be worth noting that resistance to change and fear of the unknown is

[24] The survey included questions about automatic content enrichment (ACE). This new type of technology makes texts more intelligent as it can provide external links, concepts, images etc. related to both source and target texts. ACE seems to be utterly unknown by most translators (73%). Around a fourth has heard about it but has never used it.

[25] See Olohan (2017 and 2019) on constructivist approaches to the translation technology phenomenon.

connatural to human beings, and so these could be factors also playing an important role in the way language professionals view/interact with technology.

Adoption of technology is also a gradual process. For instance, Sgourou (2019), a freelance translator with around 20 years of experience, identifies four stages of machine translation acceptance: (1) "nescience", (2) "contempt", (3) "reluctant adoption and shame", and (4) "acceptance". She urges MT developers and translators' trainers to consider them in order to help future professionals move from stage to another.

The former scale could be applied to understand the present-day situation of CAT and CAI tools, for instance. In the case of translation-related technology, almost all professionals would be already in stage 4 (including MT users). The situation is somewhat different when it comes to technology for interpreters. Remote interpreting could be considered to have pushed professionals into stages 3–4. They seem to have valid reasons to support or reject new technology-based modalities for distance and displaced interpreting. However, when it comes to tools and resources used prior to or during an interpretation, interpreters would be better placed within the realms of "nescience" (stage 1) or even "contempt" (stage 2).

In order to develop suitable products, and considering the rapid and exponential growth of technology, the changing needs and expectations of professional translators and interpreters must be constantly surveyed with up-to-date studies. In this sense, future research must also support the impact of formal education and training in the adoption technology, as well as monitor the important consequences of technology integration in the market and workflow.

Research should also keep other relevant topics on the spotlight, such as subjective perceptions about technology ('automation anxiety' and 'mixed feelings' issues), technology used by both translators and interpreters (e.g., terminology management systems, cloud systems, remote and collaborative platforms), technology specifically oriented to translators or to interpreters, technology-enriched life-long training and formal education, the potential of artificial intelligence and augmented reality for the profession, and so forth. Technology is always disruptive, but it always pays off.

Acknowledgements

This paper has been carried out in the framework of various research projects on language technologies applied to translation and interpretation (ref. FFI2016-75831-P, UMA18-FEDERJA-067, CEI-RIS3 and EUIN2017-87746).

Funding

It has also been partially funded by the Spanish Ministry of Education (ref. FPU18/05803).

References

Allen, Jeffrey. 2003. "Post-editing." In *Computers and Translation: A Translator's Guide*, ed. by Harold Somers, 297–317. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.35.19all

Alley, Erica. 2014. "Who Makes the Rules Anyway? Reality and Perception of Guidelines in Video Relay Service Interpreting". *The Interpreter's Newsletter* 19: 13–26.

Amato, Amalia, Nicoletta Spinolo, and María Jesús González Rodríguez. 2018. *Handbook of Remote Interpreting*. https://doi.org/10.6092/unibo/amsacta/5955

Arhire, Mona. 2014. *Corpus-based Translation for Research, Practice and Training* (Topics in Translation Series). Iaşi: Institutul European.

Aston, Guy. 2015. "Learning Phraseology from Speech Corpora." In *Multiple Affordances of Language Corpora for Data-driven Learning*, ed. by Agnieszka Leńko-Szymańska and Alex Boulton, 63–84. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.69.04ast

Atabekova, Anastasia A., Rimma G. Gorbatenko, Tatyana V. Shoustikova, and Carmen Valero-Garcés. 2018. "Cross-cultural Mediation with Refugees in Emergency Settings: ICT Use by Language Service Providers." *Journal of Social Studies Education Research* 9 (3): 351–369. https://jsser.org/index.php/jsser/article/view/274. (Accessed January 3, 2020).

AVIDICUS. 2019. *Video-Mediated Interpreting: Home of the AVIDICUS Projects*. http://wp.videoconference-interpreting.net/. (Accessed January 3, 2020).

Berber-Irabien, Diana. 2010. Information and Communication Technologies in Conference Interpreting. PhD thesis. Barcelona: Universitat Rovira i Virgili.

Bernardini, Silvia. 2016. "Intermodal Corpora. A Novel Resource for Descriptive and Applied Translation Studies". In *Corpus-based Approaches to Translation and Interpreting*, ed. by Gloria Corpas Pastor and Miriam Seghiri Domínguez, 129–148. Bern: Peter Lang.

Bernardini, Silvia, Adriano Ferraresi and Maja Miličević. 2016. "From EPIC to EPTIC–Exploring Simplification in Interpreting and Translation from an Intermodal Perspective". *Target* 28 (1): 61–86. https://doi.org/10.1075/target.28.1.03ber

Bernardini, Silvia, Adriano Ferraresi, Mariachiara Russo, Camille Collard and Bart Defrancq. 2018. "Building Interpreting and Intermodal Corpora: A How-to for a Formidable Task". In *Making Way in Corpus-based Interpreting Studies*, ed. by Mariachiara Russo, Claudio Bendazzoli and Bart Defrancq, 21–42. London: Springer. https://doi.org/10.1007/978-981-10-6199-8_2

Bidoli, Cynthia J. Kellet, and Sonia Vardè. 2016. "Digital Pen Technology and Consecutive Note-taking in the Classroom and beyond." In *Interchange between Languages and Cultures: The Quest for Quality*, ed. by Jitka Zehnalovà, Ondřej Molnár, and Michal Kubánek, 131–150 Olomouc: Palacký University Olomouc.

Bower, Kathryn. 2015. "Stress and Burnout in Video Relay Service (VRS) Interpreting." *Journal of Interpretation* 24 (1): 2. http://digitalcommons.unf.edu/joi/vol24/iss1/2.

Bowker, Lynne, and Gloria Corpas Pastor. 2015. "Translation Technology." In *Handbook of Computational Linguistics*, ed. by Ruslan Mitkov. Oxford: Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199573691.001.0001/oxfordhb-9780199573691-e-007 (Accessed January 3, 2020).

Braun, Sabine. 2013. “Keep your Distance? Remote Interpreting in Legal Proceedings: A Critical Assessment of a Growing Practice.” *Interpreting* 15(2): 200–228. https://doi.org/10.1075/intp.15.2.03bra

Braun, Sabine. 2015. “Remote Interpreting.” In *Routledge Encyclopedia of Interpreting Studies*, ed. by Franz Pöchhacker, 346–348. New York: Routledge.

Braun, Sabine, and Judith L. Taylor. 2012. *Videoconference and Remote Interpreting in Legal Proceedings*. Cambridge: Intersentia.

Braun, Sabine. 2018. “Video-Mediated Interpreting in Legal Settings in England: Interpreters’ Perceptions in Their Sociopolitical Context.” *Translation and Interpreting Studies* 13 (3): 393–420. https://doi.org/10.1075/tis.00022.bra

Cadwell, Patrick, Sharon O’Brien, and Carlos S. C. Teixeira. 2018. Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, 26:3, 301–321. https://doi.org/10.1080/0907676X.2017.1337210

Castilho, Sheila, Natália Resende, and Ruslan Mitkov. 2019. “What Influences the Features of Post-Editese? A Preliminary Study.” In *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, ed. by Irina Temnikova, Constantin Orasan, Gloria Corpas Pastor, and Ruslan Mitkov. https://doi.org/10.26615/issn.2683-0078.2019_003

Castilho, Sheila, Federico Gaspari, Joss Moorkens, Maja Popović, and Antonio Toral. 2019. (eds). *Machine Translation*, 33 (1–2). Special Issue on Human Factors in Neural Machine Translation.

Chen, Sijia. 2017. “Note-taking in Consecutive Interpreting: New Data from Pen Recording.” *Translation & Interpreting* 9(1): 4–23. https://doi.org/10.12807/ti.109201.2017.a02

Chen, Sijia. 2018a. “A Pen-eye-voice Approach towards the Process of Note-taking and Consecutive Interpreting: An Experimental Design.” *International Journal of Comparative Literature and Translation Studies* 6(2): 1–8. https://doi.org/10.7575/aiac.ijclts.v.6n.2p.1

Chen, Sijia. 2018b. “Exploring the Process of Note-taking and Consecutive Interpreting: a Pen-eye-voice Approach towards Cognitive Load.” *The Interpreter and Translator Trainer*, 12(4): 467–468. https://doi.org/10.1080/1750399X.2018.1535163

Corpas Pastor, Gloria. 2001. “Compilación de un corpus *ad hoc* para la enseñanza de la traducción inversa especializada.” *Trans. Revista de traductología*, 5, 155–184.

Corpas Pastor, Gloria. 2018. “Tools for Interpreters: the Challenges that Lie Ahead.” *Current Trends in Translation Teaching and Learning E*, 5: 157–182. ISSN: 2342-7205.

Corpas Pastor, Gloria, and Lily May Fern. 2016. *A Survey of Interpreters’ Needs and Practices Related to Language Technology*. Technical paper [FFI2012-38881-MINECO/TI-DT-2016-1].

Corpas Pastor, Gloria, and Miriam Seghiri Domínguez. (editors). 2016. *Corpus-based Approaches to Translation and Interpreting: From Theory to Applications*. Frankfurt: Peter Lang. ISBN 9783631609569 / E-ISBN 9783653060553. https://doi.org/10.3726/9783653060553

Corpas Pastor, Gloria, and Isabel Durán Muñoz, eds. 2018. Trends in E-Tools and Resources for Translators and Interpreters. *Approaches to Translation Studies* 45. Leiden:Brill Rodopi.

Corpas Pastor, Gloria, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. 2008. Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: October 21–25, 2008, Waikiki, Hawaii, USA*, 75–81.

Costa, Hernani, Gloria Corpas Pastor, and Isabel Durán Muñoz. 2014a. “A Comparative User Evaluation of Terminology Management Tools for Interpreters.” In *Proceedings of the 4th International Workshop on Computational Terminology*, 68–76. https://doi.org/10.3115/v1/W14-4809

Costa, Hernani, Gloria Corpas Pastor, and Isabel Durán Muñoz. 2014b. "Technology assisted Interpreting," *MultiLingual 143*, 25(3): 27–32. https://eden.dei.uc.pt/~hpcosta/docs/papers/201404-MultiLingual.pdf. (Accessed January 3, 2020).

Costa, Hernani, Gloria Corpas Pastor, and Isabel Durán Muñoz. 2017. "Assessing Terminology Management Systems for Interpreters." In *Trends in E-tools and Resources for Translators and Interpreters*, ed. by Gloria Corpas Pastor and Isabel Durán Muñoz, 7–84. Leiden: Brill Rodopi. https://doi.org/10.1163/9789004351790

Daems, Joke, Orphée De Clercq, and Lieve Macken. 2017. "Translationese and Post-editese: How Comparable Is Comparable Quality?" *Linguistica Antverpiensia, New Series: Themes in Translation Studies* 16: 89–103.

Drechsel, Alexander. 2013b. *The Tablet Interpreter Manual*. https://static1.squarespace.com/static/52d4015ce4b0eab6f2d76b6f/t/594b8b7a414fb54310f5957d/1498123132497/The+Tablet+Interpreter+Manual.pdf. (Accessed January 4, 2020)

Drechsel, Alexander, and Joshua Goldsmith. 2016. "Tablet Interpreting: The Evolution and Uses of Mobile Devices in Interpreting." In *Proceedings of the 2016 CIUTI Forum*, edited by Hannelore Lee-Jahnke. Bern: Peter Lang.

EAGLES (Expert Advisory Group on Language Engineering Standards). 1996. *Text corpora Working Group reading Guide. EAGLES Document EAG-TCWG-FR-2*, version of May 1996. http://www.ilc.cnr.it/EAGLES/corpintr/corpintr.html. (Accessed January 3, 2020).

Esperança-Rodier, Emmanuelle, and Caroline Rossi. 2019. "Time is Everything: A Comparative Study of Human Evaluation of SMT vs. NMT." In *Proceedings of the 41st Conference Translating and the Computer, ASLING, London, UK, November 21–22, 2019*, ed. by João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov, and Olaf-Michael Stefanov, 36–46. Geneva: Tradulex. ISBN: 978-2970-10957-0.

Fantinuoli, Claudio. 2016. "InterpretBank. Redefining Computer-assisted Interpreting Tools." In *Proceedings of the 38th Conference Translating and the Computer, ASLING, London, UK, November 17–18, 2016*, ed. by João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov, and Olaf-Michael Stefanov, 42–52. Geneva: Tradulex. ISBN: 978-2-9701095-0-1.

Fantinuoli, Claudio. 2017. "Computer-assisted Preparation in Conference Interpreting." *The International Journal for Translation and Interpreting Research* 9(2): 24–37. https://doi.org/10.12807/ti.109202.2017.a02

Fantinuoli, Claudio. 2018a. Computer-assisted Interpreting: Challenges and Future Perspectives. In *Trends in e-tools and resources for translators and interpreters*, ed. by Gloria Corpas Pastor and Isabel Durán Muñoz, 153–174. Leiden: Brill.

Fantinuoli, Claudio. 2018b. Interpreting and Technology: The Upcoming Technological Turn. In. *Interpreting and Technology*, ed. by Claudio Fantinuoli, 1–2- Berlin: Language Science Press.

Fantinuoli, Claudio. 2018c. "The Use of Comparable Corpora in Interpreting Practice and Training". *The Interpreters' Newsletter* 23.

Fantinuoli, Claudio and Zanettin, Federico (eds.). 2015. *New Directions in Corpus-based Translation Studies* (Translation and Multilingual Natural Language Processing 1), 133–149. Berlin: Language Science Press. https://doi.org/10.26530/OAPEN_559833

Ferraresi, Adriano. 2016. "Intermodal Corpora and the Translation Classroom: What can Translation Trainers and Trainees Learn from Interpreting?" *Linguaculture* 2016 (2): 27–51. https://doi.org/10.1515/lincu-2016-0011. (Accessed January 3, 2020).

Fumagalli, Daniela. 1999. Alla ricerca dell'interpretese. Uno studio sull'interpretazione consecutiva attraverso la corpus linguistics. Unpublished PhD thesis. Advanced School for Translators and Interpreters (SSLMIT), University of Trieste.

Gellerstam, Martin. 1986. "Translationese in Swedish Novels Translated from English." *Translation studies in Scandinavia* 1: 88–95.

Gile, Daniel. 2004. "Translation Research versus Interpreting Research". In *Translation Research and Interpreting Research*, ed. by Christina Schäffner, 10–34. Clevedon: Multilingual Matters. https://doi.org/10.21832/9781853597350-003

Girletti, Sabrina, Pierrette Bouillon, Martina Bellodi, and Philipp Ursprung. 2019. "Preferences of End-users for Raw and Post-edited NMT in a Business Environment." In *Proceedings of the 41st Conference Translating and the Computer, ASLING, London, UK, November 21–22, 2019*, ed. by João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov, and Olaf-Michael Stefanov, 47–59. Geneva: Tradulex. ISBN: 978-2970-10957-0.

Goldsmith, Joshua. 2018. "Tablet interpreting." *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association* 13(3): 342–365. https://doi.org/10.1075/tis.00020.gol

Goldsmith, Joshua, and Josephine Holley. 2015. Consecutive Interpreting 2.0: The Tablet Interpreting Experience. Unpublished MA thesis. University of Geneva.

Gough, Joanna, and Katerina Perdikaki. 2018. "Concurrent Translation-Reality or Hype?" In *Proceedings of the 40th Conference Translating and the Computer, ASLING, London, UK, November 15–16, 2018*, ed. by João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov, and Olaf-Michael Stefanov, 79–88. Geneva: Tradulex. ISBN: 978-2-9701095-5-6.

Gracia-García, Roberto A. 2002. *Telephone Interpreting: A Review of Pros and Cons*. In *Proceedings of the ATA 43rd Annual Conference, Atlanta, Georgia, November 6–9, 2002*, 195–216. Alexandria, VA: American Translators Association.

Guerberof Arenas, Ana. 2013. "What Do Professional Translators Think About Post-editing?" *JoSTrans: The Journal of Specialised Translation*, 19: 75–95.

Guerberof Arenas, Ana, and Joss Moorkens. 2019. Machine Translation and Post-editing Training as Part of a Master's Programme. *Jostrans: The Journal of Specialised Translation*, 31: 217–238.

Hiebl, Bettina. 2011. Simultanes Konsekutivdolmetschen mit dem Livescribe^{TM} Echo^{TM} Smartpen: Ein Experiment im Sprachenpaar Italienisch-Deutsch mit Fokus auf Zuhörerbewertung. MA thesis. University of Vienna. https://doi.org/10.25365/thesis.14608

Hof, Michelle R. 2012. "iPad: The Ideal Boothmate?" *Aiic.net*, November 25. http://aiic.net/p/6354. (Accessed January 3, 2020).

Hu, Kaibao. 2016. *Introducing Corpus-based Translation Studies* (New Frontiers in Translation Studies). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48218-6

Ilisei, Iustina, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. "Identification of Translationese: A Machine Learning Approach." In *Computational Linguistics and Intelligent Text Processing*, 503–511. Berlin and Heidelberg: Springer. https://doi.org/10.1007/978-3-642-12116-6_43

Ippolito, Margherita. 2013. *Simplification, Explicitation and Normalization: Corpus-Based Research into English to Italian Translations of Children's Classics*. Newcastle upon Tyne, England: Cambridge Scholars Publishing.

ISO 18578:2017. *Translation Services – Post-editing of Machine Translation Output – Requirements*. https://www.iso.org/standard/62970.html. (Accessed January 3, 2020).

Jiménez-Crespo, Miguel Ángel. 2017. *Crowdsourcing and Online Collaborative Translations*. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.131

Jiménez-Crespo, Miguel Ángel. 2018. "Crowdsourcing and Translation Quality: Novel Approaches in the Language Industry and Translation Studies." In *Translation Quality Assessment: From Principles to Practice*, ed. by Sheila Castillo, Joss Moorkens, Federico Gaspari, and Stephen Doherty, 69–93. Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7_4

Jiménez-Crespo, Miguel Ángel. 2019. "Technology and non-professional translation." In *The Routledge Handbook of Translation and Technology*, ed. by Minako O'Hagan, 220–238. London: Routledge. https://doi.org/10.4324/9781315311258-14

Kelly, Nataly. 2008. *Telephone Interpreting: A Comprehensive Guide to the Profession*. Bloomington: Trafford Publishing.

Kenny, Dorothy. 2018. "Sustaining Disruption?: On the Transition from Statistical to Neural Machine Translation." *Tradumàtica*, 16: 0059–70. https://doi.org/10.5565/rev/tradumatica.221

Kenny, Dorothy. 2019. "Machine Translation." In *The Routledge Handbook of Translation and Philosophy*, ed. by J. Piers Rawling and Philip Wilson, 428–445. London and New York, NY: Routledge.

Koskinen, Kaisa, and Minna Ruokonen. 2017. "Love Letters or Hate Mail? Translators' Technology Acceptance in the Light of Their Emotional Narratives." In *Human Issues in Translation Technology*, ed. by Dorothy Kenny, 8–24. London: Routledge.

Kostal, Nina. 2011. Die Rolle der Notizentechnik beim Konsekutivdolmetschen: Analyse mittels Livescribe^{TM} Echo^{TM} Smartpen. MA thesis. University of Vienna. https://doi.org/10.25365/thesis.16658

Leblanc, Matthieu. 2013. "Translators on Translation Memory (TM): Results of an Ethnographic Study in Three Translation Services and Agencies." *Translation & Interpreting* 5(2): 1–13. ti.105202.2013.a01. https://doi.org/10.12807/ti.105202.2013.a01

Leblanc, Matthieu. 2017. "'I can't get no satisfaction!' Should we blame translation technologies or shifting business practices?" In *Human Issues in Translation Technology*, ed. by Dorothy Kenny, 63–80. London: Routledge.

Lion, Casey K., Julie C. Brown, B. E. Ebel, Eileen J. Klein, Bonnie Strelitz, Kolleen Kays Gutman, Patty Hencz, Juan Fernandez, and Rita Mangione-Smith. 2015. "Effect of Telephone vs Video Interpretation on Parent Comprehension, Communication, and Utilization in the Pediatric Emergency Department: a Randomized Clinical Trial." *JAMA Pediatrics* 169(12): 1117–1125. https://doi.org/10.1001/jamapediatrics.2015.2630

Locatis, Craig, Deborah Williamsom, Carrie Gould-Kabler, Laurie Zone-Smith, Isabel Detzler, Jason Roberson, Richard Maisiak, and Michael Ackerman. 2010. "Comparing In-person, Video, and Telephonic Medical Interpretation. *Journal of General Internal Medicine* 25 (4), 345–350. https://doi.org/10.1007/s11606-009-1236-x

Malamatidou, Sofia. 2019. *Corpus Triangulation: Combining Data and Methods in Corpus-Based Translation Studies* (Routledge Studies in Empirical Translation). London: Routledge.

Marcus, Jessica. 2017. "Quality Improvement Project Examining Nurses' Perceptions Regarding the Use of Technology for Interpretation for Patients with Limited English Proficiency." *Doctor of Nursing Practice (DNP) Translational and Clinical Research Projects* 26. https://kb.gcsu.edu/dnp/26. (Accessed January 3, 2020).

Mikhailov, Mikhail and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Studies: A guide for research*. London: Routledge. https://doi.org/10.4324/9781315624570

Mikkelson, Holly. 2003. "Telephone Interpreting: Boon or Bane?" In *Speaking in Tongues: Language across Contexts and Users*, ed. by Luis Pérez González, 251–269. Valencia: Universitat de València.

Moser-Mercer, Barbara. 2003. "Remote Interpreting: Assessment of Human Factors and Performance Parameters." In *Joint Project International Telecommunication Union (ITU)-Ecole de Traduction et d'Interpretation, Université de Genève (ETI)*. https://ecfsapi.fcc.gov/file/7521826425.pdf. (Accessed January 3, 2020).

Mouzourakis, Panayotis. 1996. "Videoconferencing: Techniques and challenges." *Interpreting* 1(1): 21–38. https://doi.org/10.1075/intp.1.1.03mou

Mouzourakis, Panayotis. 2006. "Remote Interpreting: a Technical Perspective on Recent Experiments." *Interpreting* 8(1): 45–66. https://doi.org/10.1075/intp.8.1.04mou

Mouzourakis, Takis. 2003. "That Feeling of Being There: Vision and Presence in Remote Interpreting." *The AIIC Webzine* 23. http://aiic.net/page/print/1173. (Accessed January 3, 2020).

Mutal, Jonathan, Lise Volkart, Pierrette Bouillon, Sabrina Girletti, and Paula Estrella. 2019. "Differences between SMT and NMT Output: A Translators' Point of View." In *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, ed. by Irina Temnikova, Constantin Orasan, Gloria Corpas Pastor, and Ruslan Mitkov, 19–27. https://doi.org/10.26615/issn.2683-0078.2019_001

Naimushin, Boris. 2019. "Interviews with Translators and Interpreters." *Russian Journal of Linguistics*, 23(2): 584–590.

NAJIT. 2009. "Telephone interpreting in legal settings." http://www.najit.org/publications/Telephone%20Interpreting.pdf. (Accessed January 3, 2020).

Napier, Jemina, Robert Skinner, and Graham H. Turner. 2017. "'It's Good for them but not so for me': Inside the Sign Language Interpreting Call Centre." *Translation & Interpreting* 9(2): 1–23. https://doi.org/10.12807/ti.109202.2017.a01

Nash Fernandez, Annalisa. 2017. "Requisite or Redundant: Spanish Language Interpreters in Urban Medical Systems." *Translation Journal*. https://translationjournal.net/January-2017/requisite-or-redundant-spanish-language-interpreters-in-urban-medical-systems.html. (Accessed January 31, 2020).

Navarro-Hall, Esther. 2012. "An introduction to sim-consec." http://1culture.net/1culture/anintroduction-to-sim-consec. (Accessed April 30, 2020).

O'Brien, Sharon. 2012. "Translation as Human–Computer Interaction." *Translation Spaces* 1(1): 101–122. https://doi.org/10.1075/ts.1.05obr

Olohan, Maeve. 2017. "Technology, translation and society." *Target. International Journal of Translation Studies* 29(2): 264–283. https://doi.org/10.1075/target.29.2.04olo

Olohan, Maeve. 2019. "Sociological Approaches to Translation Technology." In *The Routledge Handbook of Translation and Technology*, ed. by Minakoed O'Hagan, 384–397. London: Routledge. https://doi.org/10.4324/9781315311258-23

Orlando, Marc. 2010. "Digital Pen Technology and Consecutive Interpreting: Another Dimension in Notetaking Training and Assessment." *The Interpreters' Newsletter* 15: 71–86.

Orlando, Marc. 2011. "Beyond Pen and Paper: Note-taking Training and Digital Technology." In *Proceedings of the "Synergise!" Biennial National Conference of the Australian Institute of Interpreters and Translators: AUSIT 2010*, ed. by Annamaria Arnall and Uldins Ozolins, 76–85. Newcastle upon Tyne: Cambridge Scholars Publishing.

Orlando, Marc. 2013. "Interpreting Training and Digital Pen Technology." http://aiic.net/p/6484. (Accessed January 3, 2020).

Orlando, Marc. 2014. "A Study on the Amenability of Digital Pen Technology in a Hybrid Mode of Interpreting: Consec-simul with Notes." *Translation & Interpreting* 6(2): 39–54.

Orlando, Marc. 2015a. "Digital Pen Technology and Interpreter Training, Practice and Research: Status and Trends." In *Interpreter Education in the Digital Age*, ed. by Suzanne Ehrlich and Jemina Napier, 125–152. Washington DC: Gallaudet University Press.

Orlando, Marc. 2015b. "Implementing Digital Pen Technology in the Consecutive Interpreting Classroom." In *To Know How to Suggest… Approaches to Teaching Conference Interpreting*, ed. by Dörte Andres and Martina Behr, 171–200. Berlin: Frank & Timme.

Orlando, Marc. 2016. *Training 21st Century Translators and Interpreters : at the Crossroads of Practice, Research and Pedagogy*. Berlin: Frank & Timme.

Ozolins, Uldins. 2012. "Telephone Interpreting: Understanding Practice and Identifying Research Needs." *Translation & Interpreting* 3(2): 33–47.

Pielmeier, Hélène and O'Mara, Paul. 2020. *The State of the Linguist Supply Chain Translators and Interpreters in 2020*. CSA Research. https://insights.csa-research.com/reportaction/305013106/Toc (Accessed January 30, 2020).

Pérez-Pérez, Pablo. 2018. "The Use of a Corpus Management Tool for the Preparation of Interpreting Assignments: A Case Study." *The International Journal of Translation and Interpreting Research* 10 (1): 137–151. https://doi.org/10.12807/ti.110201.2018.a08

Plaza Lara, Cristina. 2019. "Análisis DAFO sobre la inclusión de la traducción automática y la posedición en los másteres de la red EMT." *JosTrans: The Journal of Specialised Translation* 31: 260–280.

Porlán Moreno, R. 2019. The Use of Portable Interpreting Devices: An Overview. *Tradumàtica* 17: 45–58.

Prandi, Bianca. 2018. "An Exploratory Study on CAI Tools in Simultaneous Interpreting: Theoretical Framework and Stimulus Validation." In *Interpreting and Technology*, ed. by Claudio Fantinuoli, 29–60. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.1493281

Price, Erika Leemann, Eliseo J. Pérez-Stable, Dana Nickleach, Mónica López, and Leah S. Karliner. 2012. "Interpreter Perspectives of In-person, Telephonic, and Videoconferencing Medical Interpretation in Clinical Encounters. *Patient Education and Counseling* 87(2): 226–232. https://doi.org/10.1016/j.pec.2011.08.006

Pym, Anthony. 2011. "What Technology does to Translating." *Translation & Interpreting* 3(1): 1–9. http://www.trans-int.org/index.php/transint/article/view/121/81. (Accessed January 3, 2020).

Rabinovich, Ella and Wintner, Shuly. 2015. Unsupervised Identification of Translationese. *Transactions of the Association for Computational Linguistics* 3: 419–432. https://doi.org/10.1162/tacl_a_00148

Ruffo, Paola. 2018. "Human-Computer Interaction in Translation: Literary Translators on Technology and Their Roles." In *Proceedings of the 40th Conference Translating and the Computer, ASLING, London, UK, November 15–16, 2018*, ed. by João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov, and Olaf-Michael Stefanov, 127–131. Geneva: Tradulex. ISBN: 978-2-9701095-5-6.

Ruffo, Paola. 2019. "'I Wish They Could See the Magic': Literary Translators on Their Roles and Technology." In *EST Congress 2019: Book of Abstracts*. http://www.est2019.com/wp-content/uploads/2019/09/EST-2019-ABSTRACT-BOOK-1.pdf (Accessed January 3, 2020).

Ruiz Mezcua, Aurora (editor). 2018. *Approaches to Telephone Interpretation: Research, Innovation, Teaching, and Transference*. Bern: Peter Lang. https://doi.org/10.3726/b13326

Russo, Mariachiara, Claudio Bendazzoli, and Bart Defrancq (eds.). 2018. *Making Way in Corpus-based Interpreting Studies*. London: Springer. https://doi.org/10.1007/978-981-10-6199-8

Rütten, Anja. 2017. "Terminology Management Tools for Conference Interpreters: Current Tools and How They Address the Specific Needs of Interpreters." In *Proceedings of the 39th Conference Translating and the Computer, ASLING, London, UK, November 16–17, 2017*, ed. by João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov, and Olaf-Michael Stefanov, 98–103. Geneva: Tradulex. ISBN: 9782970109532.

Seeber, Kilian G., Laura Keller, Rhona Amos, and Sophie Hengl. 2019. "Expectations vs. Experience: Attitudes towards Video Remote Conference Interpreting." *Interpreting* 21(2): 270–304. https://doi.org/10.1075/intp.00030.see

Sgourou, Maria. 2019. "The Four Stages of Machine Translation Acceptance in a Freelancer's Life." In *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, ed. by Irina Temnikova, Constantin Orasan, Gloria Corpas Pastor, and Ruslan Mitkov, 134–135. https://doi.org/10.26615/issn.2683-0078.2019_001

Shlesinger, Miriam. 1998. "Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies." *Meta* 43(4): 486–493. https://doi.org/10.7202/004136ar

Shlesinger, Miriam. 2008. "Towards a Definition of Interpretese." In *Efforts and Models in Interpreting and Translation Research*, ed. by Gyde Hansen, Andrew Chesterman, and Heidrun Gerzymisch-Arbogast, 237–253. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.80.18shl

Shlesinger, Miriam, and Noam Ordan. 2012. "More Spoken or More Translated? Exploring a Known Unknown of Simultaneous Interpreting." *Target* 24(1): 43–60. https://doi.org/10.1075/target.24.1.04shl

Skinner, Robert, Jemina Napier, and Sabine Braun. 2018. "Interpreting via Video Link: Mapping of the Field." In *Here or there: Research on Interpreting via Video Link*, ed. by Jemina Napier, Robert Skinner, and Sabine Braun, 11–35. Washington DC: Gallaudet University Press.

Steurs, Frida. 2016. "The Translator in a New Era: Towards Collaborative Translation and New Tools." In *TETRA Conference, Date: 2016/09/30-2016/09/30, Location: Bologna, Italy*: http://wwwling.arts.kuleuven.be/qlvl/prints/Steurs_2016pres_translator_new_era.pdf. (Accessed January 3, 2020).

Straniero Sergio, Francesco, and Caterina Falbo (editors). 2012. *Breaking Ground in Corpus-based Interpreting Studies*. Bern/New York: Peter Lang. https://doi.org/10.3726/978-3-0351-0377-9

Temnikova, Irina, Constantin Orasan, Gloria Corpas Pastor, and Ruslan Mitkov (editors). 2019. *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*. https://doi.org/10.26615/issn.2683-0078.2019_001

Thompson, Paul. 2005. "Spoken Language Corpora." In *Developing Linguistic Corpora: a Guide to Good Practice*, ed. by Martin Wynne, 59–70. Oxford: Oxbow Books.

Toral, Antonio. 2019. "Post-editese: An Exacerbated Translationese." In *Proceedings of Machine Translation Summit XVII, Volume 1: Research Track*. https://www.aclweb.org/anthology/W19-66.pdf. (Accessed January 3, 2020).

Toral, Antonio, and Andy Way. 2014. "Is Machine Translation Ready for Literature?" In *Conference Chairs and Editors of the Proceedings* (p. 174).

UNHCR. 2016. *Connecting Refugees: How Internet and Mobile Connectivity can Improve Refugee Well-Being and Transform Humanitarian Action*. Geneva: UNHCR.

Veale, Tony, and Andy Way. 1997. "Gaijin: A Bootstrapping, Template-Driven Approach to Example-Based Machine Translation," in *International Conference, Recent Advances in Natural Language Processing*, 239–244.

Vieira, Lucas Nunes. 2018. "Automation anxiety and translators". *Translation Studies*, 1–21. https://doi.org/10.1080/14781700.2018.1543613

Vieira, Lucas Nunes. 2019. "Post-editing of Machine Translation." In *The Routledge Handbook of Translation and Technology*, ed. by Minako O'Hagan, 319–337. London: Routledge. https://doi.org/10.4324/9781315311258-19

Volkart, Lise, Pierrette Bouillon, and Sabrina Girletti. 2018. "Statistical vs. Neural Machine Translation: A Comparison of MTH and DeepL at Swiss Post's Language Service." In *Proceedings of the 40th Conference Translating and the Computer, ASLING, London, UK, November 15–16, 2018*, ed. by João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov, and Olaf-Michael Stefanov, 145–150. Geneva: Tradulex. ISBN: 978-2-9701095-5-6.

Wadensjö, Cecilia. 1999. "Telephone Interpreting & the Synchronization of Talk in Social Interaction." *The Translator* 5(2): 247–264. https://doi.org/10.1080/13556509.1999.10799043

Wang, Jihong. 2018a. "'It keeps me on my Toes': Interpreters' Perceptions of Challenges in Telephone Interpreting and Their Coping Strategies." *Target. International Journal of Translation Studies* 30(3): 430–462. https://doi.org/10.1075/target.17012.wan

Wang, Jihong. 2018b. "'Telephone Interpreting Should Be Used Only as a Last Resort.' Interpreters' Perceptions of the Suitability, Remuneration and Quality of Telephone Interpreting." *Perspectives* 26(1): 100–116. https://doi.org/10.1080/0907676X.2017.1321025

Wasik, Zosia. 2017. "Migrant Crisis Triggers a Wave of Tech Innovation." *Financial Times*, October 26. https://www.ft.com/content/e53197ee-8904-11e7-afd2-74b8ecd34d3b. (Accessed January 4, 2020).

Wessling, Dawn M., and Shaw Sherry. 2014. "Persistent Emotional Extremes and Video Relay Service Interpreters." *Journal of Interpretation* 23(1), article 6.

Wladyka-Leittretter, Anna Maria. 2018. "Automating Terminology Management. Discussion of IATE and Suggestions for Enhancing its Features." In *Proceedings of the 40th Conference Translating and the Computer, ASLING, London, UK, November 15–16, 2018*, ed. by João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov, and Olaf-Michael Stefanov, 151–160. Geneva: Tradulex. ISBN: 978-2-9701095-5-6.

Xu, Ran. 2018. Corpus-based terminological preparation for simultaneous interpreting. *Interpreting Research*. 20 (1). 29–58. https://doi.org/10.1075/intp.00002.xu

Zang, Jinyi, and Matsumoto, Tadahiro. 2019. Corpus Augmentation for Neural Machine Translation with Chinese-Japanese Parallel Corpora. *Applied Sciences*, 9. 2036. https://doi.org/10.3390/app9102036

Zaretskaya, Anna, Gloria Corpas Pastor, and Miriam Seghiri Domínguez. 2015. "Translators' Requirements for Translation Technologies: a User Survey." In *New Horizons in Translation and Interpreting Studies (Full papers)*, ed. by Gloria Corpas Pastor, Miriam Seghiri, Rut Gutiérrez Florido, and Miriam Urbano Mendaña. Geneva: Tradulex, 247–254.

Zaretskaya, Anna, Gloria Corpas Pastor, and Miriam Seghiri Domínguez. 2018. "User Perspective on Translation Tools: Findings of a User Survey." In *Trends in E-tools and Resources for Translators and Interpreters*, ed. by Gloria Corpas Pastor and Isabel Durán Muñoz, 37–56. Leiden: Brill Rodopi.

Zhang, Mike and Toral, Antonio. 2019. The Effect of Translationese in Machine Translation Test Sets. In *Proceedings of the Fourth Conference on Machine Translation* (Volume 1: Research Papers), 73–81. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5208

ZHEN

A directional parallel corpus of Chinese source texts and English translations

Yi Gu and Ana Frankenberg-Garcia
Centre for Translation Studies, University of Surrey

Most Chinese-English parallel corpora consist of English source texts translated into Chinese. This chapter discusses the need for corpora representative of the under-resourced Chinese into English translation direction. After a brief overview of the current Chinese-English translation scenario and an analysis of existing parallel corpora for this language pair, we discuss problems in mining contemporary Chinese to English translations and issues in Chinese to English parallel text alignment. We then introduce ZHEN, a corpus of circa one-million characters of contemporary simplified Chinese source texts from a range of text types aligned with authentic translations into English. Its aim is to contribute to our understanding of Chinese to English translation norms and of features of English translated from Chinese.

Keywords: parallel corpora, parallel text alignment, Chinese-English translation, ZHEN corpus

1. Introduction

The aim of this chapter is to present the ZHEN corpus, a unidirectional parallel corpus of contemporary Chinese (ZH) source texts aligned with authentic translations into English (EN). The chapter begins with an overview of the current Chinese to English translation scenario and the growing market for translations from Chinese into English. Next, we review existing Chinese-English parallel corpora, demonstrating that while there are a number of tools for the study of English translated into Chinese, there is a dearth of corpus resources that cater specifically for the study of Chinese translated into English. This is followed by an introduction to ZHEN, discussing a number of non-trivial issues in mining contemporary Chinese source texts and their translations into English, and in Chinese-English parallel text

https://doi.org/10.1075/btl.158.02gu

alignment. We then discuss how ZHEN can assist translation scholars, professional translators, translation trainers and trainees, and machine-translation researchers come to a more systematic understanding of the specificities of the Chinese to English translation direction.

2. A brief overview of the Chinese-English translation scenario

According to the Translators Association of China (TAC),[1] before the 1970s the translation industry in China was practically non-existent, with most translation activities being carried out within government departments and state-owned companies with a permit to translate. Chinese political essays, government documents and literature – aimed at portraying the People's Republic of China to the outside world – were translated into foreign languages by Chinese translators and edited by foreign experts with close links to the government (Yang 1999). This practice is still widely seen in China today (Translators Association of China 2019).

The first commercial translation services in China began to emerge in the 1980s, when China started to adopt the policy of opening up to the outside world. The need for translating non-literary and non-political texts out of Chinese has increased dramatically in the last few years. According to TAC, by 2002 there were more than 800 registered translation companies in Beijing alone, and by 2014, 64% of the 120 Chinese translation companies surveyed declared that more than half of their total business involved translating from Chinese into English. In 13% of the companies, Chinese to English translation accounted for 80% to 100% of their business.

TAC posits that the recent increase in demand for translation has led to a severe shortage of qualified translators, especially translators capable of translating out of Chinese and into a second language (L2). Although in European countries like France and the UK professional translation services are normally carried out into the translator's first language (L1), in China there is a long history of translating into L2, also known as inverse translation. Indeed, few people outside China have sufficient knowledge of Mandarin to be able to translate the language, and translation into L2 is so ingrained in China that, according to Wang (2011), there is no specific term to describe it. The closest equivalents that can be found in Chinese are 译入/*yiru* [inward translation], and 译出/*yichu* [outward translation],[2]

[1] http://tac-online.org.cn/en/ [accessed on 15/01/2020]

[2] Throughout this chapter, Chinese characters are followed by slash and Pinyin transcription, and a literal translation into English is provided in square brackets.

referring respectively to the translation from a foreign language into Chinese and from Chinese into a foreign language. The emphasis is thus placed on the contrast between Chinese and other languages rather than on the translator's L1 and L2, and both 译入/yiru and 译出/yichu are used to describe work carried out by Chinese L1 translators today (Hu 2006).

Despite the normality of L2 translation practice in China, little is known about it beyond common criticisms, such as those pointed out by Lao (1996: 41): translators lack qualifications; translations are often carried out without theoretical guidance; and translated texts often show features of *translationese*, i.e., an inappropriately literal rendition which clashes with accepted target language norms. When the target language is English, *translationese* is sometimes referred to as *Chinglish*, i.e., the L2 English used by L1 Chinese speakers. Pinkham (2000) described the phenomenon in the context of second language learning, while Gu (2017) found evidence of Chinglish in interpreting renditions by Chinese students pursuing an MA in Interpreting at a UK university. However, little is known about written translation in this language direction.

In fact, although there have been a number of studies looking at English translated into Chinese (for example, McEnery, Xiao, & Tono 2006; Xiao 2010), there do not seem to be many descriptive studies that systematically investigate Chinese-to-English translation phenomena.

3. Chinese-English parallel corpora

Although there are a number of excellent monolingual resources available for Chinese Translation Studies – for example, the Lancaster Corpus of Mandarin Chinese and the comparable ZJU corpus of Translational Chinese compiled by Richard Xiao (Xiao, He & Yue 2008) – to this date there is not much in the way of parallel Chinese-English corpora, not to mention resources that specifically address the Chinese to English translation direction. An analysis carried out to determine which parallel corpora containing Chinese and English were available in 2019 has enabled us to identify the six corpora in Table 1.

As can be seen, the largest Chinese-English corpus is the PKU Corpus, which began to be developed by the Institute of Computational Linguistics of Peking University in 2001 (Wang 2004a). Research based on this corpus has indeed contributed to corpus-based contrastive translation studies involving Chinese (for example, Wang 2009). However, the PKU Corpus is not publicly accessible, which makes it harder for researchers outside Peking University, let alone practicing and trainee translators (and language learners) to use. Another difficulty is that there does not seem to be a comprehensive description of the current composition of the

Table 1. Chinese English parallel corpora in 2019

Corpus	Text types	Text dates	Size in tokens	Translation direction	Availability
PKU	80% literary fiction; other sources include political speeches and legal documents	18th to 21st century	3.0 million EN 2.9. million ZH	60% EN-ZH 40% ZH-EN	Restricted
Babel	Magazine articles	2000–2001	2.5 thousand EN (no information on translated tokens)	100% EN-ZH	Restricted
OPUS (Chinese-English corpora)	95% United Nations documents; other sources include TED talks, Bible and Quran translations	Not specified	31.2 million EN & ZH (no information on ST and TT corpora size)	Not specified	Open access
Wang Lixun's English-Chinese Parallel Corpus	Essays, novels and ancient fables	Mostly 19th to 20th century	1.4 m EN source texts 576,724 ZH source texts (no information on translated tokens)	78% EN-ZH 22% ZH-EN	Open access
TRAD Chinese-English News Articles Parallel corpus	Chinese version of Voice of America professionally translated into English	2011–2012	15,000 ZH (no information on translated tokens)	100% ZH-EN	$ 150 (academic) $ 500 (commercial)
The Chinese/English Political Interpreting Corpus (CEPIC)	Transcripts of speeches by political figures from Hong Kong, Beijing, Washington DC and London	Early 21st–2017	2,578,911 ZH (1,072,368 Cantonese and 1,506,541 Mandarin) 3,815,083 EN (no information on ST and TT corpora size)	Not specified	Open access

corpus that is publicly available. According to Wang (2004a), the PKU Corpus is bi-directional, with around 40 percent of its data consisting of Chinese to English translations. However, to our knowledge there is no information on how many bitexts are represented in this translation direction or where they are sourced from. What is known from a more recent publication is that more than 80 percent of the texts within this corpus are literary fiction, with both directions counted (Wang 2009). The only two other genres represented in the corpus are political speeches and legal documents. Besides, speeches from Chinese political leaders consist of more than 90% of the bitexts represented in the Chinese-to-English section of the corpus. In addition, according to Wang (2004b), the publication dates of the texts in PKU range from the 18th to the early 21st century, during which period language use and translation practices are likely to have changed quite substantially. Therefore, despite the rich parallel text data it contains, the PKU corpus seems somewhat limited when it comes to representing contemporary Chinese to English translations.

Also shown in Table 1 is the Babel English-Chinese Parallel Corpus compiled by researchers at Lancaster University. With a total of 544,095 running words sourced from over 300 articles on a wide range of topics (e.g., from *The Foods that Fight Cancer* to *Christmas Lost and Found*) that were published in *World of English* and *Time* magazine in the early 2000s, Babel is a well-structured corpus consisting of contemporary parallel texts. Although it is an excellent resource for comparing and contrasting present-day English and Chinese, because all source texts are in English, it cannot be used to come to a better understanding of issues affecting the translation of Chinese source texts.

Another important parallel corpus containing Chinese and English is the OPUS collection of translated texts from the Web (Tiedemann 2012). OPUS is an open source repository of parallel texts from a wide range of sources. The Chinese-English sources are mostly from the United Nations Parallel Corpus, with circa 650 million tokens (Ziemsky et al. 2016) and the MultiUN corpus (Eisele and Chen 2010), with around 370 million tokens also from the United Nations Website. Together, these corpora amount to over 95% of the Chinese-English section of OPUS, although the extent to which these corpora overlap is not clear. Another problem is that the UN source language documents do not specify the source language. Therefore, it is not a valid reference for studies looking specifically at directional translation phenomena for Chinese into English translation. Other Chinese-English parallel text sources in OPUS include *Bible* and *Quran* translations, which are not useful for studies that depend on direct Chinese to English contemporary language translation data. Likewise, the parallel data available from TED talks – which are mostly delivered in English and translated by volunteers into other languages – is not appropriate for the study of contemporary Chinese translated into English. Thus although the

OPUS collection is no doubt extremely valuable for developing English-Chinese tools and resources, it is not ideal for researching directional Chinese to English translation phenomena.

In addition, Dr Wang Lixun compiled a bi-directional English and Chinese corpus at the Hong Kong Institute of Education. The English-to-Chinese part consists of famous English novels and their translations (for example *Alice in Wonderland* by Lewis Carroll and *A Tale of Two Cities* by Charles Dickens), legal documents, essays, fairy tales and a speech by Martin Luther King. The Chinese-to-English part has four modern novels, one collection of essays and several fables. However, the Chinese-to-English part amounts to just 22% of the corpus (just under 600 thousand Chinese tokens), and most of the texts are from the 19th to early 20th century (the fables are from even earlier, but with no specific publication dates), so unfortunately this corpus is not ideal for investigating present-day Chinese to English translation.

Another parallel corpus is the TRAD Chinese-English News Articles parallel corpus available from ELRA, which consists of texts from the Chinese version of *Voice of America* professionally translated into English for MT evaluation purposes. Although they may serve MT training data well, the *Voice of America* articles are very Anglo-centred, and do not cover typically Chinese topics that could entail realistic translation challenges, such as addressing the translation of culture-specific items and discoursal features of Chinese. Moreover, the English translations were specifically commissioned for the compilation of the corpus. Thus, although the source texts were professionally translated, they cannot be said to be representative of translations carried out for authentic communicative purposes.

A more recent and large-scale parallel corpus in Table 1 is the Chinese/English Political Interpreting Corpus (CEPIC) compiled at the Hong Kong Baptist University. With about 6.5 million tokens in all, it is designed for investigating political interpreting and translation between Mandarin, Cantonese and English. According to Pan & Wong (2018), CEPIC consists of transcripts of speeches delivered by political figures from Hong Kong, Beijing, Washington DC and London, as well as their corresponding translated/interpreted texts. In terms of speech types, CEPIC includes the reading of government reports such as policy addresses and budget speeches, Q&A at press conferences, parliamentary debates, as well as remarks delivered at bilateral meetings. Although most of CEPIC is oral language, a subsection of 1.2 million tokens – the Chinese PRC Reports on the Work of the Government (PRCWoG) and Press Conferences of PRC Reports on the Work of the Government (PRCWoGPC) – could be used for investigating written translation from Chinese into English. However, this covers only one genre, and is therefore not representative of current Chinese to English translation practices.

In summary, although there are large amounts of parallel text data for Chinese and English, most of the parallel text data available is in the English to Chinese translation direction, or the exact translation direction is not specified. The resources available in the Chinese to English translation direction are scattered and cannot be said to be representative of current translation practices in China.

The ZHEN corpus, whose compilation is described in the next section, is a new resource that can fill this gap. Its aim is to enable researchers to come to a better understanding of Chinese to English translation, and inform the discipline of Translation Studies with empirical evidence pertaining to an under-researched translation language direction. Additionally, the ZHEN corpus can have several practical applications, including its use as a resource for translator training, bilingual lexicography, English language teaching in China and machine translation training data. Another affordance of ZHEN is to use its translated English section as a comparable corpus that can be contrasted with non-translated English corpora to investigate distinctive features of English that has been translated from Chinese.

4. The ZHEN corpus

This section addresses the challenges encountered and decisions made when compiling ZHEN corpus. Before starting the compilation process, criteria for defining which text types would be included in the corpus were established. It is generally acknowledged that a general language corpus should aim for a wide variety of text types. However, this is not a realistic aim for parallel corpora, due to the simple fact that only a small percentage of texts in the world ever gets to be translated (Frankenberg-Garcia 2009). Moreover, not all genres are translated between certain languages (McEnery, Tono & Xiao 2006: 95), and, as discussed in Frankenberg-Garcia (2009), most parallel text collections are opportunistic in the sense that they are based on whatever texts that are available in translation. Indeed, as Mauranen (2004: 74) observes, "translations are heavily biased towards certain genres, but these biases are rarely symmetrical for any language pair". For example, Frankenberg-Garcia (2009) comments that the amount of film subtitles translated from English into Portuguese is much greater than the number of film subtitles translated from Portuguese into English. The same applies to literary translations, where English is the main source language for Norwegian-English (Johansson 1998), Italian-English (Zanettin 2014), Portuguese-English (Frankenberg-Garcia 2003), and Spanish-English (Izquierdo, Hofland & Reigem 2008). A further imbalance regarding literary translations is the relative status of the sources. For example, Zanettin (2000) noted that while a wide range of literary texts, including popular fiction, are translated from English into Italian, it is mostly more selective, quality literary Italian fiction that tends to be translated into English.

With regard to text pairs available in the Chinese into English translation direction, an official report by the Chinese Translation Bureau describing published book translations from Chinese into English from 1949 to 2009 shows that the most dominant genre is political education books about Marxism, Leninism, Mao Zedong and Deng Xiaoping theories (He 2013), with 3,045 translations falling into this category. The second most translated genre is books introducing the current status of Chinese political and legal systems, of which a total of 2,709 books have been published. These figures mirror the composition of existing Chinese-English corpora seen in the previous section. In contrast, ZHEN is not limited to these text types. As shall be seen in the next section, considerable effort was put into identifying other genres representative of contemporary Chinese-English translation.

4.1 Text selection

Following Biber (1993), the first step in the selection of texts for the corpus was to define the boundaries of the target dataset. The following criteria were used for the selection of texts for the ZHEN:

1. Ensure that only texts originally written in Chinese and then translated into English are used, excluding from the selection any text pairs translated from English to Chinese, indirect translations carried out via a pivot language, and text pairs whose translation direction is not clear.
2. Use only source texts in Simplified Chinese from mainland China, excluding the Traditional Chinese used in Hong Kong, Macau and Taiwan.
3. Limit the selection to texts published after 1990 in order to better represent present-day use of Chinese.
4. Cover a reasonable number of text types, ensuring that the corpus is representative of a variety of current Chinese-English translation scenarios.
5. Use texts that are openly available online and discard electronic files that require optical character recognition, as this would require additional resources not feasible within the scope of this project.

Not surprisingly, it was much easier to find certain text types meeting the above criteria than others. A major problem was that certain bilingual Chinese-English sources were not actual translations that could be aligned for the purposes of compiling a parallel corpus, but rather Chinese original texts paired up with English summaries or texts that had been entirely rewritten for readerships outside China. This was especially true for news from sources like *Xinhua* or *People's Daily*, where the English versions of online news are not translations of the Chinese versions, making it extremely difficult to mine bitexts that can be fully aligned.

Despite the difficulty of finding true bitexts for Chinese translated into English, it was nevertheless possible to identify the following sources that could be used for mining texts for ZHEN:

a. Government reports (GOV)
b. White papers (WHP)
c. Legal documents (LEG)
d. Public speeches (PSP)
e. Contemporary literature (LIT)
f. Movie subtitles (MOV)
g. Company websites (COM)
h. Encyclopedia entries (ENC)
i. University websites (UNI)
j. United Nations documents (UND)
k. Research abstracts (ABS)

Documents pertaining to the first three of the above text types – government reports, white papers and legal documents – were comparatively easy to find online simply by searching for "中到英/*zhongdaoying*[Chinese to English translation]" or "中英对照/*zhongyingduizhao*[Chinese and English parallel text]". The English versions are official translations published by the Chinese bodies responsible for the source texts. The government reports were collected from china.org.cn, in which Chinese and English texts were already displayed on the same page and aligned at paragraph level. The white papers were gathered from the official website of Chinese Translation Academy. The Chinese and English versions are displayed in the same webpage too. In terms of legal documents, Chinese laws and regulations and their translations can be found from pkulaw.cn. Although full-length texts are not publicly available, due to the large number of excerpts, this resource is a useful of collection present-day legal translations. In summary, categories a, b and c are comparatively easy to mine.

Another relatively rich source of texts for the corpus were public speeches by Chinese government officials and business leaders translated into English (category d). This same genre is covered in the CEPIC corpus discussed in Section 2. The speeches for ZHEN were collected from china.org.cn, but their corresponding translations needed to be retrieved from separate English websites, so it took some time to locate and pair up the parallel texts. During the collection process, care was taken to include as many topics as possible, such as economy, leisure, environment and so on.

For contemporary literature (category e), it was possible to obtain a number of texts for the corpus from Paper Republic, a registered UK charity that promotes Chinese literature in English translation, focusing on new writing from contemporary Chinese writers. The English translations provided by Paper Republic are supplied by volunteers who are professional translators (Abrahamsen 2020). Although there are a number of literary translations available, the website does not include the original Chinese texts. It was therefore only possible to include in the corpus the texts for which the Chinese source texts were found online and whose permission to use was granted by their authors. We have collected 42 texts under this category, by 40 authors and 20 translators. The publication dates range from 2000–2010, and one short story published in 1981.

Another source of suitable texts for the corpus is movie subtitles (category f). Data was collected from the well-known subtitle forum ZIMUZU.org, where English and Chinese subtitles from Chinese films can be found (Zimuzu 2020). The subtitles are uploaded by volunteers, and while it is not easy to tell whether there are professional translators among them, they still represent authentic, widely read translations. We have collected the subtitles for eight full movies from 2010 onwards.

To represent institutional texts from Chinese companies (category g), a search for the top 100 Chinese companies listed on FortuneChina was initially carried out to identify possible sources. The aim was to download bilingual Chinese-English texts from the introductory pages of Chinese enterprises with a strong international presence. However, there were several unexpected obstacles to the collection process. First, a number of companies did not have a corresponding English website. Second, the English versions of some of the websites were not full translations but simply an introduction to the company. Third, some of the English versions had not been updated and therefore no longer represented a full translation of the updated Chinese section. Fourth, some companies used complex animations and graphics, making it very hard to retrieve text alone. Despite these difficulties, it was possible to mine Chinese-English bitexts from 22 Chinese companies in a range of industries.

Another valuable source for the corpus were encyclopedia entries (category h), which can be very rich in culture-specific items. To obtain the texts, we conducted an advanced search for Chinese topics such as such as Chinese herbs, ceramic art and traditional dance in a bilingual Chinese-English encyclopedia available from gunxiong.com.

One of the most problematic genres to represent in the corpus was news articles. As discussed in the beginning of this section, Chinese news is seldom translated according to its original content into English. The English version of Chinese media usually has its own writers and editors, and the general layout and content is completely different from the original Chinese version, which makes it extremely difficult to locate parallel translated text. However, the top Chinese universities (namely PKU, Tsinghua, Tongji, SHJT and SYSU) often contain a news section in English to promote themselves to foreign students. This section has various topics, including but not limited to recent scientific discoveries, new partnerships with other institutions and delegation visits. Matching the English translations to their Chinese originals was not a straightforward procedure, however, as not all news items are translated into English, and, when they are, the English translation is usually published later in time. We therefore only included in ZHEN English translations linked to original Chinese university news items (category i).

In addition, it was possible to retrieve UN documents originally written in Chinese and then translated into English from the UN Official Document System (ODS). This online database contains full-text, UN documents published from 1993 onwards, including documents published by the Security Council, the General Assembly, the Economic and Social Council and their subsidiaries, as well as administrative issuances and other documents. By conducting a full-text search with the keyword <original: Chinese>, it was possible to retrieve 500 documents originally drafted in Chinese and pair them up with their English translations. Thus, unlike parallel Chinese-English UN documents available from OPUS, where translation direction is unclear, only Chinese to English translations are used in ZHEN (category j).

To represent academic/scientific texts in the corpus, the source used for mining parallel text was CNKI, the national repository where all journals published in China are referenced. Since Chinese journals are required to submit an English version of abstracts along with the Chinese version, the abstracts are a good source for parallel academic/scientific texts. A total of 100 abstracts in Chinese along with their English versions from ten highly cited papers in ten different subject domains were downloaded and aligned for ZHEN.

Table 2 provides a summary of composition of the ZHEN corpus, arranged by the size of each category. As can be seen, the Chinese section of the ZHEN corpus consists of 806,987 Chinese tokens and 1,003,375 English tokens. Although some text types are more represented than others, it can be seen that ZHEN covers a much wider variety of Chinese to English translations than what is usually available in this translation direction. It is also worth noting that the texts in ZHEN were published after 1990, making the corpus representative of contemporary source texts and translations. Full references and metadata to all sources of texts included in the corpus is kept in a separate document. The next section provides details about the alignment of the corpus.

Table 2. The composition of ZHEN corpus

Code	Text type	Documents	Tokens	% of the corpus
GOV	Government reports	20	210,423	26.1
LIT	Contemporary literature	42	189,340	23.5
UND	United Nations texts	50	89,644	11.1
WHP	White papers	7	64,443	8.0
PSP	Public speeches	22	62,314	7.7
LEG	Legal documents	50	61,020	7.6
MOV	Movie subtitles	8	36,042	4.5
ENC	Bilingual encyclopedia	17	33,287	4.1
UNI	University news	25	26,062	3.2
ABS	Academic abstracts	100	21,501	2.7
COM	Company websites	22	12,911	1.6
Total	**12**	**363**	**806,987**	**100.0**

4.2 Parallel text alignment

As discussed in Frankenberg-Garcia (2009: 61), alignment is the one stage in corpus compilation that is unique to parallel corpora. This section describes the alignment criteria used in ZHEN and how different alignment tools were tested to find out which performed best for Chinese and English.

4.2.1 *Alignment criteria*

Source texts and translations can be aligned at different levels, including text, paragraph, sentence, clause and word alignment, but the most common unit of alignment for parallel corpora is sentence by sentence (Frankenberg-Garcia 2009, 2019). This enables corpus users to retrieve sentence-long parallel concordances. According to Varga et al. (2007), it is often the case that a source text sentence corresponds exactly to a target text sentence. However, as discussed in Frankenberg-Garcia and Santos (2003) and Frankenberg-Garcia (2019), translators do not always translate source-text sentences with equivalent target language sentences. Therefore, 1:1 sentence alignment is not always achievable. Other ratios of sentence alignment are 1:0, when sentences are omitted from the translation; 0:1, when sentences are added to the translation without any corresponding text in the source; 1:n, when source-text sentences are split into more than one target text sentence; n:1, when more than one source-text sentence is merged into a single target text sentence; and n:n, when more than one source-text sentence matches more than one target text sentence (Frankenberg-Garcia and Santos 2003; Frankenberg-Garcia 2009; Frankenberg-Garcia 2019).

In the ZHEN corpus, the alignment criteria used are exemplified below, with <s> marking each new sentence:

1:1 alignment <s> 北京大学名列第45位，较去年提升1位，继续在中国内地高校中位列榜首。<\s>

<s>Peking University (PKU) ranks No.45, one place higher than that of last year, keeping being in the first place among all the universities and colleges in mainland China. <\s>

1:n alignment <s>欧洲科学院 (The Academy of Europe)，又称欧洲人文和自然科学院，总部设在英国伦敦，成立于1988年，是欧洲多国科学部长共同倡导创立，英国皇家学会等多个代表欧洲国家最高学术水平的国家科学院共同发起成立的一个包括东、西欧国家的国际科学组织。<\s>

<s>The Academy of Europe, or the European Academy of Sciences, Humanities and Letters, was founded in 1988 with its headquarters in London, UK. <\s> <s>An international scientific organization, across both east and west European countries, it has been proposed by ministers of science in European countries and founded by several national academies of science in Europe, including the Royal Society of UK. <\s>

n:1 alignment <s>在讲座中，菲利普•杰奎琳先生介绍了20世纪初以徐悲鸿为代表的赴法留学艺术家们在巴黎的求学经历. <\s><s>阐述了法国艺术在民国时期和中国美术界的交流，揭示了中法两国之间文化艺术交流的深厚历史渊源。<\s>

<s>In his lecture, Mr. Cinquini introduced the overseas learning of Chinese artists with Xu Beihong as the representative in France at the beginning of the 20th century, elaborated the exchanges among art circles in China and France during the Republic of China era, and recalled the history of cultural and art exchanges between the two countries. <\s>

n:n alignment <s>北京大学是国家级双创示范基地，北京大学创新创业学院参加了“双创“成果展示。<\s><s>中共中央政治局常委、国务院总理李克强参观了北大展区。<\s>

<s>Peking University is one of the national level mass entrepreneurship and innovation demonstration bases. <\s><s>On this occasion, School of Innovation and Entrepreneurship, Peking University, participated in this exhibition and was welcomed by Chinese Premier Li Keqiang. <\s>

Alignment of the type n:n normally occurs when more than one source-text sentences match more than one target-text sentences. As shown in the n:n example above, however, even though there are two sentences in the Chinese source and two sentences in the English translation, the alignment is n:n because the sentence boundaries of the two texts do not match, as shown in the literal translation provided below.

> Literal translation of n:n alignment example
>
> <s>Peking University is one of the national level mass entrepreneurship and innovation demonstration bases, and this School of Innovation and Entrepreneurship, Peking University, participated in this exhibition. <\s> <s> And it was welcomed by Chinese Premier Li Keqiang. <\s>

Although only fully translated documents rather than English summaries were included in ZHEN, during the compilation process it became evident that for some alignment units the translators added a considerable amount of information that was not present in the original, while for others the translators omitted substantial chunks of text. To account for these cases that seem to be recurrent in contemporary Chinese translated into English, extra annotation may be added to notable expansions or reductions detected in an alignment unit. This is exemplified in the Appendix, respectively with an alignment unit of 172 Chinese characters matching 381 English words, and an alignment unit of 522 Chinese characters matching just 91 English words.

Annotating distinct expansions and reductions such as these can facilitate their analysis, helping researchers better understand the phenomenon. To do so, however, it is first necessary to define what might be classified as a notable expansion or reduction, which requires, in turn, a benchmark figure for the normal rate of expansion (or contraction) in the translation of Chinese into English. Although there is some discussion about this, there does not seem to be any scientific evidence backing any claims regarding the phenomenon. Translation agencies which advertise the normal expansion or contraction rate for different language pairs simply classify Chinese into English as "variable", perhaps precisely because of the tendency to sometimes add or remove large chunks of text in the process of translation.[3] Having said this, as discussed in 3.1, according to the parallel text data that has been processed for the ZHEN corpus, the average ratio of Chinese to English tokens when texts are fully translated without significant expansion or reduction seems to be 1.5:1. Using this ratio as a benchmark, we aim to add expansion and reduction tags to alignment units that deviate substantially from the 1.5:1 ratio.

[3] For example, Kwintessential at https://www.kwintessential.co.uk/resources/expansion-retraction and Eriksen Translations at https://www.eriksen.com/language/text-expansion/ (31/01/2020)

4.2.2 *Alignment tools*

Aligning texts in two very different languages – as is the case of Chinese and English – can be particularly problematic, especially if the aligners have been conceived for relatively similar languages. Therefore, before beginning any type of alignment, it was necessary to find out which existing aligner worked best for the Chinese-English pair. A series of tools were tested with a sample of 10 corpus files of 1000 words each.

The first two aligners to be tested – TMXmall and Tsinghua Aligner – had been recommended by Chinese scholars contacted by email. However, TMXmall did not handle English paragraph separation well, despite the fact that the texts to be aligned had paragraph breaks. Another problem was its cost – 300 rmb (around £ 40) per file – which was not possible to cover in this project. Tsinghua Aligner, in turn, crashed frequently, even with small files.

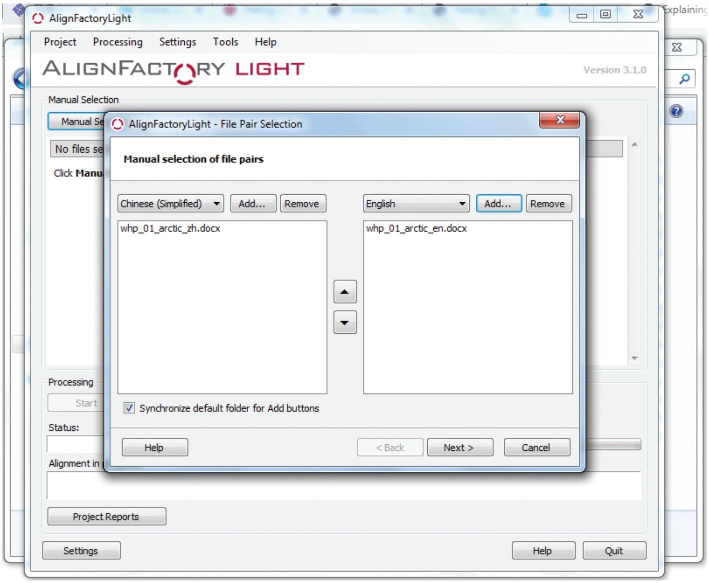
After searching for other possible aligners and raising a question about aligners that worked well for Chinese-English on the Corpora List,[4] a number of other tools were identified, although none had been designed specifically for Chinese-English. One of the aligners recommended – abbyy 2.0 – has been discontinued. Nine other aligners were tested with sample Chinese-English corpus files. The results of how the different aligners compared are summarized in Table 3.

Table 3. Comparison of alignment tools for Chinese-English

Aligner	Cost	Availability	Performance	Output	Editing interface
Abbyy1.0	Free	Stand-alone	Does not recognize Chinese	N/A	User-friendly
Abbyy 2.0	Fee	Discontinued	N/A	N/A	N/A
Tmxmall	£ 40/ page	Online	Crashes frequently	Good	User-friendly
LF aligner	Free	Stand-alone	Good	Medium	Hard to edit
Youalign	Free	Online (5 files/day)	Slow	Good	Hard to edit
Antpconc	Free	Stand-alone	Does not recognize Chinese	N/A	User-friendly
Tsinghua Aligner	Free	Stand-alone	Medium	Medium	Hard to edit
winalign	Trados liscence	with Trados	Slow with large files	Good	User-friendly
paraconc	$ 95 per user	Stand-alone	Medium	Good	User-friendly
Alignfactory light	£ 35 per license (up to 3 users)	Stand-alone	Good	Good	User-friendly

[4] See https://mailman.uib.no/public/corpora/2019-November/030775.html (31/01/2020)

The aligner selected for ZHEN was AlignFactory Light, which is reasonably priced, achieves a good level of alignment performance for Chinese-English and has a user-friendly editing interface. AlignFactory Light outputs bitexts in XML or HTML format, or generates TMX files to import directly into translation memories. The tool's text pairing screen is shown in Figure 1.



Chinese (Simplified)

whp_01_arctic_zh.docx

English

whp_01_arctic_en.docx

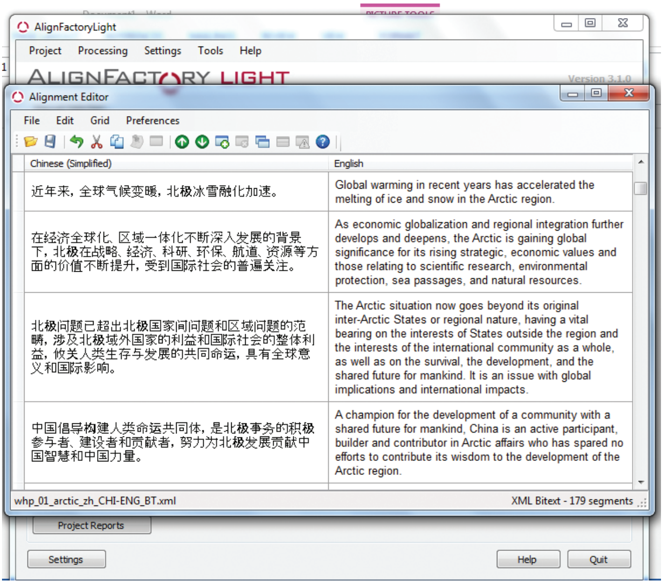
☑ Synchronize default folder for Add buttons

Figure 1. Upload of text pairs in AlignFactory light

As shown in Figure 2, after automatic alignment is achieved, the tool's alignment editor provides a user-friendly interface to view, edit and annotate different types of alignment.

4.3 Corpus compilation

A number of ready-made corpus processing tools can be used to compile parallel corpora, including, for example, WordSmith (Scott, 2008), ParaConc (Barlow, 2002), AntPConc (Anthony 2012), and Sketch Engine (Kilgarriff et al. 2014). We opted to compile the ZHEN corpus with Sketch Engine for several reasons. First, because unlike stand-alone tools like the first three of the above, Sketch Engine does not require any software to be downloaded and is easy to share online with other users. Although Sketch Engine is a proprietary tool, it is open to anyone through payment of an inexpensive subscription fee or, at no cost, from any academic institution in the European Union, thanks to the Elexis sponsorship programme.



Chinese (Simplified)	English
近年来，全球气候变暖，北极冰雪融化加速。	Global warming in recent years has accelerated the melting of ice and snow in the Arctic region.
在经济全球化、区域一体化不断深入发展的背景下，北极在战略、经济、科研、环保、航道、资源等方面的价值不断提升，受到国际社会的普遍关注。	As economic globalization and regional integration further develops and deepens, the Arctic is gaining global significance for its rising strategic, economic values and those relating to scientific research, environmental protection, sea passages, and natural resources.
北极问题已超出北极国家间问题和区域问题的范畴，涉及北极域外国家的利益和国际社会的整体利益，攸关人类生存与发展的共同命运，具有全球意义和国际影响。	The Arctic situation now goes beyond its original inter-Arctic States or regional nature, having a vital bearing on the interests of States outside the region and the interests of the international community as a whole, as well as on the survival, the development, and the shared future for mankind. It is an issue with global implications and international impacts.
中国倡导构建人类命运共同体，是北极事务的积极参与者、建设者和贡献者，努力为北极发展贡献中国智慧和中国力量。	A champion for the development of a community with a shared future for mankind, China is an active participant, builder and contributor in Arctic affairs who has spared no efforts to contribute its wisdom to the development of the Arctic region.

whp_01_arctic_zh_CHI-ENG_BT.xml XML Bitext - 179 segments

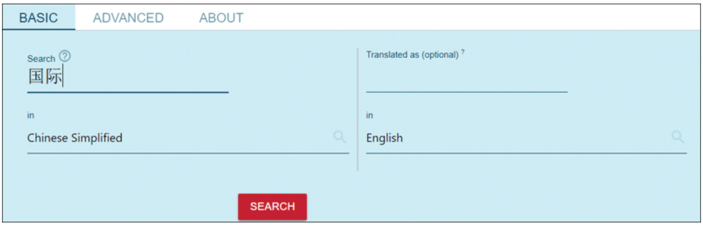
Figure 2. Alignfactory light editing interface

Another advantage of Sketch Engine is that it is very simple to use it to compile a parallel corpus, simply by uploading aligned corpus files in tmx, csv or xls format. In the case of ZHEN, we uploaded the tmx files obtained through AlignFactory Light.

Sketch Engine then automatically tokenizes both the Chinese and the English texts, and inserts part-of-speech annotation to both languages. The Chinese texts are annotated with Chinese Penn Treebank part-of-speech tagset, and English ones are tagged with the English 3.3 Penn Tree Tagger.

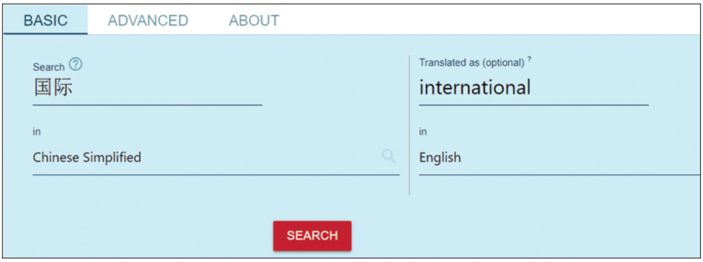
Another feature of Sketch Engine is that is very straightforward to create sub-corpora simply by specifying which corpus files should be included in them. In ZHEN, each text type in Table 2 forms a separate subcorpus of ZHEN, so that it is possible to carry out searches only within a specific category or set of categories (for example, only literary texts).

Sketch Engine has a simple, user-friendly interface for parallel corpus users who do not necessarily have advanced corpus linguistics skills. As shown in Figure 3, users only need to type in a search item in one of the corpus languages and click search. Leaving the "Translated as" option blank will retrieve all parallel concordances with the search item in question. Alternatively, users can fill in the "Translated as" box to retrieve only the concordances where a given search item has a specific equivalent in the alignment (Figure 4), and in the advanced option it is possible to search for an item *without* a specific equivalent in the alignment. The queries are by default case-insensitive and lemma-based (for example, *give* will retrieve *give, gives, gave* and *given*, in small or capital letters).



Search: 国际
in: Chinese Simplified
Translated as (optional):
in: English

Figure 3. Basic parallel concordance query in Sketch Engine

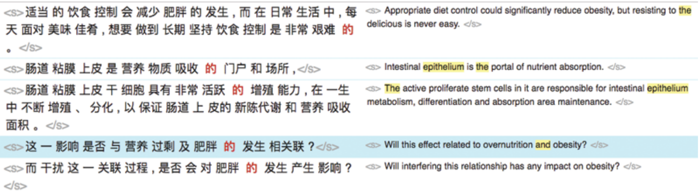


Search: 国际
in: Chinese Simplified
Translated as (optional): international
in: English

Figure 4. Basic parallel concordance query in Sketch Engine with alignment constraint

Additionally, advanced users can take advantage of all the other Sketch Engine functionalities when formulating concordance queries, such as searches involving part-of-speech tags and subcorpora.

The output of the parallel concordances is side-by-side, which facilitates scrolling down the results, as shown in Figure 5. Furthermore, parallel concordances can be saved and/or downloaded to spreadsheets for further analysis.



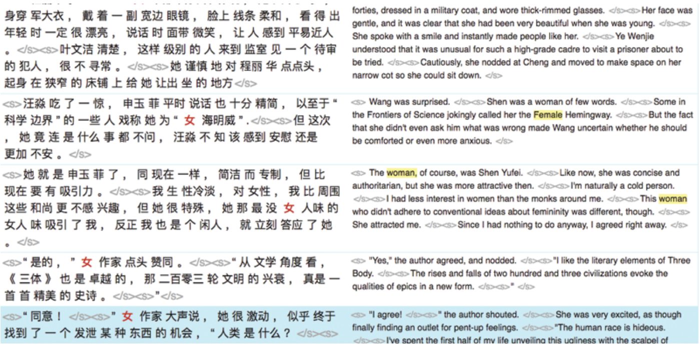
<s>适当 的 饮食 控制 会 减少 肥胖 的 发生，而 在 日常 生活 中，每 天 面对 美味 佳肴，想要 做到 长期 坚持 饮食 控制 是 非常 艰难 的 。</s>	<s> Appropriate diet control could significantly reduce obesity, but resisting to the delicious is never easy. </s>
<s>肠道 粘膜 上皮 是 营养 物质 吸收 的 门户 和 场所，</s>	<s> Intestinal epithelium is the portal of nutrient absorption. </s>
<s>肠道 粘膜 上皮 干 细胞 具有 非常 活跃 的 增殖 能力，在 一生 中 不断 增殖 、 分化，以 保证 肠道 上 皮的 新陈代谢 和 营养 吸收 面积 。</s>	<s> The active proliferate stem cells in it are responsible for intestinal epithelium metabolism, differentiation and absorption area maintenance. </s>
<s>这 一 影响 是否 与 营养 过剩 及 肥胖 的 发生 相关联？</s>	<s> Will this effect related to overnutrition and obesity? </s>
<s>而 干扰 这 一 关联 过程，是否 会 对 肥胖 的 发生 产生 影响？</s>	<s> Will interfering this relationship has any impact on obesity? </s>

Figure 5. Parallel concordances for Chinese “的/de[of]” in ZHEN using Sketch Engine

Apart from parallel concordances, it is also possible to use ZHEN to search just the Chinese source texts or just the English translations. These can in turn be compared and contrasted with monolingual non-translated Chinese and English corpora like the Chinese Web 17 corpus (with over 13 billion tokens), and the British National Corpus or the English Web corpus 15 (with over 15 billion tokens), which are also available from Sketch Engine.

5. Applications of ZHEN

Unlike corpus-like tools like Linguee and parallel corpora which do not differentiate between source texts and translations, or parallel English-Chinese corpora (with English source texts), ZHEN has been specifically compiled to enable one to understand directional shifts in Chinese to English translation. For instance, parallel concordances from Chinese to English can be used to analyse how Chinese culture-specific items have been translated into English, or to understand discourse shifts that might take place in this language direction. In a pilot study based on the English translation of the contemporary science fiction novel *The Three Body Problem* (Liu 2014), Gu (2019) explored the translation of sexist language through parallel concordances using gender-specific words such as "女/nv [female, women]". The parallel concordances like the ones shown in Figure 6 made it clear that the translator consistently made a deliberate attempt to downplay the sexist tone of the source text, which is less acceptable in the English-speaking world today. The concordances can also help Chinese-English translators with authentic examples of translation strategies used to mitigate gender bias.



身穿 军大衣， 戴 着 一 副 宽边 眼镜， 脸上 线条 柔和， 看 得 出 年轻 时 一定 很 漂亮， 说话 时 面带 微笑， 让 人 感到 平易近人 。</s><s>叶文洁 清楚， 这样 级别 的 人 来到 监室 见 一 个 待审 的 犯人， 很 不 寻常 。</s><s>她 谨慎 地 对 程丽 华 点点头， 起身 在 狭窄 的 床铺 上 给 她 让出 坐 的 地方</s>	forties, dressed in a military coat, and wore thick-rimmed glasses. </s><s>Her face was gentle, and it was clear that she had been very beautiful when she was young. </s><s> She spoke with a smile and instantly made people like her. </s><s>Ye Wenjie understood that it was unusual for such a high-grade cadre to visit a prisoner about to be tried. </s><s>Cautiously, she nodded at Cheng and moved to make space on her narrow cot so she could sit down. </s>
<s>汪淼 吃 了 一 惊， 申玉 菲 平时 说话 也 十分 精简， 以至于 " 科学 边界 " 的 一些 人 戏称 她 为 " 女 海明威 ".</s><s>但 这次 ， 她 竟 连 是 什么 事 都 不问， 汪淼 不 知 该 感到 安慰 还是 更加 不安 。</s>	<s> Wang was surprised. </s><s>Shen was a woman of few words. </s><s>Some in the Frontiers of Science jokingly called her the Female Hemingway. </s><s>But the fact that she didn't even ask him what was wrong made Wang uncertain whether he should be comforted or even more anxious. </s>
<s>她 就 是 申玉 菲 了， 同 现在 一样， 简洁 而 专制， 但 比 现在 要 有 吸引力 。</s><s>我 生 性冷淡， 对 女性， 我 比 周围 这些 和尚 更 不感 兴趣， 但 她 很 特殊， 她 那 最 没 女 人味 的 女人 味 吸引 了 我， 反正 我 也 是 个 闲人， 就 立刻 答应 了 她 。</s>	<s> The woman, of course, was Shen Yufei. </s><s>Like now, she was concise and authoritarian, but she was more attractive then. </s><s>I'm naturally a cold person. </s><s>I had less interest in women than the monks around me. </s><s>This woman who didn't adhere to conventional ideas about femininity was different, though. </s><s> She attracted me. </s><s>Since I had nothing to do anyway, I agreed right away. </s>
<s>" 是的， " 女 作家 点头 赞同 。</s><s>" 从 文学 角度 看， 《 三体 》 也 是 卓越 的， 那 二百零三 轮 文明 的 兴衰， 真是 一 首 首 精美 的 史诗 。</s><s>"</s>	<s> "Yes," the author agreed, and nodded. </s><s>"I like the literary elements of Three Body. </s><s>The rises and falls of two hundred and three civilizations evoke the qualities of epics in a new form. </s><s>" </s>
<s>" 同意！ </s><s>" 女 作家 大声说， 她 很 激动， 似乎 终于 找到 了 一 个 发泄 某 种 东西 的 机会， " 人类 是 什么？ </s><s>	<s> "I agree! </s><s>" the author shouted. </s><s>She was very excited, as though finally finding an outlet for pent-up feelings. </s><s>"The human race is hideous. </s><s>I've spent the first half of my life unveiling this ugliness with the scalpel of

Figure 6. Parallel concordances for"女/nv [female, woman]" in *The Three Body Problem*

Another possible application of ZHEN is the analysis of Chinese sense distinctions which are less straightforward to convey in English through parallel concordances in the reverse direction, i.e., from English translations to Chinese source texts. For example, in the same pilot study mentioned above, Gu (2019) found that the phrasal verb *take out* in translated English is a hypernym prompted by a range of subtle variations in meaning in the original Chinese source texts. These include words like "抡起/lunqi [swing (a belt)]" "拔出/bachu [pull out (a sword)" "搬出/banchu [carry (something heavy)]" and so on. Because translation is not symmetrical, it is unlikely that this type of distinction would surface from the analysis of English source texts translated into Chinese.

Besides the application of ZHEN as a parallel corpus, the translated English section of ZHEN (ZHEN-EN) can be contrasted with a reference corpus of non-translated English to identify features that are typical of English that has been translated from the Chinese. For example, Table 4 presents the ten most distinctive adverbs in English that has been translated from Chinese (from ZHEN-EN) using English Web 15 as reference corpus.

Table 4. Distinctive adverbs in English translated from Chinese (ZHEN-EN key adverbs using English Web 15 as a reference corpus)

	Adverb	Freq in ZHEN-EN	Freq in English Web 15	Rel freq in ZHEN-EN	Rel freq in English Web 15
1	resolutely	150	12522	135.182	0.813
2	conscientiously	101	5394	91.022	0.35
3	energetically	106	11283	95.528	0.732
4	vigorously	209	42726	188.353	2.772
5	comprehensively	139	28747	125.268	1.865
6	unswervingly	39	776	35.147	0.05
7	prudently	44	5265	39.653	0.342
8	moderately	115	47906	103.639	3.108
9	steadfastly	38	10397	34.246	0.675
10	earnestly	45	19316	40.554	1.253

As can be seen in Table 4, many of the adverbs that stand out as exceptionally frequent in English that has been translated from Chinese in relation to non-translated English suggest firm determination or purpose, which together reflect a prevailingly positive ideology. In contrast, when reversing the focus corpus and the reference corpus, to identify adverbs that are comparatively under-used in English that has been translated from Chinese, adverbs conveying uncertainty like *allegedly, reportedly, alternatively, hopefully* and *potentially* stand out.

Further contrastive analyses can be undertaken by comparing the lexis of Chinese source texts from ZHEN and the lexis of Chinese texts that do not necessarily get translated into English using a reference corpus like Sketch Engine's Chinese Web 17 corpus. This would allow one to come to a better understanding of what differentiates Chinese that is typically translated into English, and Chinese that does not normally feature in English translation.

6. Conclusion and future directions

After a brief summary of the growing demand for Chinese to English translations beyond official texts, political speeches and traditional literature, this chapter reviewed existing parallel corpora containing Chinese and English and identified a gap in the corpus resources available for the study of Chinese to English translation phenomena. This led us to compile ZHEN, a parallel corpus of contemporary Chinese from a wide range sources aligned with authentic English translations.

During our analysis of possible sources for mining texts for ZHEN, in addition to the more readily available Chinese government reports, white papers, legal documents, public speeches, and UN documents, we were able to select a reasonable amount of parallel texts, including contemporary Chinese literature, encyclopedia entries, film subtitles, news from university websites, and research abstracts. We hope, in this way, to provide users of ZHEN with original data that is not yet available from corpora of Chinese source texts aligned with authentic English translations.

Additionally, the chapter reviewed and tested a number of text aligners with the Chinese-English language pair. Although there are many aligners available in the market, to our knowledge there is not enough information about their suitability for aligning Chinese and English. The review undertaken can thus be useful to inform other projects similar to the present one.

We also discussed the alignment criteria adopted in ZHEN, including the possibility of using expansion and reduction tags, which can be useful to studies investigating a non-trivial, albeit little explored, aspect of current translation practices in China.

Additionally, as exemplified in the previous section, ZHEN can be used to explore a range of other translation phenomena, such as the mitigation of gendered language, sense distinctions like those uncovered by the different source text verbs that match the English phrasal verb *take out*, and contrastive analyses that uncover ideological trends like the notion of certainty versus uncertainty when contrasting English translated from Chinese with non-translated English can be useful not only from a scholarly perspective, but also to inform translator education, translation practice, bilingual lexicography and machine translation systems.

ZHEN was compiled using Sketch Engine, a highly sophisticated yet user-friendly corpus platform that enables corpora to be shared online. Researchers interested in obtaining access to ZHEN are welcome to contact the authors of this chapter.

References

2020. Zimuzu.net. Accessed February 6. http://www.zimuzu.net/

Abrahamsen, Eric. 2020. "Chinese Literature in Translation." *Paper Republic*. Accessed February 6. https://paper-republic.org/

Anthony, Lawrence. 2012. "AntPConc." Waseda University, Tokyo.

Barlow, Michael. 2002. "ParaConc: Concordance software for multilingual parallel corpora." *Proceedings of the Third International Conference on Language Resources and Evaluation*. Workshop on Language Resources in Translation Work and Research.

Frankenberg-Garcia, Ana. & Santos, D. 2003. "Introducing COMPARA: the Portuguese-English Parallel Corpus". In *Corpora in Translator Education*, ed. by Federico Zanettin, Silvia Bernardini & Dominic Stewart. Manchester: St. Jerome, pp 71–87.

Frankenberg-Garcia, Ana. 2009. "Are Translations Longer than Source Texts?: A Corpus-Based Study of Explicitation." In *Corpus Use and Translating*, ed. by Allison Beeby, Patricia Rodríguez & Patricia Sánchez-Gijón. 47–58. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.82.05fra

Frankenberg-Garcia, Ana. 2019. "A Corpus Study of Splitting and Joining Sentences in Translation." *Corpora* 14 (1): 1–30. https://doi.org/10.3366/cor.2019.0159

Gu, Yi. 2017. Chinglish Manifestations in Chinese to English Simultaneous Interpreting, Master dissertation. University of Surrey, England.

Gu, Yi. 2019. *Understanding Chinese to English Translation through the Compilation and Analysis of a Chinese-English Parallel Corpus (ZHEN Corpus)*, Unpublished PhD confirmation report. University of Surrey, England.

Hu, Dexiang. 2006. "Reflections on the Cultural Factors of Translating into and from Chinese" *Journal of Hainan University* 24(3): 355–359.

Johansson, Stig. 1998. "Why Change the Subject? On Changes in Subject Selection in Translation from English into Norwegian." *Target* 16(1): 29–52.

Izquierdo, Marlén, Knut Hofland, and Øystein Reigem. 2008. "The ACTRES Parallel Corpus: an English–Spanish Translation Corpus." *Corpora* 3 (1): 31–41. https://doi.org/10.3366/E1749503208000051

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine: Ten Years On." *Lexicography* 1 (1): 7–36. https://doi.org/10.1007/s40607-014-0009-9

Lao, Long. 1996. "Diudiao huanxiang lianxi shijian [My view on translatology]." *Chinese Translators Journal*. 17(2):38–41.

Liu, Cixin. 2014. *The three-body problem (Vol. 1)*. Macmillan.

Lixun, W. 2001. "Exploring parallel concordancing in English and Chinese." *Language Learning & Technology*, 5(3), 174–184.

Mauranen, Anna. 2004. "Corpora, Universals and Interference." *Translation Universals* Benjamins, 65–82. https://doi.org/10.1075/btl.48.07mau

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge. London.

Mingxing He. 2013. "Zhongguo dangdai wenxue haiwai chuban chuanbo liushi nian [60 years of overseas publication and dissemination of Contemporary Chinese Literature]". *Publish Guangjiao* 7, 18–21.

MultiUN: A Multilingual corpus from United Nation Documents, Andreas Eisele and Yu Chen, LREC 2010

ODS Databases. 2020. *United Nations*. United Nations. Accessed February 6. https://www.un.org/en/databases/

Pan, Jun, & Wong, Tak Ming. 2018. "A corpus-driven study of contrastive markers in Cantonese-English political interpreting." *BRAIN*, 9(2):168–176.

Pinkham, J. 2000. *Zhongshi yingyu zhi jian* [Lessons Drawn from Others' Mistakes of Chingish]. Beijing: Foreign Language Teaching and Research Press.

Scott, Mike. 2008. "Developing WordSmith." *International Journal of English Studies* 8.1, 95–106.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, 2214–2218.

Translators Association of China (TAC). http://tac-online.org.cn. Accessed 25 November 2019.

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. "Parallel Corpora for Medium Density Languages." *Recent Advances in Natural Language Processing IV Current Issues in Linguistic Theory*, 247–58. https://doi.org/10.1075/cilt.292.32var

Wang, Kefei. 2004a. *Shuangyu Duiying Yuliaoku: Yanzhi yu Yingyong* [A parallel corpus: Compilation and application]. Beijing: Foreign Language Teaching and Research Press.

Wang, Kefei. 2004b. "Shuangyu pingxing yuliaoku zai fanyi jiaoxue shang de yongtu [The use of parallel corpora in translator training]". *Computer-Assisted Foreign Language Education* 6: 27–32.

Wang, Kefei. 2009. "yingyihan yuyan tezheng tantao" [The features of translated Chinese from English]". *Computer-Assisted Foreign Language Education* 1.

Xiao, Richard. 2008. "Theory-driven Corpus Research: Using Corpora to Inform Aspect Theory." In *Corpus Linguistics: An International Handbook*, ed. by Anke Lüdeling & Merja Kytö, 977–1007, Berlin: Mouton de Gruyter.

Xiao, Richard, He, Lianzhen, & Yue, Ming. 2008. *The ZJU Corpus of Translational Chinese (ZCTC)*. Zhejiang University.

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): 5–35. https://doi.org/10.1075/ijcl.15.1.01xia

Yang, Zhengquan. 1999. "Xu [Introduction]." In *Zhongguo waiwenju wushinian dashiji* [A chronology of China Foreign Languages Publishing and Distribution Administration (1949–1999)], ed by Yannian Dai and Rinong Chen. Beijing: New Star Press, I–VI.

Zanettin, Federico, 2000. "Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis." In *Intercultural Faultlines*, ed. by Maeve Olohan, 105–118, London: Routledge.

Zanettin, Federico. 2014. *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. Routledge. https://doi.org/10.4324/9781315759661

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. 2016. *The United Nations Parallel Corpus, Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.

Appendix

1:n alignment + expansion

<s> “第四次工业革命”开启的信息时代又称“云时代”，国际战略以“云”的英文单词（CLOUDS)中每个字母所启 的相应单词词义为原点，构建体现北大特质、契合时代要求的国际发展理念体系，C代表Creativity (创新），是战略方针；L代表Leadership (引领)，是战略担当；O代表Openness (开放），是战略路径；U代表Uniqueness (独特），是战略底蕴；D代表Diversity (多元），是战略载体。S代表Shaping (塑造），是统领北京大学国际发展的战略使命。<\s>

(172 characters)

<s><*exapanded*> The information age opened by the “Fourth Industrial Revolution” is also called the “cloud era”, <\s><s>The Global Excellence Strategy is built based on the corresponding word meanings opened by each letter in the English word “CLOUDS”. <\s><s>C stands for creativity and it is a strategic approach to international development. <\s><s>Creativity has become the main driving force for the development of human civilization, and many problems faced by mankind rely on collaborative creativity to be resolved. <\s><s>Peking University will enhance collaborative creativity when it comes to scientific research, and generously pool in high-quality scientific research resources. <\s><s>L stands for leadership and it is a strategic role in the international development of Peking University. <\s><s>Standing up to the tide and leading the times are the roles of Peking University’s students. <\s><s>The future talents should not only have extensive background knowledge and an open, inclusive, collaborative spirit, but also be equipped with innovation, exploration skills and a lifelong learning ability. <\s><s>O stands for openness and it is the strategic path of international development at Peking University. <\s><s>Openness brings progress while closure leads to backwardness. <\s><s>The degree of interconnection of all mankind is unprecedented, and the problems and challenges faced are endless. <\s><s>Expanding openness, enhancing exchanges and cooperation are essential for the development of higher education in the world today. <\s><s>Peking University will continue to participate in exchanges and cooperation with universities and international community around the world to achieve development. <\s><s>U stands for uniqueness and it is the strategic foundation of the international development of Peking University. <\s><s>The new era of globalization is a time filled with more distinctive personality. <\s><s>Each university’s nation has its own unique historical heritage, cultural heritage and spiritual core, which encourage civilized exchanges and mutual learning. <\s><s>In the new era, Peking University will adhere to its Chinese characteristics, as it contributes unique oriental color to the forest of world-class universities. <\s><s>D stands for diversity, which is the strategic carrier of the international development of Peking University. <\s><s>The combination of diversity, uniqueness and integration creates an organic and harmonious whole. <\s><s>S stands for shaping, which is the strategic mission of the international development of Peking University. <\s><s>Its core is to enhance the ability of Peking University to lead the times and to contribute to the world through its global shaping. <\s>

(381 words)

n:n alignment + reduction

<s>六大行动计划包括：国际科研协同创新计划、全球卓越人才培养计划、全球卓越新型互联计划、国际发展特色行动计划、国际智识高地打造计划、全球合作协同推进计划。<\s><s>历经双甲子，跨越三世纪。常为新的北大一直在融汇全球的道路上阔步前行。<\s><s>1898年，京师大学堂（北京大学前身）成立，张百熙在《筹办京师大学堂情形疏》提出"为五洲万国所共观瞻"的理想胸怀。<\s><s>1918年，建校20周年，蔡元培先生提出："本校二十年之历史，仅及柏林大学五分之一，莱比锡大学二十五分之一，苟能急起直追，未尝不可与为平行之发展。" <\s><s>1998年，建校100周年，创建世界一流大学上升为国家战略，随后的20年里，北大开放办学、迈向全球，创建世界一流大学的事业不断攀登新的高峰。<\s><s>2014年5月，建校116周年，习近平总书记视察北大时提出要"认真吸收世界上先进的办学治学经验，遵循教育规律，扎根中国大地办大学"，建设世界上"第一个北大"。北大启动了创建中国特色世界一流大学的新征程。<\s><s>121年来，一代代北大人始终胸怀天下、放眼世界，以筚路蓝缕、以启山林的气魄书写了北大从中国迈向世界、矢志一流的奋斗历程和光荣传统。<\s><s>全新国际发展战略云集思想，云萃智慧，体现着北大面向未来的责任与担当。<\s><s>新时代，北大将继续引领卓越，一往无前，以全球塑造力构筑美好未来。<\s>

(522 characters)

<s><*reduced*>The six action plans include: International Research Collaborative Innovation Program, Global Excellent Talents Program, Global Excellence in New Connectivity Program, International Development Special Action Plan, International Intelligence Knowledge Hub Building Program, and Global Collaborative Initiative. Since its founding, Peking University has launched itself on a journey to create a world-class university with Chinese characteristics. <\s><s>The new "Global Excellence Strategy" gathers ideas, intelligence, and reflects the responsibility of Peking University. In the new era, Peking University will continue to strive for excellence and harness its international strength to shape a better future. <\s>

(91 words)

Word alignment in a parallel corpus of Old English prose

From asymmetry to inter-syntactic annotation

Javier Martín-Arista
Universidad de La Rioja

This chapter proposes a model of syntactic annotation for the Parallel Corpus of Old English Prose, an aligned corpus of Old English and Present Day English texts. The research focuses on areas of syntactic divergence between the aligned texts. Syntactic divergence is described in terms of four types of alignment asymmetry (markedness, constituency, order, and configuration) and is represented by means of two components: a structural description and a dependency tree. The main conclusion is that these two components constitute a historical micro-grammar that identifies stability and change with respect to specific categories and constructions.

Keywords: parallel corpus, alignment, syntactic annotation, asymmetry, Old English

1. Introduction

The last decades have witnessed, along with a growing interest in Corpus Linguistics, a thorough investigation into the points of contact between this linguistic discipline and others like Translation and Lexicography. The works by authors such as Hanks (2012), Kübler & Zinsmeister (2014), Schierholz (2015), and Faaß (2017), to cite just a few, explore these regions. Against this background, this chapter intends to be a contribution to corpora and translation research. While its topic, a parallel corpus, does not represent a completely new advance in Corpus Linguistics, the compilation and alignment of a parallel corpus that involves the modernisation of a previous diachronic stage of the target language stands up as largely virgin territory within the province of Translation.

With these premises, this chapter deals with the syntactic annotation of a parallel corpus with word alignment. The type of correspondence guides the

https://doi.org/10.1075/btl.158.03mar

design of syntactic annotation. In a parallel corpus with correspondence based on inter-linguistic translation, the linguistic distance between the source and the target language points to a degree of divergence that advises full syntactic annotation for both the source and the target. On the other hand, a parallel corpus whose correspondence relies on intra-linguistic translation or modernisation (a type of translation that involves the rendering of a text written in a previous diachronic stage of a natural language, as it is the case with Old English-Present Day English, henceforth PDE) necessarily displays a narrower linguistic distance between the source and the target language. This means that the full syntactic annotation of the source and the target language versions of a parallel corpus involving modernisation may present more points of convergence than of divergence, which, in turn, may result in some degree of descriptive inefficiency and redundancy. At the same time, it is predictable that when two texts written in different diachronic stages of the same language are compared, some mismatches arise. This may be of special relevance to English, which has significantly shifted throughout its history from a fully Germanic language to one with an outstanding Romance component, identifiable both in its morpho-syntax and lexicon.

This chapter addresses the research question of how to devise and implement a model of syntactic annotation for a parallel corpus aligned at word level. More specifically, the following sections raise the issues of the identification of the areas of divergence between the syntax of the source and the target language; the definition of the scope of an inter-syntax is which syntactic divergence is couched in terms of asymmetry; and the development of the components, categories and functions of the inter-syntax. This said, the chapter is structured as follows. Section 2 reviews previous work in parallel corpora from three perspectives: descriptive, pre-theoretical and theoretical. Section 3 lays the foundations of the Parallel Corpus of Old English Prose (hereafter ParCorOE) and presents the standards that guide its design and compilation. Section 4 proposes an inter-syntax that focuses on the syntactic divergences between the source language and the target language, which are captured in terms of alignment asymmetry. Four types of asymmetry are distinguished: marking, constituency, order and configuration. Section 5 applies this analysis of mismatches, thus identifying the main syntactic phenomena that may resist one-for-one word alignment in ParCorOE. This section also unfolds the inter-syntax of ParCorOE, which is comprised of a structural description and a dependency tree represented with graph theory. To close this work, Section 6 summarises the main conclusions and offers some avenues for future research.

2. Background

A parallel corpus is a type of bilingual or multilingual corpus that contains texts from the source language and their translations (McEnery 2003: 450). In contrast, a comparable corpus *can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representativeness* (McEnery & Xiao 2007a: 3). It is a central requirement of parallel corpora that they align the source texts and their translations, either at word or sentence level (McEnery & Xiao 2007a: 3).

According to Aijmer and Altenberg (1996, in McEnery and Xiao 2007b: 131), parallel corpora can be used for conducting a wider array of studies than monolingual corpora. Parallel corpora also have various applications to lexicography, language teaching and acquisition, as well as translation. Given that a parallel corpus offers *direct comparability* (Enrique-Arias 2013: 105), diachronic research can benefit from parallel texts because all the target language forms that express a given content from the source language can be analysed.

In spite of the advantages and applications of parallel corpora summarised in this section, a parallel corpus for English Historical Linguistics in general, and for Old English in particular, is not available at the moment. Such an undertaking should consider the state of play regarding the need to automatise corpus annotation. Some authors, such as Lu (2014), underline the importance of Natural Language Processing technology, which allows computers to annotate large corpora at different linguistic levels, so that a minimum of manual revision is required. This aim can be achieved more effectively in corpora of natural languages than in historical corpora because the latter are far smaller, thus resisting statistical processing, and, above all, because historical corpora often raise issues of spelling variation that preclude fully automatic annotation (Johnson 2009). In Historical Linguistics in general and Old English in particular, spelling variation turns the lemmatisation of the corpus -the attribution of the textual forms to the corresponding dictionary forms (Schierholz 2015)- into the central task of corpus tagging and annotation (Martín Arista 2013, 2017a, 2017b): only when a textual form has been assigned to a lemma through a normalisation procedure, is it possible to automatically provide the token in question with the relevant information from the dictionary database (Tío Sáenz 2015; Metola Rodríguez 2017; Novo Urraca and Ojanguren López 2018; García Fernández 2018).

Therefore, the compilation of an aligned parallel corpus represents a challenging project relevant for corpus and translation research, as well as an investigation with various applications to the linguistic analysis and the lexicography of Old English. From the descriptive point of view, to date there is not a large collection of annotated aligned parallel texts for the study of Old English. In pre-theoretical terms, no parallel corpus has been compiled so far that comprises a text in a historical language and its modernised version. On the theoretical side, the central aspect of a parallel corpus is alignment, in such a way that the more exhaustive tagging and annotation is required the more accurate alignment needs to be. To this effect, this chapter takes the line that the alignment in a corpus that revolves around modernisation has to be guided by the divergences between the old and the modern version of the text, given that diachronic continuity makes allowance for the exclusion of the areas of morphosyntactic convergence. These aspects are discussed in turn in the remainder of this section.

Beginning with the descriptive aspects, the most widely used corpora of Old English include the Old English segment of the *Helsinki Corpus of English Texts* (Rissanen et al. 1991), which contains around 300,000 words; *The York-Helsinki Parsed Corpus of Old English Poetry* (Pintzuk and Plug 2001), which comprises approximately 70,000 words; *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor et al. 2003; henceforth YCOE), which files ca. 1.5 million words; and the *Dictionary of Old English Corpus* (Healey et al. 2004), which gathers around three million words and was specifically compiled for the *Dictionary of Old English* (Cameron et al. 2018). These corpora are segmented by fragment and text, with tokens identified by means of a specific number or, in the case of the *Dictionary of Old English Corpus*, by means of the Cameron number (Mitchell et al. 1975, 1979). The four corpora are marked-up at text level. The *Helsinki Corpus of English Texts*, for instance, provides each fragment file with the abbreviated title, sub-period, manuscript date, dialect, text type, genre, and information on the translation, if relevant. *The York-Helsinki Parsed Corpus of Old English Poetry* and *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (henceforth YCOE) have been tagged morphologically (category and morphological case) and parsed syntactically (hierarchy and linearisation). In spite of the wealth of philological data compiled in these corpora, two major pending tasks remain for Old English linguistics, corpus analysis and lexicography: the lemmatisation of the written records and the compilation of a representative parallel corpus Old English-English.

As for the methodological questions, parallel corpora, as a general rule, compare languages from different linguistic branches, such as Portuguese (Romance) and English (Germanic) with respect to Indo-European; or language belonging to two distinct sub-branches of a linguistic family, as is the case with English (West-Germanic) and Swedish (North-Germanic) within Germanic. Even when it comes to compiling corpora for Historical Linguistics, such corpora tend to be comprised of versions from different languages, rather than presenting two diachronic stages of the same language. For example, the *ENHIGLA* (Old English – Old High

German – Latin) parallel corpus contains ca. 21,000 clauses (available at http://pelcra.pl/enhigla/corpus) from the Latin version and the Old English translation of the first twenty-five chapters from the Book of Genesis and the first ten chapters from the Gospel of Luke; the Latin original of and the Old English version of Book I and a fragment of Book II from Bede's *Historia ecclesiastica gentis anglorum*; as well as the Latin version and the Old High German translation of the first seventy-four chapters from Tatian's *Gospel Harmony, De fide catolica contra iudeos* by St. Isidor of Seville, and *Physiologus*. Put differently, parallel corpora like the ones cited above do not make a claim of continuity on the diachronic axis, which a parallel corpus comparing two diachronic stages of a language certainly does.

With regard to the theoretical aspects mentioned above, it has already been remarked that parallel corpora rely on the correspondence between the source language and the target language texts. This comparison can be established at several levels. Authors such as Kübler and Zinsmeister (2014), as well as Krause and Zeldes (2016), insist on the importance of annotation to achieve the goal of increasing searchability and put forward levels of annotation below the text level that include the sentence and the word level. Sentence level and word level alignment, however, not only beg for additional tokenisation with respect to text alignment but also demand more detailed tagging and annotation. In other words, alignment empirically demonstrates the correspondence between the texts that has been assumed as the point of departure of the research; and, ultimately, relates tokenisation to searchability, which is in need of extensive and accurate tagging and annotation at sentence and word level. Alignment also makes for the adequacy of lemmatisation, which, as pointed out above, constitutes the central task of corpus tagging and annotation. Last but not least, alignment defines the scope of the morphosyntactic comparison between the source text and the target text. Put briefly, alignment can be considered the main characteristic of a parallel corpus.

While alignment determines the scope, asymmetry emphasises certain aspects of the comparison, thus disregarding symmetric parts of the comparison between the source and the target text. Defined in these terms, asymmetry accounts for the divergence between the source and the target under comparison and, conversely, symmetry couches the convergent aspects of the comparison of the source and the target text, which is basically put aside. As for continuity on the diachronic axis, symmetry corresponds to stability while asymmetry indicates change. An inter-syntax is proposed in Sections 4 and 5 that can represent and explain the relevant aspects of morphological case marking, functional relations, argument projection and linearisation.

3. The design of an aligned parallel corpus of Old English prose

Against the background presented in the previous section, ParCorOE, an aligned parallel corpus of Old English prose, is an ongoing project that aims at compiling 300,000 words in the source language, plus the parallel version in the target language. The basic parameters of ParCorOE can be set as follows. As regards general orientation, the corpus will be historical, rather than a corpus devised for translation, comparative linguistics or second language learning. With respect to the number of languages selected, ParCorOE will be bilingual, involving Old English and PDE. As far as directionality is concerned, ParCorOE will be unidirectional: from Old English to PDE. As for the target, ParCorOE will be aimed to textual forms (tokens or inflections), instead of revolving around dictionary words or lemmas. Concerning genre, ParCorOE will select prose texts only.

With these parameters, the following standards guide the design and compilation of ParCorOE. These standards serve the general aim of increasing searchability.

Standard 1: Alignment: An aligned parallel corpus Old English-English consists of a parallel text, that is to say, an Old English text placed along its PDE modernisation, with alignment at text, sentence and word level, in such a way that every source language segment is paired with a target language segment. Word, sentence, and text alignment is in need of tokenisation at these three structural levels. Alignment parings should be marked by means of the highlighting of the source and the target segment (See Figure 2 below).

Standard 2: Annotation: Three types of annotation must be distinguished: mark up at text level, as well as syntactic annotation and morphological tagging at sentence/word level. Fragments (tokens) are comprised of at least one sentence or one syntactically independent period, identified by means of a text number, such as Mart 55.07.07, corresponding to *Ond monige menn gesegon ðæt ða deadan arison of ðæm byrgennum ond eodon geond ða halgan burh on Hierusalem, oð ðæt Crist eft aras*. (And many people saw the dead arise from their graves and walk through the holy town of Jerusalem until the resurrection of Christ).

Standard 3: Lemmatisation: The corpus must be fully lemmatised, so that all the textual attestations are grouped under the relevant lemma, and each lemma is provided with all its inflections. For example, the following inflections have been attibuted so far to the verbal lemma *niman* 'to take': *nam* (ind. pret. 3rd sing.), *naman* (ind. pret. pl.), *name* (subj. pret. sing.), *namon* (ind. pret. pl.), *namon* (subj. pres. pl.), *namon* (subj. pret. pl.), *nim* (imp. sing.), *nimað* (imp. pl.), *nimað* (ind. pres. pl.), *niman* (infinitive), *nimð* (ind. pres. 3rd sg.), *nime* (ind. pres. 1st sg.), *nime* (subj. pres. sing.), *nime* (subj. pret. sing.), *nimeð* (ind. pres. 3rd sg.), *nimen* (infinitive), *nimen* (subj. pres. pl.), *nimenne* (infl. inf.)), *nimest* (ind. pres. 2nd sg.), *nimine* (infl. inf.), *numen* (pa. part.), *nyme* (subj. pres. sing.). In token analysis, 485 inflections have been lemmatised under the verb *niman*.

Standard 4: Automation: Within the limits imposed by the available written standards and the variation that they present, the annotation of the parallel corpus must be automatic. This includes not only syntactic annotation and morphological tagging, but also the necessary lemmatisation. Lemmas and inflections must be listed dynamically, so that users have access to ablaut patterns, such as *nim-nam-nom-num* in *niman* 'to take'; elision, as in *nymaþ/nymþ* and other spelling alternatives, like *nimaþ/neomaþ/niomaþ*.

Standard 5: Feeding: The corpus must be fed with the information available from a knowledge base of Old English. The parallel corpus may retrieve information from the relational databases in the knowledge base of Old English in order to maximise the automation of the tasks of tagging, annotation and lemmatisation. For instance, additional spellings and inflections are automatically fed from the knowledge base to the lemmatisation of *niman*, including *neoman* (subj. pres. pl.), *neomendum* (pres. part. dat. pl.), *nimæð, neomaþ, nimaþ, niomað, nymþ* (ind. pres. pl.), *nimst, nimest* (pres. 2nd sg.), *niomanne, nimanne, nymenne* (infl. inf.), *nome*, (ind. pret. 2sg.), and *nomon* (ind. pret. pl.).

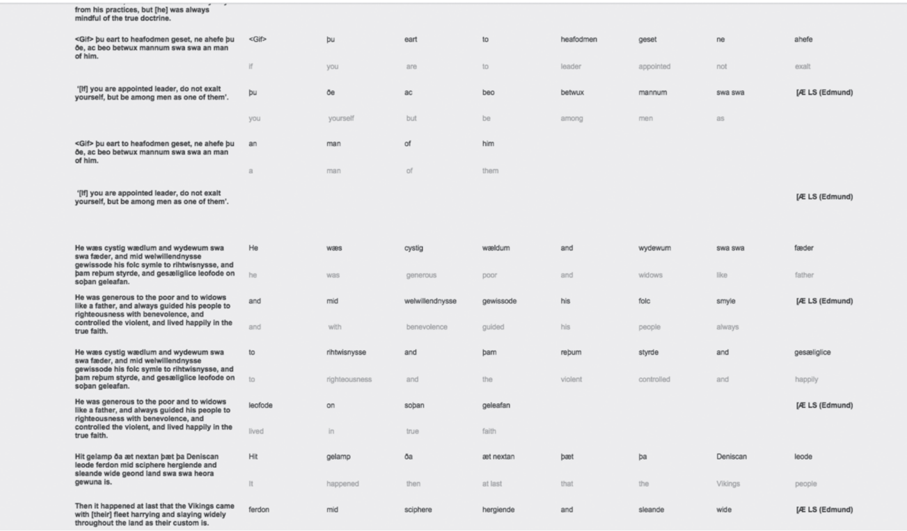
Standard 6: Searchability: The corpus must be searchable by text, fragment and word, as well as by morphological tag and syntactic annotation. Combined searches by inflectional form and lemma are also required. The corpus must be based on a concordance and an index, so that the main layouts are interconnected (see Figures 1 and 2).

Standard 7: Dissemination: The corpus must be available online in open access (see Figure 2).

To recapitulate, the background, parameters and standards presented so far point to a corpus compatible with theoretical studies as well as applications of Old English lexicography and presentations of Digital Humanities. Turning to the question of representativeness, McEnery (1996: 123) stresses the importance of the corpora of historical languages and remarks that, exactly like the corpora of natural languages, historical corpora must be quantitatively sufficient and qualitatively representative so as to offer an accurate representation of the language of analysis. Biber (2007) suggests that a corpus that has been compiled in various stages is more likely to be representative. This author recommends to design and implement a pilot corpus that gathers as much variation as possible, so that the compilers can identify specific issues and general problems. Heid (2008: 43) calls the design and implementation of a pilot corpus *preprocessing* and holds that for an approach to be corpus-based rather than corpus-driven, preprocessing is necessary.

In this line, a ten-thousand-word pilot corpus was compiled and annotated (Martín Arista 2017a, 2017b, 2018). The texts, as well as their modernisations, were extracted from Fernández Cuesta et al. (1997). The aim of the pilot corpus was to find design inadequacies and compilation shortcomings. From the quantitative point of view, ten thousand fully tokenised and annotated words suffice to raise issues in the corpus architecture as well as inconsistencies to the tokenisation and the annotation. From the qualitative point of view, a variety of prose texts were chosen. The selection of texts comprised fragments from the *Anglo-Saxon Chronicle, Orosius*, Ælfric's *Lives of Saints, Cura Pastoralis*, and Bede's *Ecclesiastical History*, thus including the major genres of historical prose, religious prose and translations from Latin. This set of texts is representative of the dialect of the vast majority of the records of Old English, which are written in the West Saxon variety. As to datation, Bede's *Ecclesiastical History*, and *Cura Pastoralis* can be dated to the 9th. century (early Old English); *Orosius* and the fragments from the *Anglo-Saxon Chronicle* can be dated to the 10th. century (classical Old English); while the *Lives of Saints* corresponds to the 11th. century (late Old English).

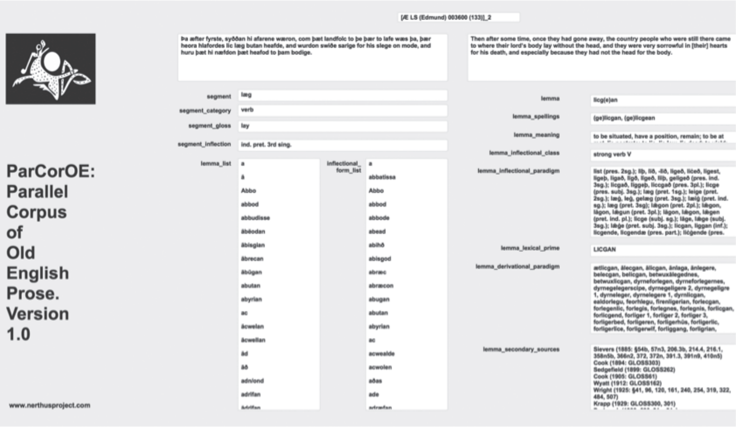
The pilot corpus has two main components: the concordance (including a word index) to the texts and the parallel corpus layouts. Two layouts have been distinguished: the static presentation and the dynamic presentation. The static presentation offers the running texts Old English-PDE, aligns them by fragment and word and provides word-for-word gloss as well as fragment modernisation. This is presented in Figure 1.



from his practices, but [he] was always mindful of the true doctrine.								
<Gif> þu eart to heafodmen geset, ne ahefe þu ðe, ac beo betwux mannum swa swa an man of him.	<Gif>	þu	eart	to	heafodmen	geset	ne	ahefe
	if	you	are	to	leader	appointed	not	exalt
'[If] you are appointed leader, do not exalt yourself, but be among men as one of them'.	þu	ðe	ac	beo	betwux	mannum	swa swa	[Æ LS (Edmund)
	you	yourself	but	be	among	men	as	
<Gif> þu eart to heafodmen geset, ne ahefe þu ðe, ac beo betwux mannum swa swa an man of him.	an	man	of	him				
	a	man	of	them				
'[If] you are appointed leader, do not exalt yourself, but be among men as one of them'.								[Æ LS (Edmund)
He wæs cystig wædlum and wydewum swa swa fæder, and mid welwillendnysse gewissode his folc symle to rihtwisnysse, and þam reþum styrde, and gesæliglice leofode on soþan geleafan.	He	wæs	cystig	wældum	and	wydewum	swa swa	fæder
	he	was	generous	poor	and	widows	like	father
He was generous to the poor and to widows like a father, and always guided his people to righteousness with benevolence, and controlled the violent, and lived happily in the true faith.	and	mid	welwillendnysse	gewissode	his	folc	smyle	[Æ LS (Edmund)
	and	with	benevolence	guided	his	people	always	
He wæs cystig wædlum and wydewum swa swa fæder, and mid welwillendnysse gewissode his folc symle to rihtwisnysse, and þam reþum styrde, and gesæliglice leofode on soþan geleafan.	to	rihtwisnysse	and	þam	reþum	styrde	and	gesæliglice
	to	righteousness	and	the	violent	controlled	and	happily
He was generous to the poor and to widows like a father, and always guided his people to righteousness with benevolence, and controlled the violent, and lived happily in the true faith.	leofode	on	soþan	geleafan				[Æ LS (Edmund)
	lived	in	true	faith				
Hit gelamp ða æt nextan þæt þa Deniscan leode ferdon mid sciphere hergiende and sleande wide geond land swa swa heora gewuna is.	Hit	gelamp	ða	æt nextan	þæt	þa	Deniscan	leode
	It	happened	then	at last	that	the	Vikings	people
Then it happened at last that the Vikings came with [their] fleet harrying and slaying widely throughout the land as their custom is.	ferdon	mid	sciphere	hergiende	and	sleande	wide	[Æ LS (Edmund)

Figure 1. The static presentation of the pilot corpus

The dynamic presentation of the parallel corpus is aligned at word level. Each word is highlighted in the source and in the target text. Full tagging and annotation are fed from the Knowledge-Base of Old English (Martín Arista and Ojanguren López 2018). The information that has been imported from the databases includes lemma, alternative spellings, lexical category, morphological class, inflectional paradigm, derivational paradigm, meaning definition, and the references of secondary sources that deal with the lemma or the inflectional form in question. As shown in Figure 2, there are two basic query options, by inflectional form and by lemma. Full inventories of inflectional forms and lemmas are available. The database software that files the corpus guarantees information retrieval through simple, combined and stepwise searches. It also facilitates open access because it makes allowance for an online publication that can be accessed and searched with an Internet browser.



ParCorOE: Parallel Corpus of Old English Prose. Version 1.0

www.nerthusproject.com

[Æ LS (Edmund) 003600 (133)]_2

Þa æfter fyrste, syððan hi afarene wæron, com þæt landfolc to þe þær to lafe wæs þa, þær heora hlafordes lic læg butan heafde, and wurdon swiðe sarige for his slege on mode, and huru þæt hi næfdon þæt heafod to þam bodige.

Then after some time, once they had gone away, the country people who were still there came to where their lord's body lay without the head, and they were very sorrowful in [their] hearts for his death, and especially because they had not the head for the body.

segment: læg

segment_category: verb

segment_gloss: lay

segment_inflection: ind. pret. 3rd sing.

lemma_list: a, ā, Abbo, abbod, abbudisse, ābēodan, ābisgian, ābrecan, ābūgan, abutan, abyrian, ac, ācwelan, ācwellan, ād, āð, adn/ond, adrīfan, ādrīfan

inflectional_form_list: a, abbatissa, Abbo, abbod, abbode, abead, abihð, abisgod, abræc, abræcon, abugan, abutan, abyrian, ac, acwealde, acwolen, aðas, ade, adræfan

lemma: licg(e)an

lemma_spellings: (ge)licgan, (ge)licgean

lemma_meaning: to be situated, have a position, remain; to be at

lemma_inflectional_class: strong verb V

lemma_inflectional_paradigm: list (pres. 2sg.); liþ, lið, -lið, ligeð, liceð, ligest, ligeþ, ligað, ligð, ligeð, liiþ, geligeð (pres. ind. 3sg.); licgað, liggeþ, liccgað (pres. 3pl.); licge (pres. subj. 3sg.); læg (pret. 1sg.); leige (pret. 2sg.); læġ, leġ, gelæg (pret. 3sg.); læiġ (pret. ind. sg.); læg (pret. 3sg); lǣgon (pret. 2pl.); lǣgon, lāgon, lǣgun (pret. 3pl.); lāgon, lǣgon, lǣgen (pret. ind. pl.); licge (subj. sg.); lāge, lǣge (subj. 3sg.); lǣġe (pret. subj. 3sg.); licgan, liggan (inf.); licgende, licgendæ (pres. part.); liċġende (pres.

lemma_lexical_prime: LICGAN

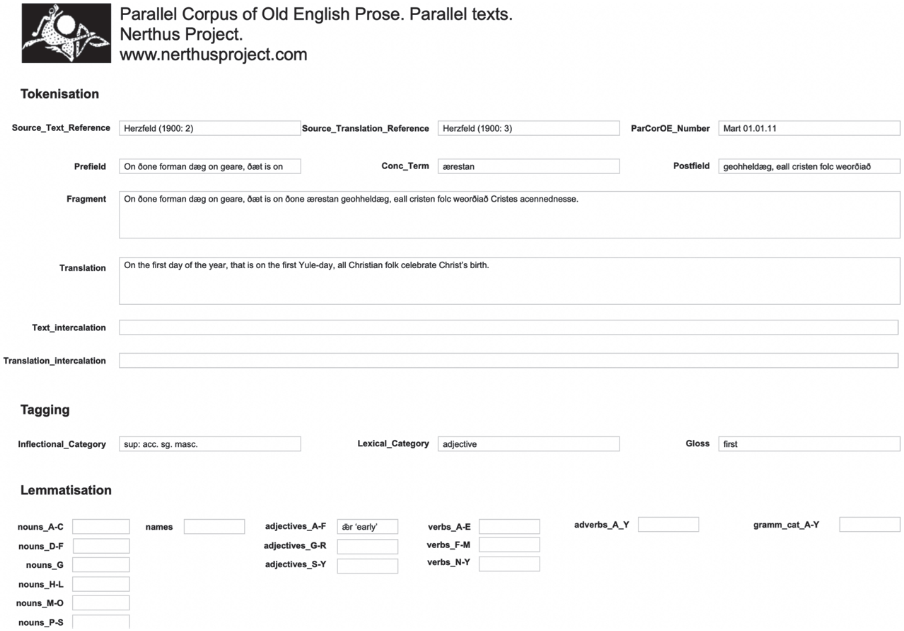
lemma_derivational_paradigm: ætlicgan, ālecgan, ālicgan, ānlaga, ānlegere, belecgan, belicgan, betwuxālegednes, betwuxlicgan, dyrneforlegen, dyrneforlegernes, dyrnegelegerscipe, dyrnegeligere 2, dyrnegeligre 1, dyrneleger, dyrnelegere 1, dyrnlicgan, ealdorlegu, feorhlegu, firenligerian, forlecgan, forlegenlic, forlegis, forlegnes, forlegnis, forlicgan, forlicgend, forliger 1, forliger 2, forliger 3, forligerbed, forligeren, forligerhūs, forligerlic, forligerlice, forligerwīf, forliggang, forligrian,

lemma_secondary_sources:
Sievers (1885: §54b, 57n3, 206.3b, 214.4, 216.1, 358n5b, 366n2, 372, 372n, 391.3, 391n9, 410n5)
Cook (1894: GLOSS303)
Sedgefield (1899: GLOSS262)
Cook (1905: GLOSS61)
Wyatt (1912: GLOSS162)
Wright (1925: §41, 96, 120, 161, 240, 254, 319, 322, 484, 507)
Krapp (1929: GLOSS300, 301)

Figure 2. The dynamic presentation of the pilot corpus

In its present state, ParCorOE consists of ca. 160,000 word files, with the new layout presented in Figure 3.

Quantitatively speaking the final corpus will comprise 300,000 words, the first half being due by March 2021. This amount represents about one tenth of all the written records of Old English. From the qualitative point of view, all the major prose genres of Old English have already been included. The words processed so far by category and text are tabulated in Table 1.



Parallel Corpus of Old English Prose. Parallel texts.
Nerthus Project.
www.nerthusproject.com

Tokenisation

Source_Text_Reference: Herzfeld (1900: 2)
Source_Translation_Reference: Herzfeld (1900: 3)
ParCorOE_Number: Mart 01.01.11

Prefield: On ðone forman dæg on geare, ðæt is on
Conc_Term: ærestan
Postfield: geohheldæg, eall cristen folc weorðiað

Fragment: On ðone forman dæg on geare, ðæt is on ðone ærestan geohheldæg, eall cristen folc weorðiað Cristes acennednesse.

Translation: On the first day of the year, that is on the first Yule-day, all Christian folk celebrate Christ's birth.

Text_intercalation:

Translation_intercalation:

Tagging

Inflectional_Category: sup: acc. sg. masc.
Lexical_Category: adjective
Gloss: first

Lemmatisation

nouns_A-C:
nouns_D-F:
nouns_G:
nouns_H-L:
nouns_M-O:
nouns_P-S:
names:
adjectives_A-F: ǣr 'early'
adjectives_G-R:
adjectives_S-Y:
verbs_A-E:
verbs_F-M:
verbs_N-Y:
adverbs_A_Y:
gramm_cat_A-Y:

Figure 3. Tokenisation, tagging and lemmatisation of ParCorOE

The source language texts and the target language translations draw on the editions cited below. Put in other words, the texts are not modernised *ad hoc*, but rather follow available PDE translations. The choice of the edition and translation at the present state of the research has been guided by copyright status. All texts are free of copyright.

The tokenisation, tagging or annotation of the texts that have been processed so far have pointed out some instances and areas of mismatch between the source language and the target language that hamper the word-for-word correspondence

Table 1. Word count by category and text

Category	Texts	Word count
Historical prose	*The Anglo-Saxon Chronicle*	25,000
Religious prose	*Ælfric's Homilies*	25,000
The Bible	*St. Mark*	25,000
Translations from Latin	*Benedictine Rule, Martyrology*	25,000
Legal prose	*Laws*	12,500
History	*Orosius*	12,500
Philosophy	*Boethius*	12,500
Medicine and herbaries	*Leechbook*	25,000

required for syntactic annotation. The solution proposed in the following sections is the implementation of an inter-syntax that has two functions: (i) to focus on the areas of syntactic divergence between the Old English text and its translation; and (ii) to map the source language tokens onto the target language ones by means of a set of tags that represent hierarchy and dependency. All the local instances of mismatch have been filed in an asymmetry bank.

4. Inter-syntax and asymmetry

The inter-syntax of ParCorOE can be described as an intermediate step between tokenisation, on the one hand, and glossing and annotation, on the other. The inter-syntax has two components, namely a structural description and a dependency tree. The structural description, in turn, comprises two labeled bracketing representations, one for the source language (extracted from the YCOE) and another one for the target language (based on the same categories as the YCOE). The dependency tree displays functional tags that relate dependent elements to their heads.

The definition of the scope of the inter-syntax requires the previous identification of the areas of divergence between the syntax of the source and the target language. This is tantamount to saying that the task is defined gradually and unidirectionally: in the search for mismatches, symmetrical and asymmetrical parings are considered, although the inter-syntax focuses on asymmetry. Old English is always the source language.

The mismatches that cause asymmetry between the source language text and its translation are described on structural grounds, but explained on a functional basis. For this reason, two sets of syntactic tags are required, categorial tags and functional tags, so that categorial tags account for hierarchy and functional tags for dependency. As has just been said, the structural description relies on the YCOE and the functional tags have been adapted from https://universaldependencies.org/u/dep/. While the structural description of the source and target language segment accounts for hierarchy and linearisation, the dependency tree aims to the relations that hold both in the source and the target language segment, as well as those that, applying in the source or the target only, constitute a description of variation on the synchronic axis or an explanation for change on the diachronic axis. The annotation in this framework, therefore, is couched in terms of an inter-syntax that maps the source language segment onto the corresponding target language segment. It must be stressed from this point that the terms *syntax* and *syntactic* are used comprehensively, so that morphological phenomena with impact on syntax, such as the assignment of morphological case, are considered.

To summarise, the structural change of the target language segment with respect to the source language segment can be derived from the tree diagrams or the labeled bracketing, but explanations based on phrasal and clausal functions require dependency relations. The areas of divergence between the source and the target language reflect a variety of syntactic phenomena that have been discussed from different angles in works like Visser (1963–73), Mitchell (1985), Denison (1993), Martín Arista (2000a, 2000b), Hogg and Fulk (2011), and Ringe and Taylor (2014).

The position held in this respect is that syntactic divergences can be captured in terms of alignment asymmetry. In this line, Scrivner (2015: 2) distinguishes the following schemas of alignment at word level: between two single words (one-to-one), between a single word and a multi-word unit (one-to-many), between a multi-word unit and a single word (many-to-one) and zero alignment. Alignment schemas are thus coached in terms of (a)symmetry: the number of slots in the source language is equal to or different from the number of slots in the target language. However, the asymmetry between the source and the target language cannot be restricted to quantity. Rather, it may be the result of marking, constituency, order or configuration. Markedness asymmetry involves more or less marked slots in the source language. Constituency asymmetry is the result of the presence of fewer slots in the source language. Order asymmetry is a consequence of a relative order in the target language different from the source language. Configuration asymmetry conveys a syntactic configuration substantially different from the source language. These four types of asymmetry may involve categories and relations and can be found at phrasal or sentential (inflectional phrase) level, although substantial changes to morphosyntactic configuration may often affect the inflectional phrase, whilst the type of asymmetry involving markedness is more likely to arise within units of the phrasal level. Consider the following example (quoted with the DOEC number).

(1) [LawWi 000500 (5)]
Gif ðæs geweorþe gesiþcundne mannan ofer þis gemot, þæt he unriht hæmed genime ofer cyngæs bebod & biscopes & boca dom, se þæt gebete his dryhtne C scillinga an ald reht;
If after this meeting, a nobleman presumes to enter into an illicit union, despite the command of the king and the bishop, and the written law, he shall pay 100 shillings compensation to his lord, in accordance with established custom.
(Attenborough 1922: 25)

In (1), several instances arise of the four types of asymmetry distinguished in this work. Beginning with markedness asymmetry, the verbal forms *geweorþe* 'please' and *gebete* 'pay' are inflected for the subjunctive, which is no longer possible in PDE by morphological means. In the noun phrase, the Old English dative *dryhtne* 'lord' requires a morphologically unmarked noun governed by a preposition, which is compatible with the definition of constituency asymmetry given above. The same can be said of the case-marked genitive noun *boca* in the noun phrase *boca dom* ‘judgement of books’. It is also of relevance for constituency asymmetry that the noun phrases *cyngæs bebod & biscopes & boca dom* ‘the command of the king and the bishop, and the judgement of books’ require a definite article functioning as determiner in PDE. Focusing on order asymmetry, the coordinate modifier in the genitive case in *cyngæs bebod & biscopes* cannot be extraposed in the PDE counterpart, thus ‘the king’s and the bishop’s command’. As regards configuration asymmetry, the verb *geweorþan* ‘to please’ selects the thematic roles Theme *ðæs* ‘of that’ (case-marked genitive) and Experiencer *gesiþcundne mannan* ‘noble man’ (inflected for the dative). Moreover, no introductory *hit* is found in the Old English fragment in sentence-initial position and the verb is complemented by a *þæt*-clause with the dependent verb in the subjunctive (*þæt he unriht hæmed genime* ‘to enter into an illicit union’), rather than by a *to*-infinitive clause realising a linked predication that shares the first argument with the matrix predication, as in ‘if it pleases a noble man to enter into an illicit union’.

Given this kind of evidence, the discussion that follows in the next section puts aside markedness asymmetry (which is rather predictable when it comes to comparing a more inflective and a less inflective language) to concentrate on constituency, order and configuration asymmetry.

5. The scope and components of the inter-syntax

This section identifies the areas that present alignment asymmetry and proposes an inter-syntax that incorporates the labeled bracketing of the YCOE but that crucially hinges around a dependency tree. Asymmetry phenomena are described with respect to the structural levels of the noun phrase and the inflectional phrase. While all the fragments have been extracted from the segment of PacCorOE that has been processed so far, they are headed by the DOEC text name and number.

In the noun phrase, a frequent asymmetry type is constituency asymmetry. It is often the case that the noun phrase is morphologically marked by means of case in the source language and by means of prepositional government in the target language. This can involve the accusative, the genitive, the dative and the instrumental, as in *lytle werede* ‘with a small force’ in (2).

(2) [ChronA (Bately) 036100 (871.30)]
Þa feng Ęlfred Ęþelwulfing his broþur to Wesseaxna rice, & þæs ymb anne monaþ gefeaht Ęlfred cyning wiþ alne þone here lytle werede æt Wiltune & hine longe on dæg gefliemde, & þa Deniscan ahton wælstowe gewald.

Then his brother Alfred, son of Æthelwulf, succeeded to the kingdom of Wessex. And one month later king Alfred fought with a small force against the entire host at Wilton, and for a long time during the day drove them off, and the Danes had possession of the place of slaughter. (Garmonsway 1972: 72)

Order asymmetry can also arise in the noun phrase. For instance, the genitive *ðæs cyninges* 'of the king' follows the nominal head *þegn* 'thane' in (3), in contradistinction to the proper name genitive *Ecgferðes* 'Ecgfrith's', which precedes the head.

(3) [Mart 5 (Kotzor) 018700 (Ma 7, B.5)]
He wæs Ecgferðes þegn ðæs cyninges, ac he forlet þa wæpna ond ða woruldlican wisan ond eode on þæt mynster ond wæs þær mæssepreost ond abbod.
He was a thane of King Ecgfrith, but he gave up his weapons and his secular life and joined the monastery and was a priest there and an abbot.
(Rauer 2013: 62)

In the inflectional phrase, the lack of do-support in the source language causes constituency asymmetry, as can be seen in (4), where *lifde* 'lived' is negated with the negative word *ne* 'not' only.

(4) [Or 3 036900 (11.82.18)]
Þagiet ne mehte se nið betux him twæm gelicgean, þeh heora na ma ne lifde þara þe Alexandres folgeras wæron, ac swa ealde swa hie þa wæron hie gefuhton: Seleucus hæfde seofon & seofontig wintra, & Lisimachus hæfde þreo & seofontig wintra.
Still the hostility between those two could not end, even though they were the only ones of Alexander's followers left, but, as old as they were, they went on fighting: Seleucus was seventy-seven years old and Lysimachus was seventy-three. (Godden 2016: 218)

The contraction of negative *bēon* 'to be', *habban* 'to have', *willan* 'will', *wītan* 'to know' and *āgan* 'ought to' also causes constituency asymmetry, given that the written representation takes one more slot in the target language than in the source language. The question is illustrated with respect to *witan* 'to know' in (5).

(5) [LawICn 003100 (6.2)]
Full georne hig witan, þæt hig nagon mid rihte þurh hæmedþingc wifes gemanan.
They know full well that they have no right to marry.
(Attenborough 1922: 22)

Another source of asymmetry in the inflectional phrase is the verbal conjugation of the source language, which does not have continuous, periphrastic or compound tenses, but presents a morphologically distinct subjunctive. This subjunctive translates as a modal periphrasis or as an indicative, as is the case with *læge* 'lay' and *bude* 'lived' in Example (6). This usually causes constituency asymmetry, but may also produce configuration asymmetry.

(6) [Or 1 008000 (1.14.5)]
He sæde þæt he æt sumum cirre wolde fandian hu longe þæt land norþryhte læge, oþþe hwæðer ænig mon be norðan þæm westenne bude.
He said that on one occasion he decided to find out how far the country extended northward, or whether anyone lived to the north of that uninhabited region. (Godden 2016: 36)

In the inflectional phrase, various phenomena cause order asymmetry, beginning with extraposition, which may involve a multiple subject or a relative clause, such as *þe ær wæs forslagen* 'which had been cut through before' in Example (7).

(7) [Æ LS (Edmund) 004700 (176)]
And his swura wæs gehalod þe ær wæs forslagen, and wæs swylce an seolcen þræd embe his swuran ræd, mannum to sweotelunge hu he ofslagen wæs.
And his neck, which had been cut through before, was healed and there was something like a red silken thread around his neck for men to remember how he had been killed. (Skeat 1881: 326)

Stranded prepositions also convey order asymmetry, as is the case with *Him com þa gangende to Godes engel* 'God's angel came to him walking' in (8).

(8) [Judg 006600 (13.3)]
Him com þa gangende to Godes engel, & cwæð ðæt hi sceoldon habban sunu him gemæne;
'And an angel of the Lord appeared to her, and said: Thou art barren and without children: but thou shalt conceive and bear a son. (*Douay Rheims Bible*: 475)

Order asymmetry also results from various fronting phenomena, including the fronting of the auxiliary verbs *bēon* and *habban*, illustrated in (9.a) and (9.b) respectively, as well as the fronting of nominative complements, such as *Themestocles* 'Themistocles' in (9.c), accusative and dative objects.

(9) a. [MkGl (Li) 000500 (1.4)]
Wæs iohannes in woestern gefulwade & bodade fulwiht hreownisses on forgefnisse synna.
John was in the desert baptizing, and preaching the baptism of penance, unto remission of sins. (Leonard 1881: 17)

b. [ChronE (Irvine) 026610 (658.3)]
Hæfde hine Penda adrefedne & rices benumene forþan þet he his swustor forlet.
Penda had expelled him and deprived him of his kingdom because he had repudiated his sister. (Garmonsway 1972: 32)

c. [Or 2 012700 (5.47.18)]
Themestocles hatte Atheniensa ladteow.
The leader of the Athenians was Themistocles. (Godden 2016: 127)

Two further characteristics of the source language bring about order asymmetry with respect to the target language, namely the V2 Rule, which places the subject after the verb in the context of an initial adverbial, as happens in *Þa comon þa menn* 'Then these men came' in (10.a); and the relatively generalised final verb in dependent clauses, which is illustrated by means of *gefultumade* 'may help' and *gehiersumade* 'may subject' in (10.b).

(10) a. [ChronA (Bately) 007400 (449.9)]
Þa comon þa menn of þrim mægþum Germanie, of Ealdseaxum, of Anglum, of Iotum.
These men came from three nations of Germany: from the Old Saxons, from the Angles, from the Jutes. (Garmonsway 1972: 12)

b. [ChronA (Bately) 032600 (853.1)]
Her będ Burgred Miercna cyning & his wiotan Ęþelwulf cyning þæt he him gefultumade þæt him Norþwalas gehiersumade.
In this year Burgred, king of Mercia, and his councilors besought king Æthelwulf that he would help them to subject the Welsh.
(Garmonsway 1972: 66)

The existence of double negation in Old English, comprising both the phrasal and the sentential levels, causes configuration asymmetry with respect to the target language. This may be due not only to the negative words themselves, but also to the lack of *do*-support in Old English. For instance, *ne* negates at sentential level and *naht* at phrasal level in (11).

(11) [Mart 5 (Kotzor) 023900 (Ma 21, B.4)]
Ond on sumum þara mynstra þe he ofergeseted wæs þa broðor him woldon sellan attor drincan forðon þe hi ne mostan for him naht unalyfedlices begangan.
And in one of the monasteries over which he presided, the brothers tried to give him poison to drink, because with him they were not allowed to do anything illicit. (Rauer 2013: 71)

Various phenomena that may be grouped under the heading of omission result in constituency asymmetry with respect to the target language. The status of the elements which are required in the target language considerably varies. In Old English, the formal subjects *there* and *it* are not compulsory, as is shown in (12.a) and (12.b), respectively.

(12) a. [Mart 2.1 (Herzfeld-Kotzor) 016700 (De 0, A.1)]
On þam twelftan monðe on geare byð an ond XXX daga.
'There are thirty-one days in the twelfth month of the year.'
(Rauer 2013: 223)

b. [Mart 2.1 (Herzfeld-Kotzor) 000800 (Ju 24, B.3)]
Þonne gelympeð þæt wundorlice on þæs sumeres sungihte on mydne dæg þonne seo sunne byð on þæs heofones mydle, þonne nafað seo syl nænige sceade.
Then amazingly, it happens during the summer solstice at midday, that when the sun is in the middle of the sky, the column does not have any shadow. (Rauer 2013: 125)

Fully lexical subjects can also be omitted, as is illustrated in (13.a), but the omission of lexical verbs is restricted, as a general rule, to *bēon*, as is presented in (13.b).

(13) a. [Mart 5 (Kotzor) 094900 (Au 30, A.2)]
Wæs in ðære ceastre þe is nemned Tubsocensi.
He lived in the city which is called Thibiuca. (Rauer 2013: 171)

b. [ÆCHom I, 7 003800 (234.79)]
Swutel is þæt ða tungelwitegan tocneowon crist. soðne man: þa ða hi befrunon. hwær is se ðe acenned is.
It is manifest that the astrologers knew Christ to be a true man, when they inquired, 'Where is he who is born?' (Thorpe 1844: 107)

As in PDE, the subject of a coordinate construction is, as a general rule, omitted in Old English. However, the object of a construction of coordination is left unexpressed far more often in the source language than in the target language of the corpus. An instance of the omission of the object of a coordinate construction is given in (14), in which the object *hine* 'him' is shared by *gebringan* 'bring' and *belucan* 'lock up'.

(14) [Bo 001200 (1.7.23)]
þa þæt ongeat se wælhreowa cyning ðeodric, þa het he hine gebringan on carcerne & þærinne belucan.
When that cruel king Theoderic discovered this, he ordered him to be put into a prison and locked up there. (Godden et al. 2009: 5)

The omission of a complementiser, such as the one depending on *secge* 'say' in (15), involves constituency asymmetry too.

(15) [Mk (WSCp) 009900 (3.29)]
Soþlice ic eow secge, se þe ðone halgan gast bysmerað, se næfð on ecnysse forgyfenesse, ac bið eces gyltes scyldig.
But he that shall blaspheme against the Holy Ghost, shall never have forgiveness, but shall be guilty of an everlasting sin. (Leonard 1881: 28)

Complex conjunctions, such as *mid þæm þe* 'when' in (16.a), and complex relatives, like *þær ðær* 'where' in (16.b), take fewer slots in the target than in the source language. This is also a matter of constituency asymmetry.

(16) a. [Or 2 003400 (2.39.6)]
Hi swaþeah heora unðances mid swicdome hie begeaton, mid þæm þe hie bædon þæt hie him fylstan mosten ðæt hie hiera godum þe ieð blotan mehten: þa hie him þæs getygðedon, þa hæfdon hi him to wifum, & heora fæderum eft agiefan noldon.
The Romans got them anyway by trickery, despite the opposition of the fathers, when they asked the Sabines to help them sacrifice, to their gods. When the Sabines agreed to this, the Romans seized the daughters as their wives and would not return them to their fathers. (Godden 2016: 107)

b. [ChronE (Irvine) 031600 (679.1)]
Her man ofsloh Ælfwine be Trentan þær ðær Egferð & Æðelred gefuhton.
In this year Ælfwine was slain beside the Trent, at the place where Ecgfrith and Æthelred fought. (Garmonsway 1972: 38)

Impersonal verbs require one more argument (a formal subject) in the target language in order to realise the Patient, Recipient or Beneficiary, which, in the source language is case marked accusative, as *hine* 'him' in (17.a) or dative, like *him* 'them' in (17.b). These are instances, therefore, of constituency asymmetry.

(17) a. [Bo 045300 (16.39.20)]
Hine lyste eac geseon hu seo burne, hu lange, & hu leohte be þære oðerre.
He wanted also to see how it burnt, how long and how brightly in comparison with the other city. (Godden et al. 2009: 26)

b. [CP 123900 (36.261.3)]
Him is to secgeanne ðæt hie unablinnendlice geðencen hu monig yfel ure Dryhten & ure Alisend geðolode mid ðam ilcan mannum ðe he self gesceop, & hu fela edwites & unnyttra worda he forbær, & hu manige hleorslægeas he underfeng æt ðæm ðe hine bismredon.
They are to be told to consider incessantly how many evils our Lord and Redeemer suffered among the same men whom he himself had created, and how much reproach and how many vain words he endured, and how many blows he received from his revilers. (Sweet 1881: 260)

Reflexives with intransitive verbs also cause constituency asymmetry. They take one more argument in the source than in the target language, either case-marked accusative, such as *hine* 'himself' in (18.a), or dative, like *him* 'themselves' in (18.b).

(18) a. [Mk (WSCp) 016800 (5.22)]
& ða com sum of heahgesamnungum Iairus hatte, & þa he hine geseah he astrehte hine to his fotum.

And there cometh one of the rulers of the synagogue named Jairus: and seeing him, falleth down at his feet. (Leonard 1881: 34)

b. [Or 1 029500 (10.29.12)]
Hi þa þæt lond forleton, & him hamweard ferdon.
Then they left that land and went home. (Godden 2016: 79)

As can be seen in (19), there is configuration asymmetry between an instance of the verb *hātan* 'to be called', which occurs in active sentences in the source language, and the corresponding passive in the target language.

(19) [Or 1 003600 (1.11.1)]
Seo Ægyptus þe us near is, be norþan hire is þæt land Palastine, & be eastan hiere Sarracene þæt land & be westan hire Libia þæt land, & be suþan hire se beorg þe mon hæt Climax.
The part of Egypt that is nearer to us has Palestine to the north, and to the east is the Saracen land, and to the west is Libya, and to the south is a mountain called Climax. (Godden 2016: 30)

Example (19) also illustrates the configuration asymmetry holding with respect to the indefinite pronoun *mon* 'someone' in the source language and in the target language, which frequently calls for a passive.

Considering the areas of asymmetry presented in this section, the inter-syntax of ParCorOE comprises a set of dependency relations that links the hierarchy and linearisation of the source language representation to the target language. Hierarchy and linearisation are displayed by labeled bracketing representations of the type adopted by the YCOE, which is illustrated in Figure 4, representing *Aristoteles hit gerehte on þære bec þe Fisica hatte* 'Aristotle explained it in the book which is entitled Physics' (Godden et al. 2009). Figure 5 shows the structural description of the target language segment.

```
(IP-MAT-SPE (NP-NOM (NR^N Aristoteles)
            (NP-ACC (PRO^A hit))
            (VBPS gerehte)
            (PP (P on)
                (NP-DAT (D^D+t+are) (N^D bec)
                        (CP-REL-SPE (WNP-NOM-1 0)
                                    (C+te)
                                    (IP-SUB-SPE (NP-NOM *T*-1)
                                                (NP-PRD (NR Fisica))
                                                (VBD hatte))))))
      (..)) (ID coboeth, Bo:40.140.8.2794))
```

Figure 4. Source language labeled bracketing from the YCOE

```
(IP-MAT-SPE (NP (NR Aristotle)
                (VBPS explained)
                (NP (PRO it))
                (PP (P in)
                    (NP (D the) (N book)
                        (CP-REL-SPE (WNP-1 0)
                                    (C that)
                                    (IP-SUB-SPE (NP *T*-1)
                                                (VBD is entitled))))))
                                                (NP-PRD (NR Physics))
```

Figure 5. Target language bracketing based on the YCOE

With the relations of hierarchy and linearisation that arise in Figures 4–5, the representation of dependency put forward in Figure 8 has two main properties: explicitness and compatibility with the labeled bracketing provided by the YCOE. At the present stage, the tags and relations of dependency include the ones listed in Figure 6.

DET	Determiner of
MOD	Modifier of
QUANT	Quantifier of
SUB	Subject of
OBJ	Object of
EXPLSUBJ	Expletive Subject of
EXPLOBJ	Expletive Object of
COMP	Complement of
ADV	Adverbial of
ADP	Adposition to
GVN	Governed by

Figure 6. Dependency tags and relations

These tags and relations rely on a concept of dependency that involves argumenthood (Subject of, Object of, Expletive Subject of, Expletive Object of), complementation (Complement of), government (Governed by) and obligatoriness (Determiner of, Modifier of, Adposition to).

The annotation procedure calls for the manual selection of the relevant tag and relation from a scroll-down menu on a database implemented in Filemaker. This procedure, which is fully manual at the moment, will be partly automatised once a larger segment of the corpus has been processed, so that certain associations between lexical items and dependency tags and relations can be predicted on a statistical basis.

The inter-syntactic representation shown in Figure 7 resorts to graph theory in order to increase searchability and favour visualisation. In graph theory, a graph consists of vertices and tokens. Binary graphs relate two tokens to each other. In directed graphs the relationship holds in one direction only. In the inter-syntactic representation presented in Figure 8, each of the two constituents between which a relation of dependency holds is a node. The arc represents the dependency type. It is directed, which means that it points from the dependent to the head of the dependency relation. In Figure 7, higher arcs represent main sentence relations, while lower arcs are used for dependent clausal relations. The structural level of the clause is displayed over the linguistic segment and the level of the phrase is represented under the linguistic segment.

Graphs are generated with RAWGraphs from an Excel spreadsheet displaying the following columns: dependent, dependent token number, head, head token number, dependency tag and structural level. The data filed in the Excel spreadsheet is then imported to Filemaker, which allows for searches by lexical item (e.g. *hit*) and by dependency relation. Both types of searches can be simple (e.g. GVN) or complex (e.g. GVN and phrase level)

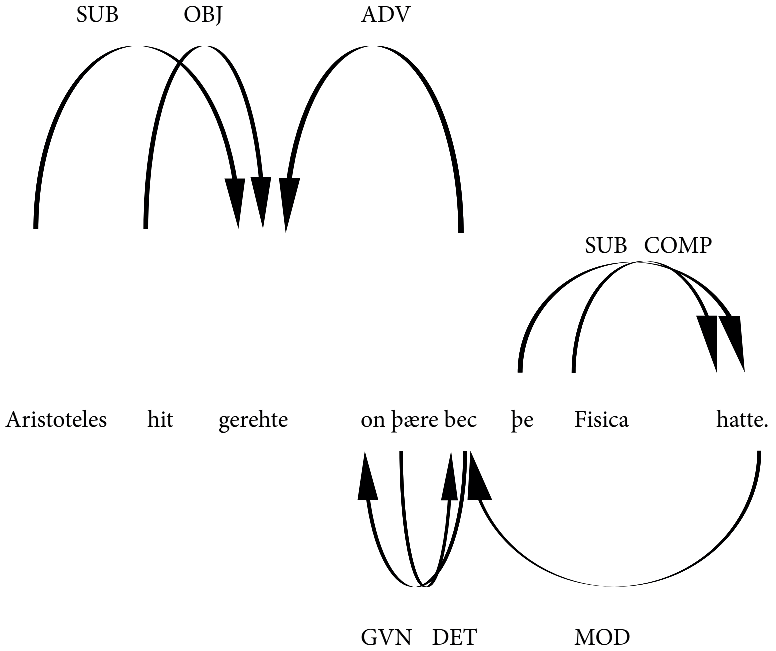


Figure 7. Inter-syntactic representation by means of a dependency tree

The comparison of the structural description in Figures 4, 5 and the dependency tree in Figure 7 indicates areas of stability as well as areas of change: whereas the phrasal and clausal relations of dependency remain, thus SUB, OBJ, ADV, COMP, GVN, DET, and MOD; change concentrates on the areas of morphological case (the accusative *hit*, the dative *bec* and the nominative *Fisica* are marked in the source language version) and linearisation (the Object *hit* and the Complement *Fisica* precede their respective verbs in the Old English text). The examples discussed above also display changes to the relations of dependency presented in Figure 7, which stresses the need for an inter-syntactic model that specifies both clausal and phrasal dependency relations.

6. Conclusion and further research

This chapter has addressed the question of how to devise and implement an inter-syntactic model for ParCorOE that equips the corpus with syntactic annotation compatible with word alignment. The fact that ParCorOE is a corpus of intra-linguistic translation has guided this solution. On the one hand, two diachronic stages of English are compared, which predicts a considerable amount of convergence between the source language and the target language. On the other, alignment at word level calls for a level of correspondence that excludes local mismatches. The balanced solution described in this chapter restricts the syntactic annotation to the areas of divergence between the source and the target language.

Syntactic divergences have been explained on the basis of asymmetry and with respect to all structural levels: markedness asymmetry (generalised); constituency asymmetry (noun phrase, reflexive pronominal phrase, inflectional phrase, complementiser, conjunction); order asymmetry (noun phrase, prepositional phrase, inflectional phrase, adverbial phrase); and configuration asymmetry (noun phrase, inflectional phrase both active and passive, complementiser). The inter-syntax comprises the structural description of the source and the target language segments (with YCOE labels, in order to guarantee compatibility) as well as a dependency tree. The comparison of the structural description the dependency tree constitutes a historical micro-grammar, in the sense that it distinguishes syntactic stability from change, including the change of dependency relations. The dependency tree is represented by means of graph theory so as to increase explicitness and to facilitate searchability. Overall, ParCorOE can be searched for text, fragment and token and, above all, for lexical items, morphological categories and dependency relations.

This model of inter-syntax has been found adequate to represent all the local mismatches that have arisen so far, but more research will be necessary as the corpus processing advances. In this line, the identification of more areas of asymmetry is pending. It also remains for future research to determine whether alignment at word level may increase the exhaustivity of annotation and boost automation: although the syntactic annotation procedure is manual at the moment, it is expected that it will be partially automatised in the near future.

The following abbreviations are used in this section

ind.	indicative
sub.	subjunctive
imp.	imperative
infl. inf.	inflected infinitive
pres. part.	present participle
pa. part.	past participle
pres.	present
pret.	preterite
sg.	singular
pl.	plural
dat.	dative

Funding

This research has been funded through the grant FFI2017-83360P, which is gratefully acknowledged.

References

Attenborough, Frederick L. (ed. and trans.). 1922. *The Laws of the Earliest English Kings*. Cambridge: Cambridge University Press.

Biber, D. 2007. "Representativeness in corpus design", in: W. Teubert and R. Krishnamurthy (eds.), *Corpus linguistics: Critical concepts in linguistics* (Vol. II). London: Routledge. 134–165.

Cameron, Angus, Ashley C. Amos, and Antonette diPaolo Healey (eds.). 2018. *The Dictionary of Old English in Electronic Form A-I*. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Denison, David. 1993. *English Historical Syntax: Verbal Constructions*. London: Longman.

Enrique-Arias, Andrés. 2013. "On the usefulness of using parallel texts in diachronic investigations: insights from a parallel corpus of Spanish medieval Bible translations." In *New Methods in Historical Corpora*, ed. by Paul Durrell, Martin Scheible, Silke Whitt, and Richard J. Bennett, 105–116. Tübinguen: Gunter Narr.

Faaß, G. 2017. "Lexicography and corpus linguistics." In *The Routledge Handbook of Lexicography*, ed. by Pedro A. Fuertes-Olivera, 123–137. Abingdon: Routledge. https://doi.org/10.4324/9781315104942-9

Fernández Cuesta, Julia, Nieves Rodríguez Ledesma, and Gloria Alvárez Benito (eds. and trans.). 1997. *Prosa anglosajona*. Sevilla: Universidad de Sevilla.

García Fernández, Laura. 2018. "Preterite-present verb lemmas from a corpus of Old English." In *Verbs, Clauses and Constructions: Functional and Typological Approaches*, ed. by Pilar Guerrero Medina, Roberto Torre Alonso, and Raquel Vea Escarza, 59–76. Newcastle: Cambridge Scholars Publishing.

Garmonsway, George N. (ed. and trans.). 1972. *The Anglo-Saxon Chronicle*. London: Dent & Sons LTD.

Godden, Malcom. 2016. *The Old English History of the World. An Anglo-Saxon Rewriting of Orosius*. Cambridge, Massachusetts: Dumbarton Oaks.

Godden, Malcom, Susan Irvine (eds.), with Mark Griffith, and Rohini Jayatilaka. 2009. *The Old English* Boethius. Volume II. Oxford: Oxford University Press.

Hanks, Patrick. 2012. "Corpus Evidence and Electronic Lexicography." In *Electronic Lexicography*, ed. by Sylviane Granger, and Magali Paquot, 57–82. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199654864.003.0004

Healey, A. diPaolo (ed.) with John P. Wilkin, and Xin Xiang. 2004. *The Dictionary of Old English Web Corpus*. Toronto: Dictionary of Old English Project, Centre for Medieval Studies, University of Toronto.

Heid, Ulrich. 2008. "Corpus linguistics and lexicography." In *Corpus Linguistics. An International Handbook* (Volume 1), ed. by Anke Lüdeling and Merja Kytö, 132–153. Berlin: Mouton de Gruyter.

Hogg, Richard M. and Robert D. Fulk. 2011. *A Grammar of Old English. Volume 2: Morphology*. Oxford: Blackwell.

Johnson, B. 2009. Using the Levenshtein algorithm for automatic lemmatization in Old English. MA Thesis, The University of Georgia.

Krause, Thomas, and Amir Zeldes. 2016. ANNIS3: "A new architecture for generic corpus query and visualization." *Literary and Linguistic Computing* 31(1): 118–139. https://doi.org/10.1093/llc/fqu057

Kübler, Sandra, and Heike Zinsmeister. 2014. *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury.

Leonard, Henry C. (ed. and trans.). 1881. *A Translation of the Anglo-Saxon Version of St. Mark's Gospel*. London: James Clarke & Co.

Lu, Xiaofei. 2014. *Computational Methods for Corpus Annotation and Analysis*. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-8645-4

Martín Arista, Javier. 2000a. "Sintaxis medieval inglesa I: complementación, caso y sintaxis verbal." In *Lingüística histórica inglesa*, ed. by Isabel de la Cruz Cabanillas and Javier Martín Arista, 224–312. Barcelona: Ariel.

Martín Arista, Javier. 2000b. "Sintaxis medieval inglesa II: funciones, construcciones y orden de constituyentes." In *Lingüística histórica inglesa*, ed. by Isabel de la Cruz Cabanillas and Javier Martín Arista, 313–377. Barcelona: Ariel.

Martín Arista, Javier. 2013. Nerthus. "Lexical Database of Old English: From word-formation to meaning construction." Lecture delivered at the Research Seminar, School of English, University of Sheffield.

Martín Arista, Javier. 2017a. "Toward a parallel corpus of Old English prose. Preliminary questions and initial design." Lecture delivered at the Departmental Colloquium Series at the Department of Language and Linguistic Science, University of York.

Martín Arista, Javier. 2017b. "The Nerthus Project at the crossroads. From lexical database to parallel corpus of Old English." Lecture delivered at the 2017 International Conference of SELIM, held at the University of Málaga.

Martín Arista, Javier. 2018. "The design and implementation of a pilot parallel corpus of Old English." In *Aspects of Medieval English Language and Literature*, ed. by Michiko Ogura and Hans Sauer, 111–134. Berlin: Peter Lang.

Martín Arista, Javier, and Ana E. Ojanguren López. 2018. "Doing Electronic Lexicography of Old English with a Knowledge-Base." Workshop delivered at the CLASP Project (University of Oxford).

McEnery, Tony. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, Tony. 2003. "Corpus linguistics." In *Oxford handbook of computational linguistics*, ed. by Ruslan Mitkov, 448–463. Oxford: Oxford University Press.

McEnery, Tony & Richard Xiao. 2007a. Parallel and comparable corpora: What are they up to? *Incorporating Corpora: Translation and the Linguist. Translating Europe*. Clevedon: Multilingual Matters. https://doi.org/10.21832/9781853599873-005

McEnery, Tony, and Richard Xiao. 2007b. "Parallel and Comparable Corpora-The State of Play." In *Corpus-Based Perspectives in Linguistics*, ed. by Yuji Kawaguchi, Toshihiro Takagaki, Nobuo Tomimori, and Yoichiro Tsuruga, 131–146. Amsterdam: John Benjamins. https://doi.org/10.1075/ubli.6.11mce

Metola Rodríguez, Darío. 2017. "Strong Verb Lemmas from a Corpus of Old English. Advances and issues." *Revista de Lingüística y Lenguas Aplicadas* 12: 65–76. https://doi.org/10.4995/rlyla.2017.7023

Mitchell, Bruce. 1985. *Old English Syntax* (2 vols.). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198119357.001.0001

Novo Urraca, Carmen, and Ana E. Ojanguren López. 2018. "Lemmatising Treebanks. Corpus Annotation with Knowledge Bases." *RAEL* 17: 99–120.

Pintzuk, Susan, and Leendert Plug (comp.). 2001. *The York-Helsinki Parsed Corpus of Old English Poetry*. Department of Language and Linguistic Science, University of York.

Rauer, Christiane. (ed. and trans.). 2013. *The Old English Martyrology*. Cambridge: D. S. Brewer.

Ringe, Don, and Ann Taylor. 2014. *A Linguistic History of English Volume II: The Development of Old English*. Oxford: Oxford University Press.

Rissanen, Matti, Merja Kytö, L. Kahlas-Tarkka, Matti Kilpiö, Saara Nevanlinna, Irma Taavitsainen, Tertu Nevalainen and Helena Raumolin-Brunberg (comp.). 1991. *The Helsinki Corpus of English Texts*. Department of Modern Languages, University of Helsinki.

Schierholz, Stefan J. 2015. "Methods in Lexicography and Dictionary Research." *Lexikos*: 25(1): 323–352.

Scrivner, Olga. 2015. "Tools for Digital Humanities: Parallel Corpus and Visualization." Paper presented at the Conference Corpora 2015, held at Saint-Petersburg, Russia.

Skeat, Walter W. (ed.) 1881. *Ælfric's Lives of Saints. Volume I*. Oxford: Oxford University Press.

Sweet, Henry (ed.). 1881. *King Alfred's West-Saxon Version of Gregory's Pastoral Care*. London: Trübner & Co.

Taylor, Ann, Anthony Warner, Susan Pintzuk and Frank Beths (comp.) 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. Department of Language and Linguistic Science, University of York.

The Holy Bible Translated from the Latin Vulgate (Douay Rheims Version) 1971 (1899). *Rpt. Rockford*, Illinois: Tan books.

Thorpe, Benjamin. (ed. and trans). 1844. *The Homilies of the Anglo-Saxon Church. Volume I*. London: Red Lion Court.

Tío Sáenz, Marta. 2015. "The Regularization of Old English Weak Verbs". *Revista de lingüística y lenguas aplicadas* 10: 78–89.

Visser, Ferdinand. 1963–1973. *An Historical Syntax of the English Language* (4 vols.). Leiden: Brill.

Semantic textual similarity based on deep learning

Can it improve matching and retrieval for Translation Memory tools?

Tharindu Ranasinghe, Ruslan Mitkov, Constantin Orăsan and Rocío Caro Quintana

This study proposes an original methodology to underpin the operation of new generation Translation Memory (TM) systems where the translations to be retrieved from the TM database are matched not on the basis of Levenshtein (edit) distance but by employing innovative Natural Language Processing (NLP) and Deep Learning (DL) techniques. Three DL sentence encoders were experimented with to retrieve TM matches in English-Spanish sentence pairs from the DGT TM dataset. Each sentence encoder was compared with Okapi which uses edit distance to retrieve the best match.[1] The automatic evaluation shows the benefit of the DL technology for TM matching and holds promise for the implementation of the TM tool itself, which is our next project.

Keywords: machine translation, translation memory, deep learning, Okapi, textual similarity, semantic similarity

1. Introduction

The Translation Memory (TM) tools revolutionised the work of professional translators and the last three decades have seen dramatic changes in the translation workflow. The concept of TM systems is, essentially, a simple one: the translator has access to a database of previous translations (referred to as a *Translation Memory database*), which he or she may consult, usually on a sentence-by-sentence basis, in order to find something similar enough to the current sentence to be translated. One of the most important functions of TM systems is their ability to match a

[1] The Okapi Framework is a cross-platform and free open-source set of components and applications that offer extensive support for localising and translating documentation and software. It is available on https://okapiframework.org/. We specifically used the Rainbow application available in the framework which allows bulk matching and retrieval from a translation memory.

https://doi.org/10.1075/btl.158.04ran

sentence to be translated against the database. Where there is an exact match, the user can simply reuse the corresponding target language segment. If there is no exact match, the system displays one or more close matches with the differences which the translator can edit to deliver the correct translation thus saving time to complete the translation job (Mitkov 2020).

Most commercial TM systems are able to quantify the goodness of the match with a 'fuzzy score' or 'fuzzy match'. While most systems operate on character-string similarity, some of them incorporate additional heuristics such as formatting or indicative words but no specific details are revealed due to commercial reasons. The character-string similarity is also referred to as 'string edit distance', simply 'edit distance' or more formally 'Levenshtein distance' (Levenshtein 1966). The edit distance is the minimal number of insertions, deletions or substitutions necessary to change one string of characters into another. For instance, in order to convert *memory* into *memories* one needs one deletion (y) and three insertions (i, e and s). While edit distance has been extensively used in a variety of applications in addition to TM systems, such as speech or image processing, it has a number of shortcomings.

While TM systems have revolutionised the translation industry, these tools are far from being perfect. A serious shortcoming has to do with the fact that the (fuzzy) matching algorithm of most commercial TM systems is based on edit distance and no language processing is employed. Also, segmentation is carried out at sentence level which makes it more difficult to suggest matches. Among the first ones to discuss the shortcomings were Macklovitch and Russell (2000) who maintained that Translation Memory technology was limited by the rudimentary techniques employed for approximate matching. They cite Planas and Furuse's (1999) comments that unless a TM system can perform morphological analysis, it will have difficulty recognising that sentence (3) below is more similar to input sentence (1) than (2) is.

(1) The wild child is destroying his new toy.

(2) The wild chief is destroying his new tool.

(3) The wild children are destroying their new tool.

Most commercial TM systems have similarity based on the number of shared character (or more generally, edit distance between strings) and will conclude that (1) and (2) are more similar as they differ only in four letters, whereas (1) and (3) differ in nine letters. A better way forward would be for TM systems to allow inflectional variants of a word to match (e.g. *child, children*).

The above shortcomings paved the way to the development of second-generation TM tools which had some language processing capabilities such as grammatical pattern recognition and performed limited segmentation at sub-sentence level. However, there are only a few commercially available second-generation TM systems such as *Similis* (Planas 2005), *Translation Intelligence* (Grönroos and Becks, 2005) and *Meta Morpho* TM system, Morphologic (Hodász and Pohl, 2005). *Similis* (Planas, 2005) performs linguistic analysis in order to split sentences into syntactic chunks or syntagmas, making it easier for the system to retrieve matches. Mitkov (2005) noted that another shortcoming has to do with the fact that TM systems usually perform matching at sentence level and do not go down to subsentential level (i.e., they do not process and compare clauses which might have already been translated) or up to 'suprasentential' level (i.e. they do not process sentences composed of already translated sentences). He illustrated the former shortcoming by the example that if a translator has already used a TM system to translate the complex sentence (4) and if he/she has to translate either the simple sentence (clause) (5) or the simple sentence (6), due to inability of TM tools to perform parsing or clause splitting, a sufficiently high match will not be shown and the translator cannot benefit from the fact that part of the complex sentence has already been translated. Mitkov illustrated the latter shortcoming by reversing the example: if the simple sentences (7) and (8) already had their translations in the TM database, but now the complex sentence (9) were to be translated, the TM tool would not score a high enough match to offer the translations of each of the clauses. He noted that former shortcoming is partially covered by second-generation TM systems. Mitkov went on to make the point that a new generation of TM was needed which would be able to process segments not only at surface or syntactic level, but also at semantic level. By way of illustration, if a current TM tool has been used to translate (10), then due to the insufficiently high match, translating (11) would have to be from scratch even though (10) and (11) are the same sentences semantically.

(4) Select 'Shut down' from the menu and click on 'Shut down'.

(5) Select 'Shut down' from the menu.

(6) Click on 'Shut down'.

(7) Select 'Shut down' from the menu.

(8) Click on 'Shut down'.

(9) Select 'Shut down' from the menu and click on 'Shut down'.

(10) Microsoft developed Windows 10.

(11) Windows 10 was developed by Microsoft.

Therefore, Mitkov (2005) proposed the so-called *third-generation of Translation Memory tools* which are capable of performing 'semantic matching'. He also proposed the integration of a TM system with WordNet in order to be able to identify synonyms: if for example the sentence (12) has been translated, but after that the sentence (13) is under consideration, then *vacuum cleaner* would be identified as a synonym of *hoover* and both sentences would be regarded as synonymous. Finally, he suggested that the next generation of TM systems be capable of identifying named entities and temporal information and to semantically normalise sentences such as (14) and (15) into the semantic classes <person-m> flew to <city> on <date> so that an existing translation in the translation memory would be re-used/edited for sentences with a high 'semantic' match.

(12) Unplug the hoover.

(13) Unplug the vacuum cleaner.

(14) John Smith flew to Brussels on February 3rd.

(15) Dr Johnson flew to Rome on 7 January 2006.

Further to Mitkov's (2005) observations on the future of TM systems, Pekar and Mitkov's (2007) work went beyond string matching and partial syntactic analysis and developed an algorithm which matches sentences which are similar not only syntactically, but also semantically. While this promising work was the first example of matching algorithms for future third-generation TM systems, the described approach was not deemed suitable for practical applications due to its very long processing time (it could take days to compare matches). Later work included Timonera and Mitkov (2015) who experimented with clause splitting and paraphrasing, seeking to establish whether these NLP tasks would improve the performance of TM systems in terms of matching.

Further work towards the development of third-generation TM systems included the more recent studies conducted by members of the Research Group in Computational Linguistics, University of Wolverhampton (Gupta et al. 2016a; Gupta et al., 2016b) who experimented with paraphrasing the TM with a view to securing more matches. The authors sought to embed information from PPDB, a database of paraphrases (Ganitkevitch et al., 2013), in the edit distance metric by employing dynamic programming (DP)[2] as well as dynamic programming and greedy approximation (DPGA).[3] The same research group in Wolverhampton have

[2] Segments may contain numerous words and phrases that can be paraphrased, hence it is not practical to generate all the possible variants for a given input segment and compare them to all the segments in the TM. This is particularly important given that the number of variants increases exponentially with the length of the input segment. For this reason, it was necessary to find an efficient way to decide which paraphrases to consider. The *dynamic programming* approach splits the problem into subproblems and solves each of them independently, combining their solutions in an optimal way.

[3] The *greedy approximation* is a technique which may not lead to the best possible solution because it takes decisions on the basis of the information available at a certain point, but in most of the cases it is a solution which is good enough and which can be obtained in a reasonable amount of time.

been experimenting with and proposing new semantic textual similarity (STS) metrics with a view to incorporating them in new generation TM systems. Gupta et al. (2014) developed a machine learning approach for semantic similarity and textual entailment based on features extracted using typed dependencies, paraphrasing, machine translation, evaluation metrics, quality estimation metrics and corpus pattern analysis which performed well with reported 0.711 Pearson correlation[4] for the semantic relatedness task and 78.52% accuracy for the textual entailment task (Marelli et al., 2015). This similarity method was experimented with to retrieve the most similar segments from a translation memory but the evaluation results showed that the approach was too slow to be used in a real scenario and that it did not perform better than edit distance (Gupta et al., 2014b). Latest work of the group includes Ranasinghe et al.'s (2019b) study which put forward a new semantic similarity metric based on Siamese networks.[5] More specifically, the authors employed a Gated Recurrent Unit (GRU)[6] which is a recurrent neural network architecture similar to Long Short-Term Memory (LSTM)[7] but has fewer parameters and which performs better on smaller datasets than LSTM. In particular, GRU-based architectures handle active-passive equivalence better than other models and the experiments report that Ranasinghe et al.'s (2019b) semantic similarity metric outperforms other state of the art STS metrics.

[4] The Pearson correlation has a value between −1 and 1 which represents the extent to which two variables are linearly related. Pearson correlation of −1 indicates that the two variables are perfectly negatively linearly related. Pearson correlation of 0 corresponds to the case when the two variables do not have any linear relation and Pearson correlation of 1 means that two variables are perfectly positively linearly related.

[5] Siamese networks are a special class of Neural networks. These networks contain two or more identical sub-networks. The networks are identical in the sense that they have the same configuration with the same parameters and weights. In addition, parameter updating is mirrored across these sub-networks. They are popular among tasks that involve finding similarity or a relationship between two comparable things.

[6] GRU can be considered as a variation on the LSTM because both are designed similarly and, in some cases, produce equally good results. GRU has a less complex structure with fewer gates than LSTM which makes GRU computationally more efficient.

[7] LSTMs are a special kind of Recurrent Neural Network (RNN). RNN is widely used neural network architecture for Natural Language Processing (NLP). An RNN maintains a memory based on history information, which enables the model to predict the current output conditioned on long distance features: which is the reason why RNNs perform well in NLP tasks. Although an RNN can learn dependencies, it can only learn about recent information. LSTM can help to solve this problem as it can understand context along with recent dependency. LSTM networks are similar to RNNs with one major difference that hidden layer updates are replaced by memory cells. This makes them better at finding and exposing long range dependencies in data which is imperative for sentence structures.

This study seeks to develop the operational basis of a new generation of Translation Memory (TM) systems where the sentences to be translated against the TM database will be matched not on the basis of the traditional Levenshtein (edit) distance fuzzy matching metric but by employing innovative Natural Language Processing (NLP) and Deep Learning (DL) methodology.

The new generation TM system which will be underpinned by our novel matching technology, will be able to return for revision not only sentences from the TM database which are syntactically similar to the sentences to be translated, but also *semantically close* ones. As an illustration, if the sentences (16) (17) and (18) are fed into any of the commercial TM systems, sentences (16) and (17) would be rated to be more similar than (16) and (18). The TM tool to be developed will be expected to return higher similarity for (16) and (18) than (16) and (17) as (16) and (18) are semantically very close, as opposed to (16) and (17).

(16) I like Madrid which is such an attractive and exciting place.

(17) I dislike Madrid which is such an unattractive and unexciting place.

(18) I love Madrid as the city is full of attractions and excitements.

More specifically, in this study we explore several deep learning based semantic textual similarity methods to improve performance of future translation memory tools. We experiment with several word embeddings and sentence embeddings on DGT's translation memory dataset.

2. Methodology

Over the years, researchers have proposed many algorithms to calculate string similarity. The most common algorithm is Levenshtein distance (Levenshtein 1966) which may also be referred to as *edit distance* although that term may also denote a larger family of distance metrics known collectively as edit distance. As we explained before the Levenshtein distance between two strings is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one string into the other (Levenshtein 1966). There are other variants to this such as Damerau–Levenshtein distance which allows insertion, deletion, substitution, and the transposition of two adjacent characters as edits (Damerau 1964) and longest common subsequence (LCS) distance which allows only insertion and deletion, not substitution as edits. There are a few variants of the Levenshtein Distance that considers multiple-characters differentiating from the single-character approach in the Levenshtein Distance. Sørensen–Dice Coefficient is such an algorithm that works by comparing the number of identical character pairs between the two strings (Sørensen 1948; Dice 1945). However, none of these algorithms can capture the *semantic* similarity between two strings when the words are different.

There are a number of approaches which rely on information from Wordnet to calculate similarity between sentences. For example, Wali et al. (2017) propose an approach which exploits the WordNet ”is a” taxonomy to estimate the semantic similarity between words, which in turn is used to determine the similarity between sentences. Even though their method provides better results than edit distance variants, it is no longer the state of the art when computing the semantic textual similarity.

Recent research in semantic textual similarity relies on deep learning and various vector-based representations of sentences to calculate the semantic textual similarity between two sentences. The rationale for employing vector representations is that that computing similarity between vectors is more straightforward than computing similarity between texts. It is easy to calculate how close or distant two vectors are by using well understood mathematical distance metrics. In recent years, researchers have experimented with a number of methods such as averaging word embeddings (Ranasinghe et al. 2019a), smooth inverse frequency (Arora et al. 2019), sent2vec (Pagliardini et al. 2018) etc. to represent variable length text in fixed size vectors. In addition, specific recently developed deep learning architectures have emerged as particularly effective in representing sentence vectors. The deep learning community have termed these deep learning architectures ‘sentence encoders’. The sentence encoders take variable length text as input and produce a fixed size vector as output. The output vector is defined as sentence embedding or sentence vector which is a meaningful numerical representation of the input sentence.

In this study we employ sentence encoders to derive numerical representations for all sentences in the translation memory. These numerical representations will be referred to as *translation memory sentence vectors*. We use the same sentence encoder to encode the input sentences, i.e. the sentences to be translated by the translators. These sentences will be referred to as *input sentence vectors*. For each of the input sentence vectors, the cosine similarity[8] with each of the translation memory sentence vectors is computed and the one with minimal distance is returned.

Three sentence encoders are used to acquire the sentence vectors. The sentence encoders chosen for this study were the ones performing very well for Semantic Textual Similarity tasks.

[8] Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1.

2.1 InferSent

InferSent is an NLP technique for universal sentence representation developed by Facebook which uses supervised training to produce high quality sentence vectors (Conneau et al. 2017). The authors explore 7 different architectures for sentence encoding including Long short-term memory (LSTM) (Hochreiter & Schmidhuber 1997), Gated recurrent unit (GRU) (Chung et al. 2014), Bi directional Long Short-Term Memory (BiLSTM)[9] with mean/max pooling,[10] Self-attentive network and Hierarchical Deep Convolutional Neural Network (Conneau et al. 2017). They evaluate the quality of the sentence representation by using them as features in 12 different transfer tasks[11] like Binary and multi-class text classification, semantic textual similarity, Paraphrase detection etc. The results indicate that the BiLSTM with the max-pooling operation performs best on these tasks (Conneau et al. 2017).

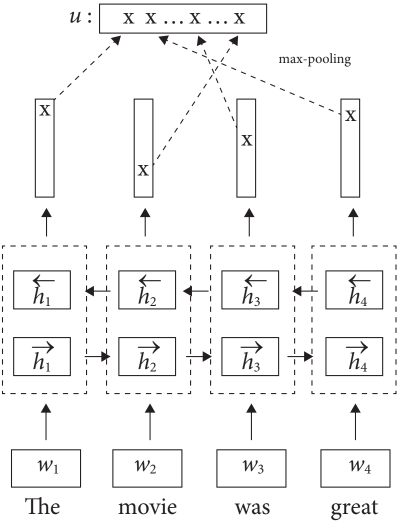


Figure 1. Bi-LSTM max-pooling network. Source: Supervised learning of universal sentence representations from natural language inference data. (Conneau 2017)

[9] As the name suggests, these networks are bidirectional, that is, they have access to both past and future input features for a given time. In the context of language processing, this means it considers both the context before and after a word.

[10] Pooling layers provide an approach to down sampling feature maps by summarising the presence of features. Two common pooling methods are mean pooling and max pooling that summarise the average presence of a feature and the most activated presence of a feature respectively.

[11] Transfer Tasks are the tasks in Transfer Learning. Motivation for Transfer learning used for Machine Learning and Deep Learning is based on the fact that people can intelligently apply knowledge learned previously for a different task or domain that can be used to solve new problems faster or with better solutions. These new problems are the transfer tasks.

Facebook released two models which derive the sentence embeddings. One model is trained with GloVe[12] (Pennington et al. 2014) which in turn has been trained on text preprocessed with the PTB tokeniser[13] and the other model is trained with fastText[14] (Mikolov et al. 2019) which has been trained on text preprocessed with the MOSES tokeniser.[15] We used the model trained on fastText (Mikolov et al. 2019) as character embeddings provide better coverage than the standard embeddings (Mikolov et al. 2019).

2.2 Universal sentence encoder

The Universal Sentence Encoder (Cer et al. 2018) released by Google is the second sentence encoder we employed in this study. It comes with two versions i.e. one trained with Transformer encoder and other trained with Deep Averaging Network (DAN). Both architectures are outlined briefly below. The two have a trade-off of accuracy and computational resource requirement. While the one with Transformer encoder has higher accuracy, it is computationally more expensive. The one with DAN encoding is computationally less expensive but with slightly lower accuracy.

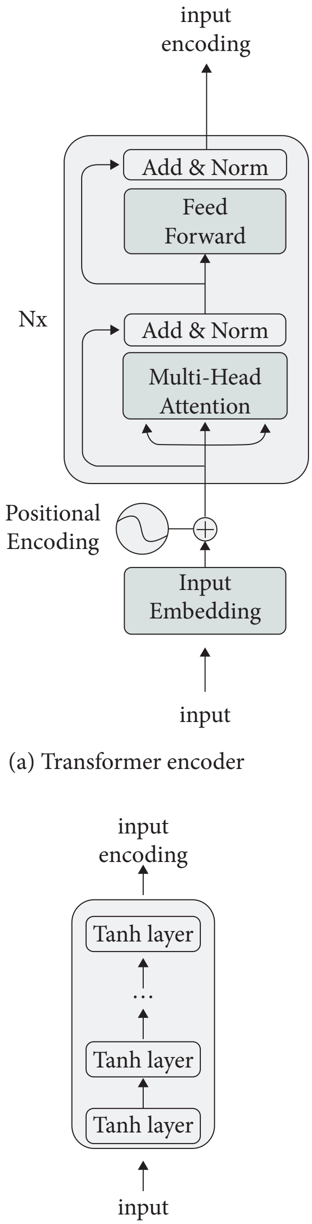
The original Transformer encoder model constitutes an encoder and decoder. Since our research is only focussed on encoding sentences to vectors, we only use its encoder part. The encoder is composed of a stack of $N = 6$ identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. Cer et al. (2018) also employed a residual connection around each of the two sub-layers, followed by layer normalisation. Since the model contains no recurrence and no convolution, for the model to make use of the order of the sequence, it must inject some information about the relative or absolute position of the tokens in the sequence, that is what the "positional encodings" does. The transformer-based encoder achieves the best overall transfer task performance. However, this comes at the cost of computing time and memory usage scaling dramatically with sentence length.

[12] GloVe is an unsupervised learning algorithm for obtaining vector representations for words.

[13] PTBTokenizer is a an efficient, fast, deterministic tokeniser implemented as a finite automaton, produced by JFlex. While deterministic, it employs good heuristics and can successfully decide when single quotes are parts of words, when periods do and do not imply sentence boundaries, etc. It is developed by Christopher Manning, Tim Grow, Teg Grenager, Jenny Finkel, and John Bauer (https://nlp.stanford.edu/software/tokenizer.html).

[14] FastText is a library for learning of word embeddings and text classification created by Facebook's AI Research lab. The model allows to create an unsupervised learning or supervised learning algorithm for obtaining vector representations for words. Facebook makes available pretrained models for 294 languages.

[15] Moses is a statistical machine translation system that allows you to automatically train translation models for any language pair. Moses Tokeniser is the tokeniser that is included in this system.



(a) Transformer encoder

(b) DAN encoder

Figure 2. Side by side model architectures comparison for the transformer and DAN sentence encoders

Deep Averaging Network (DAN) is much simpler where input embeddings for words and bi-grams are first averaged together and then passed through a feedforward deep neural network (DNN) to produce sentence embeddings. The primary advantage of the DAN encoder is that computation time is linear in the length of the input sequence.

Since the Transformer encoder model had higher accuracy and had enough efficiency for the task, we used this architecture to obtain the sentence encodings. (Efficiency of each sentence encoder will be discussed in Section 3).

2.3 Sentence BERT

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) has set a new state-of-the-art performance benchmark on many natural language processing tasks like question answering, natural language inference, text classification and many others. BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling. The results show that a language model which is bidirectionally trained can have a deeper sense of language context and flow than single-direction language models. The attention model learns contextual relations between words (or sub-words) in a text.

The BERT model has achieved state of the arts performance in semantic textual similarity task too. However, it requires that both sentences to be fed into the network, which causes a massive computational overhead: finding the most similar pair in a collection of 10,000 sentences requires about 50 million inference computations which will take approximately 65 hours with BERT. This is not good enough for a translation memory tool.

Therefore, we used Sentence-BERT (SBERT) (Reimers & Gurevych, 2019), a modification of the pretrained BERT network which uses Siamese and triplet network structures (Reimers & Gurevych, 2019) to derive semantically meaningful sentence embeddings that can be compared using cosine similarity. This reduces the effort for finding the most similar pair from 65 hours with to about 5 seconds with SBERT, while maintaining the accuracy from BERT.

We used pre trained 'bert-large-nli-stsb-mean-tokens' SBERT model.[16] The model had best performance on Semantic Textual Similarity (STS) benchmark[17]

[16] https://github.com/UKPLab/sentence-transformers/blob/master/docs/pretrained-models/nli-models.md

[17] The STS Benchmark (http://ixa2.si.ehu.eus/stswiki) contains sentence pairs with human gold score for their similarity.

with 85.29% Spearman correlation coefficient.[18] It was first trained on AllNLI Dataset[19] and then was fine-tuned on STS benchmark training set.

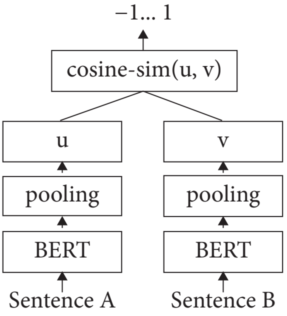


Figure 3. SBERT architecture at inference, for example, to compute similarity scores

3. Dataset and experiments

For the experiments of this study we used the DGT-Translation Memory[20] which has been made publicly available by the European Commission's (EC) Directorate General for Translation (DGT) and the EC's Joint Research Centre. It consists of sentences and their professional translations covering twenty-two official European Union (EU) languages and their 231 language pair combinations (Steinberger et al., 2012). It is typically used by translation professionals in combination with TM software to improve the speed and consistency of their translations. We should note that the DGT TM is a valuable resource for translation studies and for language technology applications, including statistical machine translation, terminology extraction, named entity recognition, multilingual classification and clustering, among others.

[18] Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. It is a popular evaluation metric in Semantic Textual Similarity Tasks.

[19] The AllNLI dataset is the concatenation of the SNLI dataset (https://nlp.stanford.edu/projects/snli/) and the MultiNLI dataset (https://www.nyu.edu/projects/bowman/multinli/). The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labelled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of natural language inference (NLI). The Multi-Genre Natural Language Inference (MultiNLI) corpus is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information. The corpus is modelled on the SNLI corpus, but differs in that covers a range of genres of spoken and written text.

[20] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

While we chose English-Spanish sentence pairs for the experiments of this study,[21] our approach is easily extendable to any language pair. In this particular study, 2018 Volume 1 was used as experimental translation memory and 2018 Volume 3 was used as input sentences. The translation memory we built from 2018 Volume 1 featured 230,000 sentence pairs whilst, 2018 Volume 3 had 66,500 sentence pairs which we used as input sentences.

The experimental process contains the following five steps:

1. The sentence embeddings for the sentence in the translation memory were generated using one of the sentence encoders described above (InferSent). The generated sentence embeddings were stored in the random-access memory of the computer in order to enable fast access to them.
2. For each input sentence, the sentence embedding from InferSent (Same sentence encoder as the first step) was acquired.
3. The cosine similarity between the embedding of each input sentence and the embeddings of the sentences from the translation memory was computed and the embedding with the highest similarity score from the translation memory for each input sentence was returned as the best match for that input sentence.
4. Repeat above three steps for all the input sentences
5. Repeat above four steps for the other two sentence encoders too. (Universal Sentence Encoder and SBERT)

At the end of the process, a file is generated for each sentence encoder which features the best matches for all input sentences for each sentence computed by this sentence encoder.

Given that speed of retrieval is one of the main features expected from a translation memory, Table 1 below discusses the efficiency of each sentence encoder. The experiments were done in Intel(R) Core (TM) computer with i7-8700[22] CPU and 3.20GHz clock speed.[23] While the performance of the sentence encoders would be more efficient in a GPU (Graphics processing unit), we carried our experiments in CPU (Central Processing Unit) since the translators using translation memory tools would not have access to GPU on daily basis.

[21] In order to keep the study manageable, we decided to choose one language pair only and the main reason for opting for English-Spanish was because 3 native Spanish speakers were available to assist us with the evaluation experiments.

[22] Core i7-8700 is a 64-bit hexa-core high-end performance x86 desktop microprocessor introduced by Intel in 2017.

[23] Clock speed is the rate at which a processor executes a task and is measured in Gigahertz (GHz). A higher number means a faster processor.

Table 1. Time taken to acquire similar sentences with each sentence encoder

Embedding type	Time to get embeddings for 230,000 sentences	Time to get embedding for the new sentence	Time to retrieve the best match
InferSent	496s	0.022s	0.52s
Universal Sentence Encoder	108s	1.23s	0.42s
Sentence-BERT	1102s	0.052	0.45s

The translation memory was processed in batches of 256 sentences with a view to obtaining sentence embeddings. As seen in Table 1, the *Universal Sentence Encoder* was the most efficient encoder delivering sentence embeddings within 108 seconds for 230,000 sentences. At the other end was *Sentence-BERT* which took 1102 seconds to derive the sentence embeddings for the same number of sentences in the translation memory. Even though these times may appear to very long, we should keep in mind that this process needs to be done only once. In a real-world scenario, these sentence embeddings would be kept in a database, so they do not need to be computed again.

The second column of the above table reports the time needed for each sentence encoder to embed a single sentence. Input sentences were not processed in batches as was done for the TM sentences. The rationale behind this decision was the fact that translators translate sentences one by one. An interesting observation is that while the Universal Sentence Encoder was the most efficient in generating sentence embeddings in batches, it was the least efficient encoder one to derive the embeddings for single sentences. It took 1.23 seconds to encode a single sentence. InferSent was the fastest sentence encoder for a single sentence.

The third column reports the time needed to retrieve the best match from the translation memory. This includes the time taken to compute cosine similarity between the embeddings of TM sentences and the embeddings of the input sentence. Also, it includes the time taken to sort the similarities, get the index of the highest similarity and retrieve the TM sentence considered as perfect match for the input sentence. As shown in Table 1, all sentence encoders needed approximately 0.5 seconds only to perform this operation. As a whole, to identify the best match from the translation memory, InferSent and Sentence-BERT encoders did not take more than 1 second while Universal sentence encoder took 1.6 seconds which is considered good enough for an operational translation memory tool. These numbers provide evidence that the proposed methods are fast enough to be used in a real-world environment.[24]

[24] Note that we did not implement the algorithms in such a way to maximise their speed. This means that the reported times can be improved ever further.

4. Evaluation and results

In this section we report the results of the evaluation of the performance of the three selected sentence encoders. We ran automatic evaluation experiments by comparing the matches returned by Okapi[25] and the matches returned by each of the sentence encoders.

In order to assess whether our methods perform better than traditional TM tools, we compare our results to those obtained by Okapi, which uses simple edit distance to retrieve matches from the translation memory. For all the experiments we used *DGT-TM 2018 Volume 1* as the translation memory and *2018 Volume 3* – as the source for input sentences. With a view to measuring the quality of a retrieved segment, the METEOR score[26] (Lavie & Agarwal 2007) was computed between the translation of the incoming sentence as present in the DGT-TM 2018 and the translation of the match retrieved from the translation memory. This process was repeated for both the segments retrieved by our deep learning methods and those retrieved using Okapi. We used the METEOR score since it can capture the semantic similarity between two segments better than BLEU score[27] (Lavie & Agarwal 2007).

For a better comparison, we first removed the sentences where the matches provided by Okapi and the sentence encoders were same. Next, in order to analyse the results, we divided the results into 5 partitions according to the fuzzy match score retrieved from Okapi: 0.8–1, 0.6–0.8, 0.4–0.6, 0.2–0.4, and 0–0.2. The ranges were selected to understand the behaviour of the sentence encoders in TM matching and retrieval task. The first partition contained the matches derived from Okapi that had a fuzzy match score[28] between 0.8 and 1. We calculated the average METEOR score for the segments retrieved from Okapi and for the

[25] See footnote 1.

[26] METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric for the evaluation of machine translation output. It has several features that are not found in other metrics, such as stemming and synonymy matching, along with the standard exact word matching. The metric was designed to fix some of the problems found in the more popular BLEU metric, and also produces good correlation with human judgement at the sentence or segment level. While METEOR is a better choice than BLUE because it can also account for semantic variations between segments, METEOR is still very much an overlap metric. This means it is not able to identify that two sentences talk about the same thing when they have different word orders.

[27] BLEU (BiLingual Evaluation Understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. BLEU was one of the first metrics to claim a high correlation with human judgements of quality, and remains one of the most popular automated and inexpensive metrics.

[28] Fuzzy match score in Okapi is based on Sørensen-Dice's Coefficient with 3-grams. The Sørensen–Dice coefficient is a statistic used to measure the similarity of two strings. The algorithm works by comparing the number of identical character pairs between the two strings.

segments retrieved from each of the sentence encoders in this particular partition. We repeated this process for all the partitions. Table 2 below lists the results for each sentence encoder and Okapi.

Table 2. Mean METEOR score from each of the sentence encoders and from Okapi for each METEOR score range. The best score in each range has marked as bold.

Fuzzy match score range	Okapi meteor score mean	InferSent meteor score mean	Universal Sentence Encoder meteor score men	SBERT meteor score mean
0.8–1.0	**0.931**	0.864	0.854	0.843
0.6–0.8	0.693	**0.743**	0.702	0.698
0.4–0.6	0.488	**0.630**	0.594	0.602
0.2–0.4	0.225	**0.347**	0.318	0.316
0–0.2	0.011	**0.134**	0.128	0.115

As can be seen from Table 2, for the fuzzy match score range 0.8–1.0, Okapi METEOR score mean is higher than any of the mean METEOR score of the sentence encoders which indicates that matches returned in that particular fuzzy match score range by Okapi are better than the matches returned by any of the sentence encoders. However, in rest of the fuzzy match score ranges the sentence encoders outperform Okapi which shows that for the fuzzy match score ranges below 0.8, the sentence encoders offer better matches than Okapi. From the sentence encoders InferSent performs better than both the Universal Sentence Encoder and SBERT.

The results in Table 2 show that when there are close matches in the Translation Memory, edit distance delivers better matches than the sentence encoders. However, when the edit distance fails to find a proper match in the Translation Memory, the match offered by the sentence encoders will be better.

5. Analysis of typical errors

As aforementioned, a file was generated for each sentence encoder (one for InferSent, one for Universal Sentence Encoder, and one for SBERT). These files consisted of several source segments, their human translation, the translation provided by the sentence encoder and the translation provided by Okapi. Three native Spanish speakers with background in translation went through each file and compared the matches provided by the sentence encoders and the matches provided by Okapi. The usual pattern they found was that the sentence encoders returned better results; however, there were a limited number of cases where Okapi performed better. The native speakers analysed more than one thousand segments and below is a brief analysis of the typical error cases they found.

In a number of cases InferSent performed better than Okapi because the latter proposed translations which contained information which did not appear in the English input segment. As an illustration of this typical case, for the input segment (19) for which the correct translation is (20) Okapi retrieved (21) whilst InferSent selected (22), which is more appropriate.

(19) The audit shall include.

(20) La evaluación incluirá.

(21) Los indicadores de rendimiento incluirán. (Key performer indicators shall include)

(22) El informe incluirá. (The report shall include)

In other cases, Okapi selects segments which capture only part of the meaning of the input segment correctly but fails to provide its correct whole meaning. For example, for the input segment (23) Okapi selects (24). Due to its exclusive reliance on edit distance, Okapi selects a segment which has the correct temporal expression (16 June/16 de junio) but the rest of the retrieved translation does not have any connection with the original. In contrast, InferSent is able to retrieve a segment which conveys the meaning correctly but has the incorrect date (25). From the point of view of the effort required to produce accurate translation, the segment selected by InferSent requires less effort (as the translator would have to correct the date only) than the one selected by Okapi.

(23) It shall apply in all Member States from 16 June 2020.

(24) A partir del 16 de junio de 2024, los Estados miembros utilizarán la función de registro centralizada. (Member States will use the centralised registration function from 16 June 2024)

(25) Los Estados miembros aplicarán dichas disposiciones a partir del 21 de diciembre de 2020. (These provisions shall apply in all Member States from 21 December 2020)

The advantage of sentence encoders can also be observed when comparing the performance of Okapi with the Universal Sentence Encoder. Okapi often recognises a part of the English sentences only, so the match that is suggested is only partially correct. As an illustration, the short sentence (26) is retrieved from Okapi as (27). The word *brief* does not appear in the retrieved text and, additionally, Okapi adds *de las mercancías*. The translation retrieved by the Universal Sentence Encoder is correct (28). This pattern can also be seen when comparing Okapi with SBERT. By way of example, while proposed match for (29) by SBERT is correct (30), Okapi only recognises one word of the segment, as the retrieved translation is (31).

(26) Brief description

(27) Descripción de las mercancías (Goods description)

(28) Breve descripción

(29) Test equipment

(30) Los equipos de ensayo (The test equipment)

(31) Equipo informático (IT equipment)

In general, and on a number of occasions, Okapi omits some of the information that the sentence encoders identify. The equivalent of the sentence (32) is retrieved by Okapi as (33) with *Edible offal* missing in Okapi's proposal. The sentence retrieved from InferSent however, conveys this information (34).

(32) Edible offal of bovine animals, frozen

(33) De la especial bovina, congelados (Bovine animals, frozen)

(34) Carne de animales de la especie bovina, congelada.

Okapi often fails not only with whole sentences but also with segments that only contain one word. When retrieving the translation of the word (35) the sentence encoder InferSent suggest (36), whereas Okapi also adds the word *Lugar* (37). This also happens with (38), which InferSent returns as (39), whereas Okapi retrieves (40); the word *elección* (choice) does not appear in the English sentence. In addition, Okapi often fails with multiword expressions. The translation of the multiword expression (41) is retrieved by Okapi as (42) and in this case the proposed match features redundant information. The segment retrieved by SBERT represents the best solution (43).

(35) Date

(36) Fecha

(37) Lugar y fecha (Place and date)

(38) Fuel

(39) Combustible

(40) elección del combustible (choice of fuel)

(41) Engine type

(42) Potencia del motor principal en KW: Marca: Tipo (Main engine power in KW: Make: Type)

(43) Tipo de motor

There are cases where the segment retrieved from the sentence encoder is similar to the one retrieved from Okapi, but the sentence encoder is better at conveying subtle nuances. For instance, the proposed translation for sentence (44) by Okapi is (45) and the sentence retrieved from the Universal Sentence Encoder is (46). The nuance refers to the proposed translation for *as appropriate*. Okapi returns (47), whereas the Universal Sentence Encoder retrieves the correct translation (48). Another similar example where Okapi fails is the retrieved translation of (49) as (50); the Universal Sentence Encoder acts correctly on this occasion.

(44) This Decision shall be kept under constant review and shall be renewed or amended, as appropriate, if the Council deems that its objectives have not been met.

(45) Se prorrogará o modificará, si procede, en caso de que el Consejo estime que no se han cumplido los objetivos de la misma. (This Decision shall be renewed or amended, if appropriate, if the Council deems that its objectives have not been met)

(46) Será prorrogada o modificada, según proceda, si el Consejo considera que no se han cumplido sus objetivos. (This Decision shall be renewed or amended, as appropriate, if the Council deems that its objectives have not been met)

(47) si procede (if appropriate)

(48) según proceda (as appropriate)

(49) if applicable

(50) no procede (not applicable)

There are a number of cases where Okapi returns a completely incorrect translation as opposed to the sentence encoders. For (51) Okapi proposed (52) which has nothing to do with the original meaning. The Universal Sentence Encoder offers a simple, yet good solution (53). Another similar example is the suggested match for (54): Okapi returns a completely wrong translation (55), the proposal by SBERT while incorrect, is semantically similar (56).

(51) None of the above

(52) Veánse los considerandos 92 a 94 (See items 92 to 94)

(53) Ninguna (None)

(54) limes

(55) Reducir al mínimo los tiempos de permanencia (Minimise dwell times)

(56) de lima (made of limes)

There are a limited number of cases where Okapi fares better than the sentence encoders. One such example is when encoders retrieve matches of named entities. By way of illustration, the translation the Universal Encoder retrieves for (57) is (58) instead of (59); SBERT retrieves (60) when the original segment is (61); and the proposal by InferSent for (62) is (63).

(57) Japan

(58) Israel

(59) Japón

(60) Singapur (Singapore)

(61) Philippines

(62) within municipality of Sitovo

(63) en el municipio de Alfatar (within municipality of Alfatar)

Finally, and occasionally, sentence encoders too could propose translations featuring redundant information which does not appear in the English original segment. The match InferSent returns for (64) is (65) and in this case Okapi retrieves a correct translation (66). On another, isolated occasion, SBERT also adds a redundant word *mixto* (joint) by proposing (67) as translation for (68). In this particular instance the retrieved match by Okapi is correct (69).

(64) Requirements

(65) Requisitos del Eurosistema (Eurosystem requirements)

(66) Requisitos

(67) El Comité mixto adoptará su reglamento interno (The joint Committee shall establish its own rules of procedures)

(68) The Committee shall establish its own rules of procedures

(69) El Comité dispondrá su reglamento interno

During this error analysis the Spanish native speakers did not count how many times each of the types of errors listed above occur. A more detailed quantitative analysis of these errors is planned for the future.

6. Conclusion

We evaluated the performance of three sentence encoders to retrieve translation memory matches in English-Spanish sentence pairs and used the DGT translation memory as dataset for our experiments. We compared the results from each of the sentence encoders with the results from Okapi which uses edit distance to acquire the best match from the translation memory; the results using fuzzy match score were evaluated across several ranges. The results show that for sentences with a fuzzy match score less than 0.8 in Okapi, the sentence encoders return better matches than simple edit distance. Of the sentence encoders, InferSent fares best. We also discuss the results of the analysis of typical errors where three native Spanish speakers analysed the matches proposed by the sentence encoders and Okapi.

The main limitation of the approach is that it requires considerable RAM as the sentence embeddings for the TM is stored in RAM. When the TM becomes very large, it would require a larger RAM which would not be practical or feasible for commercial TM systems. Therefore, we plan a future study in which we use a database[29] to store the sentence embeddings for the TM which in turn, would address the RAM challenge. We have already experimented with AquilaDB[30] which is a database specially designed to store vectors/embeddings.

The second limitation is the time taken to retrieve a match which is high and in fact impractical with a large TM. This is a common problem for Deep learning applications which is usually solved by employing GPUs. However, in this case it would not be feasible to use GPUs since they are expensive and translators using translation memory tools would not have access to GPU on daily basis. To overcome this impediment, we envisage the deployment of algorithms to filter out the sentences from the TM before the retrieval and to make the calculation of cosine similarity between vectors a computationally less intensive process. Faster algorithms generating sentence embeddings like averaging word embeddings (Ranasinghe et al., 2019a) will be used in these experiments.

In conclusion, this study shows that deep learning techniques improve the matching and retrieval in TM tools and hold promise for future research. In particular, experiments with other deep learning architectures such as Siamese Neural Networks (Ranasinghe et al., 2019b) which perform well in text similarity tasks, are worth considering and are part of our immediate plans.

[29] A database is an organised collection of data, generally stored and accessed electronically from a computer system. Usually this data is stored on the hard disk of the database server.

[30] AquilaDB is a Vector Database to store Feature Vectors/ embeddings. It is available in https://aquiladb.xyz/.

Acknowledgements

We would like to express our gratitude to Encarnación Núñez Ignacio from the University of Wolverhampton and Lucía Bellés-Calvera from Jaume I University, Castellón, who compared the results between the sentence encoders and Okapi.

Funding

This study has been partially carried out in the framework of the research projects UMA18-FEDERJA-067, FFI2016-75831-P and MI4ALL (CEI-RIS3).

References

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2019. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings". *Proceedings of the 5th International Conference on Learning Representations (ICLR'2017)*.

Cer, D., Yang, Y., Kong, S. yi, Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., Sung, Y. H., Strope, B., & Kurzweil, R. 2018. "Universal sentence encoder for English". *Proceedings of EMNLP 2018 – Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, 169–174. https://doi.org/10.18653/v1/D18-2029

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. 2014. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". *NIPS 2014 Workshop on Deep Learning, December 2014*. http://arxiv.org/abs/1412.3555

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. 2017. "Supervised learning of universal sentence representations from natural language inference data". *EMNLP 2017 – Conference on Empirical Methods in Natural Language Processing, Proceedings*, 670–680. https://doi.org/10.18653/v1/D17-1070

Damerau, F. J. 1964. "A technique for computer detection and correction of spelling errors". *Communications of the ACM*, 7(3), 171–176. https://doi.org/10.1145/363958.363994

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://github.com/tensorflow/tensor2tensor

Dice, Lee R. 1945. "Measures of the Amount of Ecologic Association Between Species". *Ecology*. 26 (3): 297–302. https://doi.org/10.2307/1932409

Ganitkevitch, Juri, Van Durme Benjamin, and Chris Callison-Burch. 2013. "PPDB: The paraphrase database". In *Proceedings of NAACL-HLT*, 758–764, Atlanta, Georgia.

Gow, Francie. 2003. Metrics for Evaluating Translation Memory Software. PhD thesis. University of Ottawa.

Grönroos, Mickel, and Ari Becks. 2005. "Bringing Intelligence to Translation Memory Technology". *Proceedings of the International Conference Translating and the Computer 27*. London: ASLIB.

Gupta, R., Bechara, H., El Maarouf, I. and Orasan, C., 2014, August. UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation* (SemEval 2014) (pp. 785–789). https://doi.org/10.3115/v1/S14-2139

Rohit Gupta, Hanna Bechara, and Constantin Orăsan. 2014b. Intelligent Translation Memory Matching and Retrieval Metric Exploiting Linguistic Technology. In *Proceedings of the thirty sixth Conference on Translating and Computer*, London, UK.

Gupta, R., Orăsan, C., Zampieri, M., Vela, M., Mihaela Vela, van Genabith, J. and R. Mitkov. 2016a. "Improving Translation Memory matching and retrieval using paraphrases", *Machine Translation*, 30(1), 19–40. https://doi.org/10.1007/s10590-016-9180-0

Gupta, R., Orăsan, C., Liu, Q. and R. Mitkov. 2016b. "A Dynamic Programming Approach to Improving Translation Memory Matching and Retrieval using Paraphrases". Lecture Notes in Computer Science book series (LNCS, volume 9924). *Proceedings of the 19th International Conference on Text, Speech and Dialogue (TSD)*, Brno, Czech Republic. Springer. https://doi.org/10.1007/978-3-319-45510-5_30

Hochreiter, S., & Schmidhuber, J. 1997. "Long Short-Term Memory". *Neural Computation*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hodász, G. and Pohl, G., 2005, September. MetaMorpho TM: a linguistically enriched translation memory. In *International Workshop: Modern Approaches in Translation Technologies* (pp. 26-30).

Lavie, A., & Agarwal, A. 2007. "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments". *Proceedings of the Second Workshop on Statistical Machine Translation, June*, 228–231. http://acl.ldc.upenn.edu/W/W05/W05-09.pdf#page=75. https://doi.org/10.3115/1626355.1626389

Levenshtein, V. I., 1966, February. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, No. 8, pp. 707–710).

Macklovitch, E. and Russell, G., 2000, October. What's been forgotten in translation memory. In *Conference of the Association for Machine Translation in the Americas* (pp. 137–146). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-39965-8_14

Marelli, Marco, Bentivogli, Luisa, Baroni, Marco, Bernardi, Raffaella, Menini, Stefano and Zamparelli, Roberto, 2014, August. SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (Sem Eval 2014)* (pp. 1–8). Dublin, Ireland: Association for Computational Linguistics. https://www.aclweb.org/anthology/S14-2001. https://doi.org/10.3115/v1/S14-2001

Mikolov, Tomas, Grave, Edouard, Bojanowski, Piotr, Puhrsch, Christian and Joulin, Armand, 2018, May. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1008

Mitkov, R. 2005. 'New Generation Translation Memory systems'. Panel discussion at the 27th international Aslib conference 'Translating and the Computer'. London..

Mitkov, R. "Translation Memory". 2020. In S. Deane-Cox and A. Spiessens (Eds), *The Routledge Handbook of Translation and Memory*. Basingstoke: Routledge.

Pagliardini, M., Gupta, P. and Jaggi, M., 2018, June. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers) (pp. 528–540). https://doi.org/10.18653/v1/N18-1049

Pekar, V. and Mitkov, R. 2007. "New Generation Translation Memory: Content-Sensitive Matching". *Proceedings of the 40th Anniversary Congress of the Swiss Association of Translators, Terminologists and Interpreters*. Bern: ASTTI, 2007.

Pennington, J., Socher, R. and Manning, C. D., 2014, October. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1162

Planas, Emmanuel. 2005. "SIMILIS: Second-generation translation memory software". proceedings of the 27th International Conference Translating and the Computer. London.

Planas, Emmanuel and Furuse, Osamu. 2003. "Formalizing Translation Memory". In Michael Carl and Andy Way (Eds), *Recent Advances in Example-Based Machine Translation* (pp. 157–188). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-010-0181-6_5

Ranasinghe, T., Orasan, C. and Mitkov, R., 2019, September. Enhancing Unsupervised Sentence Similarity Methods with Deep Contextualised Word Representations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 994–1003).

Ranasinghe, T., Orasan, C. and Mitkov, R., 2019, September. Semantic textual similarity with Siamese neural networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)* (pp. 1004–1011).

Reimers, N. and Gurevych, I., 2019, November. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3973–3983). https://doi.org/10.18653/v1/D19-1410

Sørensen, T. 1948. "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons". *Kongelige Danske Videnskabernes Selskab*. 5 (4): 1–34.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. 2012. "DGT-TM: A freely available translation memory in 22 languages". *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 454–459. http://eur-lex.europa.eu/

Timonera, K. and R. Mitkov. 2015. "Improving Translation Memory Matching through Clause Splitting". *Proceedings of the RANLP'2015 workshop 'Natural Language Processing for Translation Memories'*. Hissar, Bulgaria.

Wali, W., Gargouri, B. and Hamadou, A. B. 2017. "Sentence similarity computation based on WordNet and VerbNet". *Computación y Sistemas*, 21(4), 627–635.

TAligner 3.0

A tool to create parallel and multilingual corpora

Zuriñe Sanz-Villar and Olaia Andaluz-Pinedo
University of the Basque Country (UPV/EHU)

In 2010 the TRALIMA/ITZULIK research group began developing a tool to create parallel and multilingual corpora. This would serve (among other functions) to explain, describe, and predict translational behaviour in terms of translational norms and laws (Toury 2012). Instead of relying on already existing resources, it was decided to invest in the creation of a new tool. As a result, TAligner 3.0 enables users to create corpora simultaneously aligning any number of texts (as required by the research) as well as compile corpora consisting of different text types. Furthermore, unlike most tools the aligned corpora can be queried within the program. The aim of this paper is to present, firstly, TAligner 3.0 and secondly, two parallel corpora built using this tool.

Keywords: TAligner, parallel corpora, German-into-Basque translations, theatre translations

1. Introduction

In 2000, Kenny argued that "[t]he emergence of a corpus-based approach to translation studies is arguably one of the most promising developments in the area in recent years" (2000: 143). Now, almost two decades later, the increased amount of studies conducted within Corpus-based Translation Studies (CBTS) bear witness to Kenny's words. In the conduction of such studies, corpora have been built using a variety of tools (the tools used depending on the requirements of each body of research). As obvious as it seems, it is necessary to make a distinction between corpus tools and corpus data, since "there is a continuing tendency within the field [of corpus linguistics] to ignore the tools of analysis and to consider the corpus data itself as an unchanging 'tool' that we use to directly observe new phenomenon in language" (Anthony 2013: 143).

https://doi.org/10.1075/btl.158.05san

The TRALIMA/ITZULIK (Translation, Literature and Audiovisual Media)[1] group has a long tradition of conducting research within the descriptive branch of translation studies (e.g. Merino-Álvarez 1994b; Barambones 2009; Manterola 2011; Zubillaga 2013; Sanz-Villar 2015; Cabanillas 2016; Arrula 2018). The three branches proposed by Holmes in 1972 are interconnected (Toury 2004: 19), with the descriptive branch providing empirical data for the other two and thus serving as a bridge between the theoretical and applied branches. Toury (2012: XII) maintains that in order to obtain such data, research within translation studies should be conducted as follows: description, explanation and prediction. Descriptive Translation Studies (DTS) brought a significant change from prescription to description:

> DTS marked a self-conscious departure from much previous scholarship in the area, which had been highly speculative and prescriptive in its orientation, concerned as it was with determining what an ideal translation should strive to be (…). Proponents of DTS sought instead to engage with real translation phenomena, to describe translations as they actually occur, and to account for observed features of translations with reference to the literary, cultural and historical contexts in which they were produced. (Kenny 2001: 49)

Corpora became an excellent tool towards achieving this goal, allowing researchers to observe and describe real texts systematically. As described by Xiao and Yue (2009: 239): "The marriage between DTS and corpora is only natural, in that corpus linguistics, as a discipline stemming from the description of real linguistic performance, supplies DTS with a systematic method and trustworthy data". From the 1990's on, corpora have also been applied in fields beyond lexicography such as translation studies.

As mentioned above, members of the TRALIMA/ITZULIK research group have been working within DTS for a long time. Due to the benefits obtained through the use of corpora, priority was given to corpus-based descriptive translation studies. After trying commercial tools for the creation of corpora, the group identified the need to develop their own program. The analysis of available tools showed that they have some limitations for descriptive translation studies such as the ones carried out by members of the research group. Manterola (a member of TRALIMA/ITZULIK), for instance, built a multilingual corpus consisting of 12 original Basque literary texts with corresponding translations into seven languages (Manterola 2011: 191). She used WordSmith Tools (Scott 2004). This was the only option to simultaneously align more than two texts at that time as well as a powerful

[1] Further information about the research group can be found here: https://www.ehu.eus/en/web/tralimaitzulik/home, https://addi.ehu.es/handle/10810/34436.

tool to create and analyse corpora. However, some other features of the program encouraged the research group to begin developing their own corpus creation and query tool. These included the fact that WordSmith Tools is not an open-source tool, it uses its own file formats and the process of aligning the texts at sentence level is quite complicated (Uribarri 2016: 251). Although time consuming, developing the new program provided the opportunity to build it according to the group's specific research requirements, as well as the option to continuously improve it in the future. The aim was to create a tool that would allow precise alignment of multiple texts. While automatic alignment may be useful for some studies, others might need a higher precision rate that can only be achieved through manual or supervised alignment. In addition, research on indirect translations, or on several retranslations of an original, requires simultaneous alignment of more than two texts. For us researchers in Translation Studies (with limited knowledge on computer science), it was important that the tool integrates parallel corpus building functions (mainly annotation, structural mark-up, metadata and alignment) and corpus analysis options within it. Since different text types are analysed among researchers of the group, the software should also cater for text genre specificities. For instance, theatre plays require particular structural mark-up and analysis options, and alignment at utterance level is more suitable than at sentence level due to stage translations typical variability.

To achieve these goals, an interdisciplinary approach and the collaboration of researchers from the fields of computer science and translation studies were key factors. As mentioned by Anthony (2013: 144), "[t]o develop a tool for corpus linguistics requires an understanding of not only of human languages but also programming languages, computer algorithms, data storage methods, character encodings, and user-interface visual designs". Since the members of the group lack such knowledge, the programming (and many other) skills of a computer technician were indispensable. Iñaki Albisua has been in charge of this task since 2011.

In summary, while some scholars suggest that corpus linguistic researchers should learn programming in order to create their own tools, our research team adopted the method proposed by Anthony: "I propose here that researchers in corpus linguistics should also work more closely with members of the science and engineering community, such as computer scientists and software engineers, in order to design and build the next generation of corpus tools" (2013: 156).

As will be explained in Section 2, we have spent several years developing the latest version of TAligner 3.0, a third generation tool (McEnery and Hardie 2012). It enables the user to create corpora consisting of different text types (for now narrative and theatrical) and to simultaneously align as many texts as desired.

We are aware of the fact that there is a growing tendency to create fourth generation tools. However, as stated by Anthony (2013: 153), both third and fourth generation tools have their advantages and disadvantages.[2] While third-generation tools run on the desktop, fourth-generation systems are web-based. Thus, the latter are more suitable to deal with very large corpora. As mentioned by McEnery and Hardie (2012: 44), "any corpus software designed for a PC is still subject to the limitations of the memory and processing power of a given user's computer". However, the same authors point out that fourth-generation tools are not significantly better than the third in terms of corpus analysis tools. They are meant to overcome three main issues: "[T]he limited power of desktop PCs, problems arising from non-compatible PC operating systems, and legal restrictions on the distribution of corpora".

As for the first point, this will always depend on the research characteristics and the goals of the researcher. Our studies were not limited due to this issue: Corpora of up to 3.5 million words were created with TAligner and while search queries were quite slow in the second version of the program (see Section 2), this has not been a problem with the latest version where query results are obtained within a few seconds. Secondly, TAligner is a Java-written program; so the only requirement to run the tool is to have this program correctly installed on the desktop. Thus, it can run in operating systems other than Windows. Regarding copyright issues, McEnery and Hardie (2012) argue that since concordance lines do not usually exceed sentence length in web-based systems and this falls "within the level of 'fair use' allowed under copyright law" (2012: 44), authors' copyrights are not violated. However, they also admit that "the legality of such 'fair use' redistribution has yet to be comprehensively tested" (2012: 44). Thus, as we see it, the problem has not yet been satisfactorily resolved.

Sanjurjo-González (2018) analyses eleven computer applications ('frameworks', according to his terminology) that were created by users without programming knowledge to process linguistic corpora (2018: 24). We could add TAligner 3.0 (see Table 1) to the list of third generation tools (WordSmith, AntConc and AntPConc, MonoConc and ParaConc, and LancsBox) and describe it based on the features suggested by Sanjurjo-González (2018: 27–29).

Regarding TAligner's positive attributes, it is necessary to emphasize its user-friendliness and how simply it allows the user to create corpora (see Section 3). Technical assistance is not required to use the program and it has a simple graphical interface. It is not language bound and it allows the compilation and processing of multilingual and parallel corpora, making queries within them without necessarily

[2] Anthony (2013) thoroughly explains said advantages and disadvantages.

Table 1. Features of TAligner 3.0

	TAligner 3.0
Requires technical assistance	No
Corpus compilation	Yes
Processes multilingual corpora	Yes
Processes parallel corpora	Yes
Integrated aligner	Assisted
Processes comparable corpora	No
Indexing and querying technology	Java
Statistics	
– Frequency list	Yes
– Collocations	No
– Keyword list	No
– n-grams	No
Graphical interface	Yes
Generation	Third
Web hosted	No
Annotation support	Yes
Built-in linguistic taggers	
– Grammatical	No
– Semantic	No
– Rhetoric	No
Availability	Downloadable

resorting to other tools (see Section 3.2). Since none of the researchers in the group works with comparable corpora, it was not necessary to include the option to create them. The latest version of the program is downloadable from http://corpusnet.unileon.es/[3] and the digital repository of the University of the Basque Country (https://addi.ehu.es/handle/10810/42445). Additionally, the latest version includes the option of using comments for both narrative and theatrical texts. This may be regarded as a kind of manual annotation. Depending on the purposes of the research, the simple character of TAligner 3.0 may also have its limitations. As will be explained in Section 3.1, the alignment is almost manual, and plain, unannotated text is used to create corpora. Apart from frequency lists, the program does not include any other statistical option. As previously mentioned with regard to corpus size, the program has thus far been used to create small corpora (consisting of

[3] As specified on the webpage, "CorpusNet is a hub of bilingual and multilingual corpora and related resources featuring any of the languages of Spain (Spanish, Catalan, Galician, and Basque) alongside other languages". TRALIMA/ITZULIK has contributed to it as a team member of this network for fostering excellence in research.

around 3.5 million words) and no limitations were observed regarding speed when querying them. As mentioned by Anthony (2017), third-generation tools "are often ideal tools for use with small general corpora".

After explaining the evolution of the tool and steps involved in the building and querying of corpora in TAligner, Section 4 will describe examples of corpora created with TAligner and the paper will conclude with some final remarks and future lines.

2. Evolution of TAligner 3.0

The current version of TAligner 3.0 is the result of a progressive development which involved computer science engineers and members of TRACE (University of León) and TRALIMA/ITZULIK (University of the Basque Country) research groups. The first version of this tool was conceived in 2010 within the framework of TRACE joint research projects. Since then, the software's modifiablility, as well as feedback obtained through its application in a variety of studies led to a number of improvements which better facilitate the tasks required of TAligner. This section will briefly report on the changes made to subsequent versions[4] of the program in an attempt to show the evolution of this corpus management tool.

2.1 TRACE Corpus Tagger/Aligner 1.0©

The first version, called TRACE Corpus Tagger/Aligner 1.0©, was developed at the University of León in collaboration with the computer technician Roberto Pérez González. The aim of the software was to facilitate descriptive-comparative analysis of pairs of original texts and translations. It was designed to work with the different text types studied in TRACE projects: narrative, drama (including theatre, cinema, and television), and poetry/songs (Gutiérrez Lanza, Bandín Fuertes, García González and Lobejón Santos 2015).

These underlying ideas are reflected in the options available in each part of the program: Tagging and aligning. Tagging involves marking different structural units in each type of text for alignment and possible searches (Gutiérrez Lanza et al. 2015). In narrative texts they indicate paragraphs, sentences, headings, and dialogues; in drama, utterances, headings, character names, stage directions and dialogues; in poetry/songs, verses and strophes. Relevant metadata for the projects

[4] Further information about the program's origin and its main features can be found in Uribarri (2016).

is also added to the resulting XML file. In the aligning section, automatic alignment based on structural units' tags (paragraphs or sentences for narrative, utterances for drama, and paragraphs for poetry/songs) is adjusted through different options, such as inserting blank units. This is finally saved to a TMX or HTML format.[5] This tool was applied in Lobejón-Santos's doctoral thesis (2013) to create a parallel corpus of poems.

The design of the tagging and aligning options (which perdured throughout the tool's successive versions), offers advantages in comparison to other corpus programs. On the one hand, the possibility to adjust the structural mark-up and alignment to different text types is an innovation. Apart from traditional alignment in sentences and paragraphs for prose and poetry, it includes the possibility of aligning texts at the utterance level, the unit considered most adequate for drama comparison (Merino-Álvarez 1992: 285, 1994b: 44). Additionally, it is possible to mark up different parts of a text (such as dialogues, stage directions or character names) to facilitate future searches. The tool's orientation towards creating aligned corpora thus contributes to the relatively small number of tools offering this option.

2.2 TRACEAligner 2.0

Computer expert Iñaki Albisua continued developing the program at the University of the Basque Country. Modifications were introduced related to the use of the tool within TRALIMA/ITZULIK. Researchers built and analysed parallel corpora of literary and philosophical texts as part of their doctoral thesis and other studies (Zubillaga 2013; Sanz-Villar 2015; Zubillaga, Sanz-Villar and Uribarri 2015). These experiences informed improvements of the tool and the following options were introduced:

- Alignment of three texts: This was essential in the study of assumed indirect translations. It allows the alignment of original and intermediary texts along with the assumed indirect translations. The option to align more than two texts is an advantage in comparison to most other aligners.
- Extension of metadata fields: New entries were added to register extra information found useful for later corpus queries, such as codes to identify each text.
- More options to manually adjust the automatic alignment: These features facilitate the aligning process and solve potential problems such as the "undo" button.

[5] For more details about TRACE Corpus Tagger/Aligner 1.0©, see Gutiérrez Lanza et al. (2015).

Additionally, Iñaki Albisua created other instruments to complement the functions of this version:

- A cleaning application which solved some recurrent format errors such as double spaces.
- A database for aligned texts and a search engine which enables the user to define queries through different filters.

2.3 TAligner 3.0

Iñaki Albisua developed the tool to the latest version, called TAligner 3.0. It has been used to build and query parallel corpora of narrative texts in some studies, including another doctoral thesis (Arrula 2018). In addition, the program was used to align theatrical texts (Merino-Álvarez and Andaluz-Pinedo 2017; Merino-Álvarez and Andaluz-Pinedo 2020). Another corpus of translated theatre was created for a further doctoral thesis. These processes were used to introduce further improvements to the tool. Some enhancements included in this version are:

- Integration of the aforementioned options to clean and search the corpus as part of the program. Perhaps one of the software's most positive aspects is that it allows for the creation of aligned corpora and the direct querying of them within that same tool.
- Possibility to align several texts: An original text and several retranslations can be aligned simultaneously.
- Automatic numbering of each utterance for theatrical texts. This facilitates mapping and comparison of the theatrical texts' macrostructure, as suggested by Merino-Álvarez for the analysis of theatrical translations (1994b: 47).
- Option of adding or deleting tags that indicate character names, stage directions, and dialogue. This allows the user to solve possible issues which may arise.
- Option to write (and delete) comments on aligned texts that can later be searched for in the query section. This option enables the user to create, apply and search for personalised annotation systems.
- Options to make the alignment visualisation more adaptable: Zooming in and out, adjustment of column width to view as many texts as are in alignment.

To conclude, since the first version of the tool, efforts have been made to develop it according to specific research needs. As a result, the program is the product of a collaborative and cyclical process; the tool's application in various studies allows problems to be detected and the identification of areas which need improvement. These are communicated to the computer expert who updates the tool to maintain its functionality.

Some of TAligner's beneficial options include: The possibility of creating parallel corpora by taking into account different structural units, the alignment of as many texts as necessary and querying those corpora directly within the tool. This integration of alignment and query functions in one tool is an important advantage. According to the analysis of corpus software conducted by Sanjurjo-González (2018: 29–48), there are not many available applications to process parallel corpora and in most cases an external aligner needs to be used, a task which requires technical knowledge. The following section will describe the features of TAligner's latest update in more detail.

3. Building and querying corpora in TAligner 3.0

This section will thoroughly explain how multilingual and parallel corpora can be compiled using TAligner 3.0. The corpora in question consist of narrative and theatrical texts,[6] since these are the text types currently supported by the tool.[7] We will explicitly reference differences between the two as relevant.

3.1 Building corpora

As mentioned in the introduction to this paper, the compilation process of corpora in TAligner 3.0 is simple and intuitive. It consists of three steps: Cleaning (*limpiar*), tagging (*etiquetar*), and aligning (*alinear*). These steps are easily identified in the upper part of TAligner's graphical interface, as shown in Figure 1:

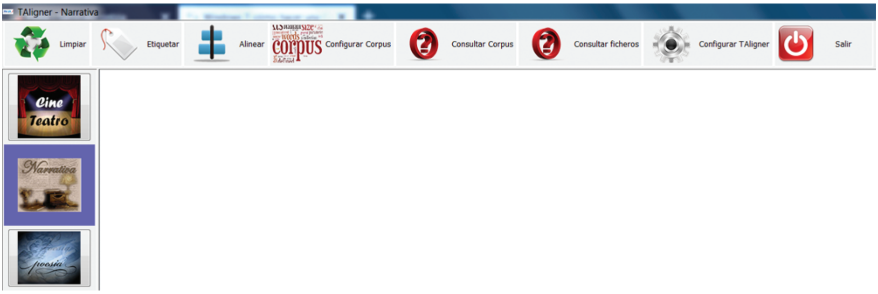
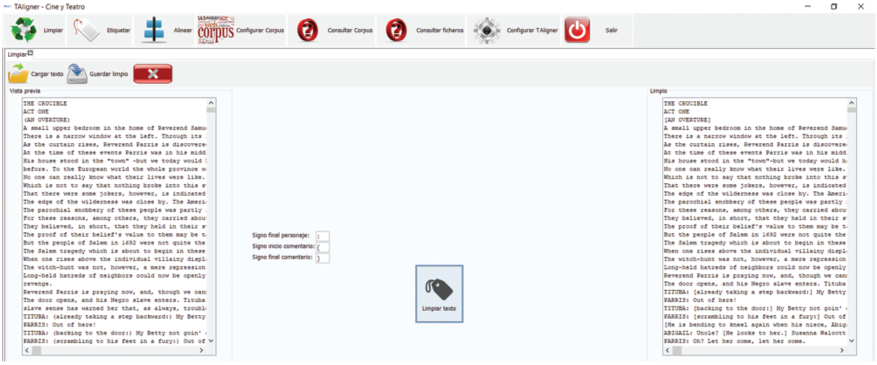


Figure 1. Interface of TAligner 3.0

[6] Theatrical texts have their own specific characteristics that are taken into account for the tagging and aligning processes. However, this is not the case for narrative texts. Therefore, other types of texts (e.g. journalistic) have also been included within the "narrative" option.

[7] There is an ongoing dissertation about poetry translation. This will be considered when adapting the tool to work with such text types.

Following selection of the type of text to include in the corpus from the vertical bar, the texts must be cleaned. This creates standardised texts, reducing the amount of errors in subsequent steps. Files in TXT format will be uploaded on the left hand side of the 'Cleaning screen' (*cargar texto*), as can be seen in Figure 2. The program will subsequently correct formatting errors. The right column shows the cleaned output, which will also be saved in TXT format (*guardar limpio*). As a special feature of theatrical texts, signs which delimit characters' names and stage directions are standardised, as can be seen in Figure 2:



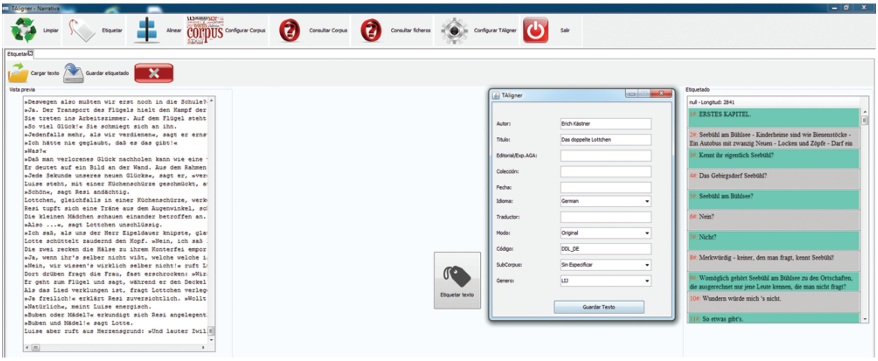
Signo final personaje: :
Signo inicio comentario: (
Signo final comentario:)

```
THE CRUCIBLE
ACT ONE
(AN OVERTURE)
A small upper bedroom in the home of Reverend Samu
There is a narrow window at the left. Through its
As the curtain rises, Reverend Parris is discovere
At the time of these events Parris was in his midd
His house stood in the "town" -but we today would
before. To the European world the whole province w
No one can really know what their lives were like.
Which is not to say that nothing broke into this s
That there were some jokers, however, is indicated
The edge of the wilderness was close by. The Ameri
The parochial snobbery of these people was partly
For these reasons, among others, they carried abou
They believed, in short, that they held in their s
The proof of their belief's value to them may be t
But the people of Salem in 1692 were not quite the
The Salem tragedy which is about to begin in these
When one rises above the individual villainy displ
The witch-hunt was not, however, a mere repression
Long-held hatreds of neighbors could now be openly
revenge.
Reverend Parris is praying now, and, though we can
The door opens, and his Negro slave enters. Tituba
slave sense has warned her that, as always, troubl
TITUBA: (already taking a step backward:) My Betty
PARRIS: Out of here!
TITUBA: (backing to the door:) My Betty not goin'
PARRIS: (scrambling to his feet in a fury:) Out of
```

```
THE CRUCIBLE
ACT ONE
[AN OVERTURE]
A small upper bedroom in the home of Reverend Samu
There is a narrow window at the left. Through its
As the curtain rises, Reverend Parris is discovere
At the time of these events Parris was in his midd
His house stood in the "town"-but we today would h
No one can really know what their lives were like.
Which is not to say that nothing broke into this s
That there were some jokers, however, is indicated
The edge of the wilderness was close by. The Ameri
The parochial snobbery of these people was partly
For these reasons, among others, they carried abou
They believed, in short, that they held in their s
The proof of their belief's value to them may be t
But the people of Salem in 1692 were not quite the
The Salem tragedy which is about to begin in these
When one rises above the individual villainy displ
The witch-hunt was not, however, a mere repression
Long-held hatreds of neighbors could now be openly
Reverend Parris is praying now, and, though we can
The door opens, and his Negro slave enters. Tituba
TITUBA: [already taking a step backward:] My Betty
PARRIS: Out of here!
TITUBA: [backing to the door:] My Betty not goin'
PARRIS: [scrambling to his feet in a fury:] Out of
[He is bending to kneel again when his niece, Abig
ABIGAIL: Uncle? [He looks to her.] Susanna Walcott
PARRIS: Oh? Let her come, let her come.
```

Figure 2. Cleaning process of a theatrical text

Following this, narrative texts are uploaded as cleaned TXT files in the 'Tagging screen' (through the option *cargar texto* as shown in Figure 3). Narrative texts are divided into sentences and theatrical texts into utterances, which are in turn subdivided into character names, stage directions, and dialogue. The resulting XML file is saved (*guardar etiquetado*) with the corresponding metadata, such as author, title, translator, code, etc. As will be seen later (Section 3.2), the more detailed the saved metadata, the more accurately the corpus can be searched.



```
»Deswegen also mußten wir erst noch in die Schule?
»Ja. Der Transport des Flügels hielt den Kampf der
Sie treten ins Arbeitszimmer. Auf dem Flügel steht
»So viel Glück!« Sie schmiegt sich an ihn.
»Jedenfalls mehr, als wir verdienen«, sagt er erns
»Ich hätte nie geglaubt, daß es das gibt!«
»Was?«
»Daß man verlorenes Glück nachholen kann wie eine
Er deutet auf ein Bild an der Wand. Aus dem Rahmen
»Jede Sekunde unseres neuen Glücks«, sagt er, »ver
Luise steht, mit einer Küchenschürze geschmückt, a
»Schön«, sagt Resi andächtig.
Lottchen, gleichfalls in einer Küchenschürze, werk
Resi tupft sich eine Träne aus dem Augenwinkel, sc
Die kleinen Mädchen schauen einander betroffen an.
»Also ...«, sagt Lottchen unschlüssig.
»Ich saß, als uns der Herr Eipeldauer knipste, gla
Lotte schüttelt zaudernd den Kopf. »Nein, ich saß
Die zwei recken die Hälse zu ihrem Konterfei empor
»Ja, wenn ihr's selber nicht wißt, welche welche i
»Nein, wir wissen's wirklich selber nicht!« ruft L
Dort drüben fragt die Frau, fast erschrocken: »Wir
Er geht zum Flügel und sagt, während er den Deckel
Als das Lied verklungen ist, fragt Lottchen verleg
»Ja freilich!« erklärt Resi zuversichtlich. »Wollt
»Natürlich«, meint Luise energisch.
»Buben oder Mädel?« erkundigt sich Resi angelegent
»Buben und Mädel!« sagt Lotte.
Luise aber ruft aus Herzensgrund: »Und lauter Zwil
```

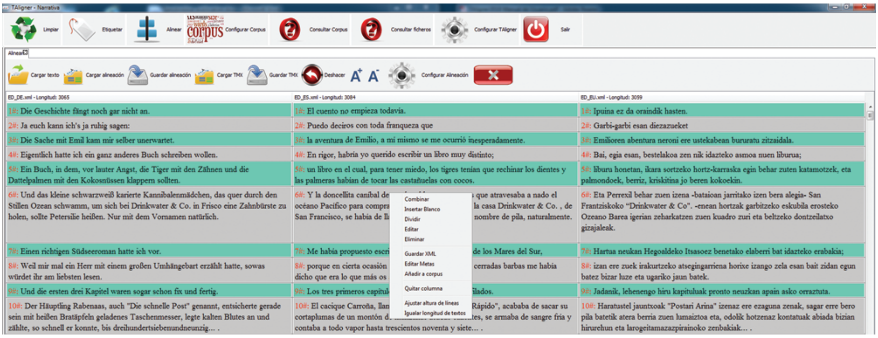
Autor: Erich Kästner
Título: Das doppelte Lottchen
Editorial/Exp.AGA:
Colección:
Fecha:
Idioma: German
Traductor:
Modo: Original
Código: DDL_DE
SubCorpus: Sin Especificar
Genero: LIJ

null - Longitud: 2841

1#: ERSTES KAPITEL.
2#: Seebühl am Bühlsee - Kinderheime sind wie Bienenstöcke - Ein Autobus mit zwanzig Neuen - Locken und Zöpfe - Darf ein
3#: Kennt ihr eigentlich Seebühl?
4#: Das Gebirgsdorf Seebühl?
5#: Seebühl am Bühlsee?
6#: Nein?
7#: Nicht?
8#: Merkwürdig - keiner, den man fragt, kennt Seebühl!
9#: Womöglich gehört Seebühl am Bühlsee zu den Ortschaften, die ausgerechnet nur jene Leute kennen, die man nicht fragt?
10#: Wundern würde mich 's nicht.
11#: So etwas gibt's.

Figure 3. Tagging of a narrative text

Once the texts are tagged, they are aligned in the 'Aligning screen' (Figure 4). Within TAligner 3.0, this means that texts are divided automatically into sentences and that the user must manually make all the necessary adjustments. For that purpose, a menu of options (displayed using right click on the screen) is available as shown in Figure 4. With these options, cells can be merged (*combinar*), split (*dividir*), edited (*editar*), or deleted (*eliminar*). If necessary, empty cells can also be added with the option *insertar blanco*.

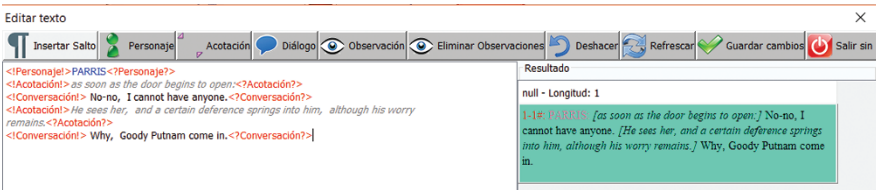


ED_DE.xml - Longitud: 3065	ED_ES.xml - Longitud: 3084	ED_EU.xml - Longitud: 3059
1#: Die Geschichte fängt noch gar nicht an.	1#: El cuento no empieza todavía.	1#: Ipuina ez da oraindik hasten.
2#: Ja euch kann ich's ja ruhig sagen:	2#: Puedo deciros con toda franqueza que	2#: Garbi-garbi esan diezazueket
3#: Die Sache mit Emil kam mir selber unerwartet.	3#: la aventura de Emilio, a mí mismo se me ocurrió inesperadamente.	3#: Emilioren abentura neroni ere ustekabean bururatu zitzaidala.
4#: Eigentlich hatte ich ein ganz anderes Buch schreiben wollen.	4#: En rigor, habría yo querido escribir un libro muy distinto;	4#: Bai, egia esan, bestelakoa zen nik idazteko asmoa nuen liburua;
5#: Ein Buch, in dem, vor lauter Angst, die Tiger mit den Zähnen und die Dattelpalmen mit den Kokosnüssen klappern sollten.	5#: un libro en el cual, para tener miedo, los tigres tenían que rechinar los dientes y las palmeras habían de tocar las castañuelas con cocos.	5#: liburu honetan, ikara sortzeko hortz-karraska egin behar zuten katamotzek, eta palmondoek, berriz, kriskitina jo beren kokoekin.
6#: Und das kleine schwarzweiß karierte Kannibalenmädchen, das quer durch den Stillen Ozean schwamm, um sich bei Drinkwater & Co. in Frisco eine Zahnbürste zu holen, sollte Petersilie heißen. Nur mit dem Vornamen natürlich.	6#: Y la doncellita caníbal de que atravesaba a nado el océano Pacífico para compra la casa Drinkwater & Co. , de San Francisco, se había de lla nombre de pila, naturalmente.	6#: Eta Perrexil behar zuen izena -bataioan jarritako izen bera alegia- San Frantziskoko "Drinkwater & Co". -enean hortzak garbitzeko eskubila erosteko Ozeano Barea igerian zeharkatzen zuen kuadro zuri eta beltzeko dontzeilatxo gizajaleak.
7#: Einen richtigen Südseeroman hatte ich vor.	7#: Me había propuesto escri de los Mares del Sur,	7#: Hartua neukan Hegoaldeko Itsasoez benetako elaberri bat idazteko erabakia;
8#: Weil mir mal ein Herr mit einem großen Umhängebart erzählt hatte, sowas würdet ihr am liebsten lesen.	8#: porque en cierta ocasión cerradas barbas me había dicho que era lo que más os	8#: izan ere zuek irakurtzeko atsegingarriena horixe izango zela esan bait zidan egun batez bizar luze eta ugariko jaun batek.
9#: Und die ersten drei Kapitel waren sogar schon fix und fertig.	9#: Los tres primeros capítul ilados.	9#: Jadanik, lehenengo hiru kapituluak pronto neuzkan apain asko orraztuta.
10#: Der Häuptling Rabenaas, auch "Die schnelle Post" genannt, entsicherte gerade sein mit heißen Bratäpfeln geladenes Taschenmesser, legte kalten Blutes an und zählte, so schnell er konnte, bis dreihundertsiebenundneunzig... .	10#: El cacique Carroña, llan Rápido", acababa de sacar su cortaplumas de un montón d tes, se armaba de sangre fría y contaba a todo vapor hasta trescientos noventa y siete... .	10#: Haratustel jauntxoak "Postari Arina" izenaz ere ezaguna zenak, sagar erre bero pila batetik atera berria zuen lumaiztoa eta, odolik hotzenaz kontatuak abiada bizian hirurehun eta larogeitamazazpirainoko zenbakiak... .

Figure 4. Simultaneous alignment of three narrative texts

This is admittedly a time-consuming task that could have been automatized or semi-automatized. However performed manually, the error rate of the alignment process is reduced to zero, thus ensuring that the user obtains query results without alignment errors.

With theatrical texts, the option of editing a cell enables the user to modify the text and also to manage tags which mark character names, stage directions, and dialogue. This provides an opportunity to solve possible tagging errors directly at this stage (observed in Figure 5). Without this option, a misidentified section of text or a missing tag would have to be corrected in the TXT file, necessitating that the whole process of tagging and aligning be repeated.



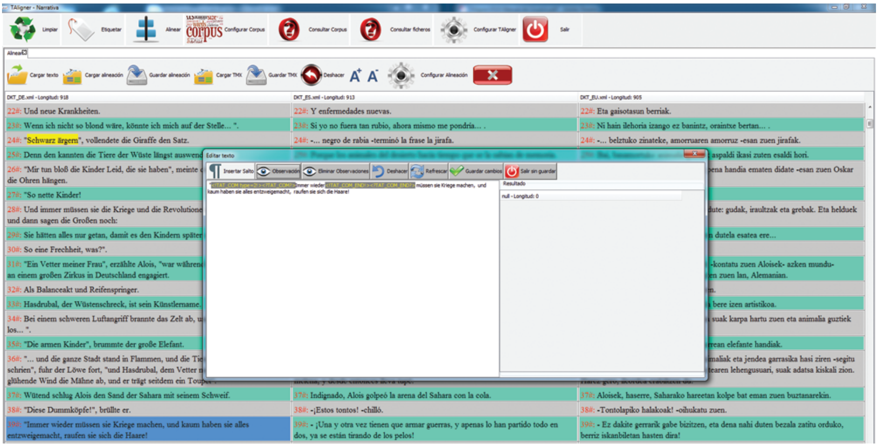
<!Personaje!>PARRIS<?Personaje?>
<!Acotación!>as soon as the door begins to open:<?Acotación?>
<!Conversación!> No-no, I cannot have anyone.<?Conversación?>
<!Acotación!>He sees her, and a certain deference springs into him, although his worry remains.<?Acotación?>
<!Conversación!> Why, Goody Putnam come in.<?Conversación?>

null - Longitud: 1

1-1#: PARRIS: *[as soon as the door begins to open:]* No-no, I cannot have anyone. *[He sees her, and a certain deference springs into him, although his worry remains.]* Why, Goody Putnam come in.

Figure 5. Editing cells' content and tags in theatrical texts

As outlined in Section 2, another interesting option included in the latest version of the program is that of adding tailor-made comments during the alignment process. In the 'Configure TAligner screen', the user can create labels for as many comments as they wish. Returning to the 'Alignment screen' and clicking on the edit option of the drop-down, the tool allows the user to edit the content (and tags, in case of theatrical texts) of the sentence, as well as to mark any part of the sentence and to tag it as a comment (*observación*). In the example shown in Figure 6, for instance, the German word combination *immer wieder* was highlighted with the label PU (phraseological unit).



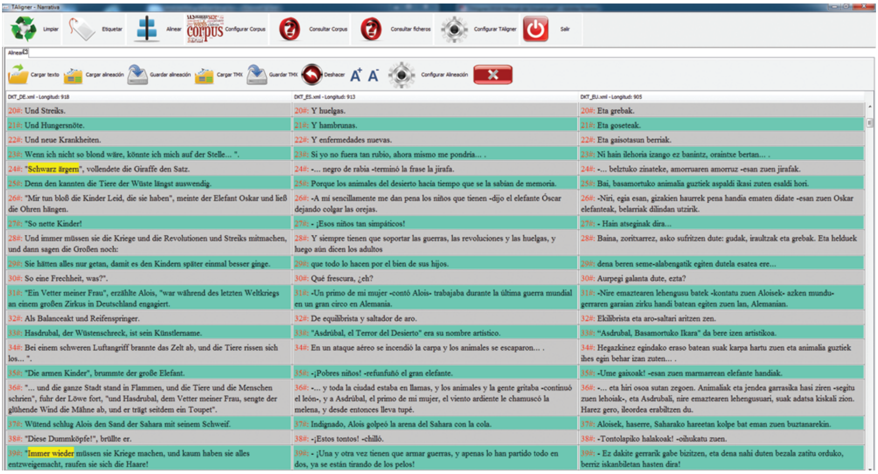
"<!TAT_COM type=2!><?TAT_COM?>Immer wieder<!TAT_COM_END!><?TAT_COM_END?> müssen sie Kriege machen, und kaum haben sie alles entzweigemacht, raufen sie sich die Haare!

null - Longitud: 0

DKT_DE.xml - Longitud: 918	DKT_ES.xml - Longitud: 913	DKT_EU.xml - Longitud: 905
22#: Und neue Krankheiten.	22#: Y enfermedades nuevas.	22#: Eta gaisotasun berriak.
23#: Wenn ich nicht so blond wäre, könnte ich mich auf der Stelle... ".	23#: Si yo no fuera tan rubio, ahora mismo me pondría... .	23#: Ni hain ilehoria izango ez banintz, oraintxe bertan... .
24#: "Schwarz ärgern", vollendete die Giraffe den Satz.	24#: -... negro de rabia -terminó la frase la jirafa.	24#: -... belztuko zinateke, amorruaren amorruz -esan zuen jirafak.
25#: Denn den kannten die Tiere der Wüste längst auswend		aspaldi ikasi zuten esaldi hori.
26#: "Mir tun bloß die Kinder Leid, die sie haben", meinte d die Ohren hängen.		pena handia ematen didate -esan zuen Oskar
27#: "So nette Kinder!		
28#: Und immer müssen sie die Kriege und die Revolutione und dann sagen die Großen noch:		dute: gudak, iraultzak eta grebak. Eta helduek
29#: Sie hätten alles nur getan, damit es den Kindern später		n dutela esatea ere...
30#: So eine Frechheit, was?".		
31#: "Ein Vetter meiner Frau", erzählte Alois, "war währen an einem großen Zirkus in Deutschland engagiert.		-kontatu zuen Aloisek- azken mundu- en zuen lan, Alemanian.
32#: Als Balanceakt und Reifenspringer.		n.
33#: Hasdrubal, der Wüstenschreck, ist sein Künstlername.		bere izen artistikoa.
34#: Bei einem schweren Luftangriff brannte das Zelt ab, u los... ".		suak karpa hartu zuen eta animalia guztiek
35#: "Die armen Kinder", brummte der große Elefant.		rrean elefante handiak.
36#: "... und die ganze Stadt stand in Flammen, und die Tie schrien", fuhr der Löwe fort, "und Hasdrubal, dem Vetter m glühende Wind die Mähne ab, und er trägt seitdem ein Toup		maliak eta jendea garrasika hasi ziren -segitu tearen lehengusuari, suak adatsa kiskali zion.
37#: Wütend schlug Alois den Sand der Sahara mit seinem Schweif.	37#: Indignado, Alois golpeó la arena del Sahara con la cola.	37#: Aloisek, haserre, Saharako hareetan kolpe bat eman zuen buztanarekin.
38#: "Diese Dummköpfe!", brüllte er.	38#: -¡Estos tontos! -chilló.	38#: -Tontolapiko halakoak! -oihukatu zuen.
39#: "Immer wieder müssen sie Kriege machen, und kaum haben sie alles entzweigemacht, raufen sie sich die Haare!	39#: - ¡Una y otra vez tienen que armar guerras, y apenas lo han partido todo en dos, ya se están tirando de los pelos!	39#: - Ez dakite gerrarik gabe bizitzen, eta dena nahi duten bezala zatitu orduko, berriz iskanbiletan hasten dira!

Figure 6. Adding comments in narrative texts

If we accept the addition of a comment (*guardar cambio*), the added comments are highlighted in yellow, as shown in Figure 7.

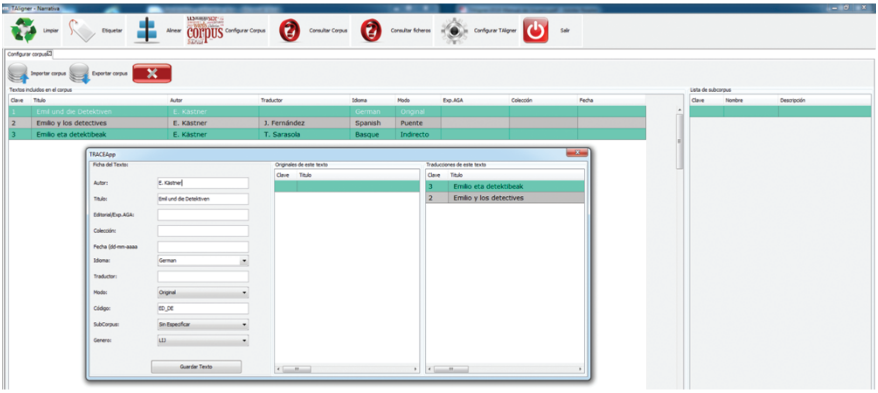


DKT_DE.xml - Longitud: 918	DKT_ES.xml - Longitud: 913	DKT_EU.xml - Longitud: 905
20#: Und Streiks.	20#: Y huelgas.	20#: Eta grebak.
21#: Und Hungersnöte.	21#: Y hambrunas.	21#: Eta goseteak.
22#: Und neue Krankheiten.	22#: Y enfermedades nuevas.	22#: Eta gaisotasun berriak.
23#: Wenn ich nicht so blond wäre, könnte ich mich auf der Stelle... ".	23#: Si yo no fuera tan rubio, ahora mismo me pondría... .	23#: Ni hain ilehoria izango ez banintz, oraintxe bertan... .
24#: "Schwarz ärgern", vollendete die Giraffe den Satz.	24#: -... negro de rabia -terminó la frase la jirafa.	24#: -... belztuko zinateke, amorruaren amorruz -esan zuen jirafak.
25#: Denn den kannten die Tiere der Wüste längst auswendig.	25#: Porque los animales del desierto hacía tiempo que se la sabían de memoria.	25#: Bai, basamortuko animalia guztiek aspaldi ikasi zuten esaldi hori.
26#: "Mir tun bloß die Kinder Leid, die sie haben", meinte der Elefant Oskar und ließ die Ohren hängen.	26#: -A mí sencillamente me dan pena los niños que tienen -dijo el elefante Óscar dejando colgar las orejas.	26#: -Niri, egia esan, gizakien haurrek pena handia ematen didate -esan zuen Oskar elefanteak, belarriak dilindan utzirik.
27#: "So nette Kinder!	27#: - ¡Esos niños tan simpáticos!	27#: - Hain atseginak dira...
28#: Und immer müssen sie die Kriege und die Revolutionen und Streiks mitmachen, und dann sagen die Großen noch:	28#: Y siempre tienen que soportar las guerras, las revoluciones y las huelgas, y luego aún dicen los adultos	28#: Baina, zoritxarrez, asko sufritzen dute: gudak, iraultzak eta grebak. Eta helduek
29#: Sie hätten alles nur getan, damit es den Kindern später einmal besser ginge.	29#: que todo lo hacen por el bien de sus hijos.	29#: dena beren seme-alabengatik egiten dutela esatea ere...
30#: So eine Frechheit, was?".	30#: Qué frescura, ¿eh?	30#: Aurpegi galanta dute, ezta?
31#: "Ein Vetter meiner Frau", erzählte Alois, "war während des letzten Weltkriegs an einem großen Zirkus in Deutschland engagiert.	31#: -Un primo de mi mujer -contó Alois- trabajaba durante la última guerra mundial en un gran circo en Alemania.	31#: -Nire emaztearen lehengusu batek -kontatu zuen Aloisek- azken mundu-gerraren garaian zirku handi batean egiten zuen lan, Alemanian.
32#: Als Balanceakt und Reifenspringer.	32#: De equilibrista y saltador de aro.	32#: Ekilibrista eta aro-saltari aritzen zen.
33#: Hasdrubal, der Wüstenschreck, ist sein Künstlername.	33#: "Asdrúbal, el Terror del Desierto" era su nombre artístico.	33#: "Asdrubal, Basamortuko Ikara" da bere izen artistikoa.
34#: Bei einem schweren Luftangriff brannte das Zelt ab, und die Tiere rissen sich los... ".	34#: En un ataque aéreo se incendió la carpa y los animales se escaparon... .	34#: Hegazkinez egindako eraso batean suak karpa hartu zuen eta animalia guztiek ihes egin behar izan zuten... .
35#: "Die armen Kinder", brummte der große Elefant.	35#: -¡Pobres niños! -refunfuñó el gran elefante.	35#: -Ume gaixoak! -esan zuen marmarrean elefante handiak.
36#: "... und die ganze Stadt stand in Flammen, und die Tiere und die Menschen schrien", fuhr der Löwe fort, "und Hasdrubal, dem Vetter meiner Frau, sengte der glühende Wind die Mähne ab, und er trägt seitdem ein Toupet".	36#: -... y toda la ciudad estaba en llamas, y los animales y la gente gritaba -continuó el león-, y a Asdrúbal, el primo de mi mujer, el viento ardiente le chamuscó la melena, y desde entonces lleva tupé.	36#: -... eta hiri osoa sutan zegoen. Animaliak eta jendea garrasika hasi ziren -segitu zuen lehoiak-, eta Asdrubali, nire emaztearen lehengusuari, suak adatsa kiskali zion. Harez gero, ileordea erabiltzen du.
37#: Wütend schlug Alois den Sand der Sahara mit seinem Schweif.	37#: Indignado, Alois golpeó la arena del Sahara con la cola.	37#: Aloisek, haserre, Saharako hareetan kolpe bat eman zuen buztanarekin.
38#: "Diese Dummköpfe!", brüllte er.	38#: -¡Estos tontos! -chilló.	38#: -Tontolapiko halakoak! -oihukatu zuen.
39#: "Immer wieder müssen sie Kriege machen, und kaum haben sie alles entzweigemacht, raufen sie sich die Haare!	39#: - ¡Una y otra vez tienen que armar guerras, y apenas lo han partido todo en dos, ya se están tirando de los pelos!	39#: - Ez dakite gerrarik gabe bizitzen, eta dena nahi duten bezala zatitu orduko, berriz iskanbiletan hasten dira!

Figure 7. Comments in narrative texts highlighted in yellow

3.2 Querying corpora

Once the texts are aligned, they must be uploaded to the corpus and it is necessary to manually establish the relationships between the texts (i.e. to determine the source and target texts). For that, each aligned text must first be uploaded to the corpus using the option *añadir a corpus*, as can be seen in the drop-down in Figure 4. Then, in the 'Configure Corpus screen', the source and target texts are manually identified (Figure 8). In this specific case, it has been determined that *Emilio eta detektibeak* and *Emilio y los detectives* are translations of the source text *Emil und die Detektiven*. As a result of doing this, the aligned texts appear together when querying the corpus.



Textos incluidos en el corpus

Clave	Título	Autor	Traductor	Idioma	Modo	Exp.AGA	Colección	Fecha
1	Emil und die Detektiven	E. Kastner		German	Original			
2	Emilio y los detectives	E. Kästner	J. Fernández	Spanish	Puente			
3	Emilio eta detektibeak	E. Kästner	T. Sarasola	Basque	Indirecto			

Lista de subcorpus

Clave	Nombre	Descripción

Ficha del Texto:

Autor: E. Kästner
Título: Emil und die Detektiven
Editorial/Exp.AGA:
Colección:
Fecha (dd-mm-aaaa
Idioma: German
Traductor:
Modo: Original
Código: ED_DE
SubCorpus: Sin Especificar
Genero: LIJ

Originales de este texto

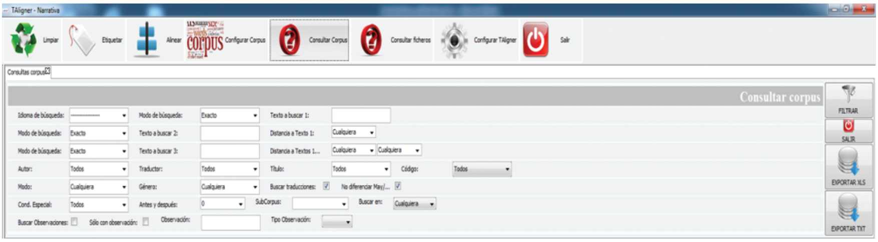
Clave	Título

Traducciones de este texto

Clave	Título
3	Emilio eta detektibeak
2	Emilio y los detectives

Figure 8. Establishing the relationship between the uploaded texts

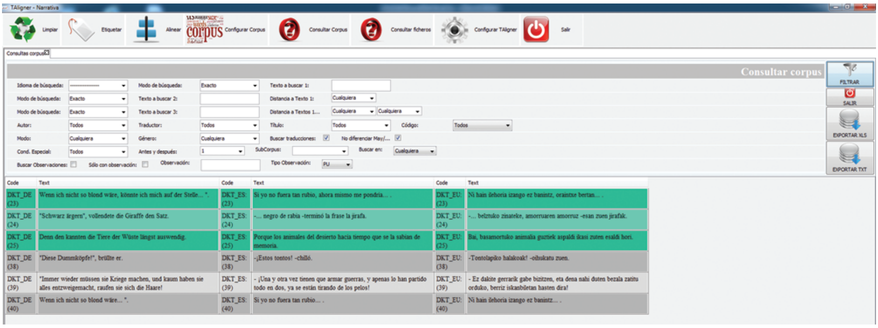
Once all texts are uploaded and the textual relations have been established, the user can start searching the corpus in the 'Queries screen' (*consultar corpus*). Simple and general queries can be made, but the tool also allows for more restricted queries due to the metadata included during the tagging process and the added comments. Figure 9 represents the searching interface of TAligner.



Idioma de búsqueda: ------------
Modo de búsqueda: Exacto
Texto a buscar 1:
Modo de búsqueda: Exacto
Texto a buscar 2:
Distancia a Texto 1: Cualquiera
Modo de búsqueda: Exacto
Texto a buscar 3:
Distancia a Textos 1...: Cualquiera Cualquiera
Autor: Todos
Traductor: Todos
Título: Todos
Código: Todos
Modo: Cualquiera
Género: Cualquiera
Buscar traducciones: ☑
No diferenciar May/...: ☑
Cond. Especial: Todos
Antes y después: 0
SubCorpus:
Buscar en: Cualquiera
Buscar Observaciones: □
Sólo con observación: □
Observación:
Tipo Observación:

Figure 9. Query options in TAligner 3.0

The user can enter any word, words (or part of words) in the entire corpus or apply filters such as author, translator, mode of translation, genre, title, or code. In addition, if comments were added during alignment, these can also be searched for (Figure 10). The two sentences containing the words highlighted as PUs (*schwarz ärgern* and *immer wieder*) can be observed in the search results in Figure 10, together with their corresponding Spanish and Basque translations. In order to widen the context, it was explicitly requested (*antes y después*) that previous and subsequent sentences be shown. Furthermore, in theatre corpora, a filter allows us to define if we are interested in searching whole texts, only stage directions or only dialogues. Previously added comments can also be searched. All results can be exported as XLS or TXT files.



Idioma de búsqueda: ----------------- Modo de búsqueda: Exacto Texto a buscar 1:
Modo de búsqueda: Exacto Texto a buscar 2: Distancia a Texto 1: Cualquiera
Modo de búsqueda: Exacto Texto a buscar 3: Distancia a Textos 1... Cualquiera Cualquiera
Autor: Todos Traductor: Todos Título: Todos Código: Todos
Modo: Cualquiera Género: Cualquiera Buscar traducciones: ☑ No diferenciar May/... ☑
Cond. Especial: Todos Antes y después: 1 SubCorpus: Buscar en: Cualquiera
Buscar Observaciones: □ Sólo con observación: □ Observación: Tipo Observación: PU

Code	Text	Code	Text	Code	Text
DKT_DE (23)	Wenn ich nicht so blond wäre, könnte ich mich auf der Stelle... ".	DKT_ES: (23)	Si yo no fuera tan rubio, ahora mismo me pondría... .	DKT_EU: (23)	Ni hain ilehoria izango ez banintz, oraintxe bertan... .
DKT_DE (24)	"Schwarz ärgern", vollendete die Giraffe den Satz.	DKT_ES: (24)	-... negro de rabia -terminó la frase la jirafa.	DKT_EU: (24)	-... belztuko zinateke, amorruaren amorruz -esan zuen jirafak.
DKT_DE (25)	Denn den kannten die Tiere der Wüste längst auswendig.	DKT_ES: (25)	Porque los animales del desierto hacía tiempo que se la sabían de memoria.	DKT_EU: (25)	Bai, basamortuko animalia guztiek aspaldi ikasi zuten esaldi hori.
DKT_DE (38)	"Diese Dummköpfe!", brüllte er.	DKT_ES: (38)	-¡Estos tontos! -chilló.	DKT_EU: (38)	-Tontolapiko halakoak! -oihukatu zuen.
DKT_DE (39)	"Immer wieder müssen sie Kriege machen, und kaum haben sie alles entzweigemacht, raufen sie sich die Haare!	DKT_ES: (39)	- ¡Una y otra vez tienen que armar guerras, y apenas lo han partido todo en dos, ya se están tirando de los pelos!	DKT_EU: (39)	- Ez dakite gerrarik gabe bizitzen, eta dena nahi duten bezala zatitu orduko, berriz iskanbiletan hasten dira!
DKT_DE (40)	Wenn ich nicht so blond wäre... ".	DKT_ES: (40)	Si yo no fuera tan rubio... .	DKT_EU: (40)	Ni hain ilehoria izango ez banintz... .

Figure 10. Search result in TAligner

4. Corpora created using TAligner 3.0

Although other corpora (Zubillaga et al. 2015; Arrula 2018; Merino-Álvarez and Andaluz-Pinedo 2017) were built using TAligner, the following sections will describe two corpora created by the authors of this paper.

4.1 Narrative corpora

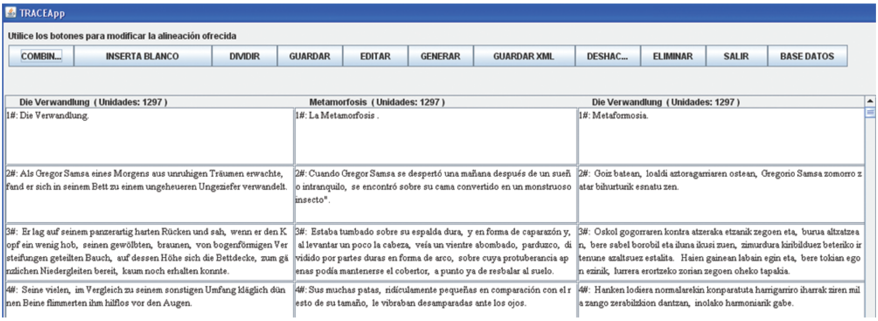
A corpus-based study to analyze the translation of phraseological units in German-into-Basque literary translations was performed by Sanz-Villar (2015). For that purpose, the AleuskaPhraseo parallel and multilingual corpus was compiled using TRACEAligner 2.0. Prior to the compilation of the corpus, several steps were taken following the TRACE research project's proposal.[8] In Sanz-Villar's (2015) case, the Aleuska catalogue, which had previously been built by Uribarri and

[8] A detailed description of these steps can be found in Sanz-Villar (2018).

updated by Zubillaga (2013), was updated. This catalogue consists of German literary texts translated into Basque. A part of the catalogue consisting of adult literature texts and children's literature texts was then described according to different criteria. Taking into account the features of this subcatalogue and the goals of the study, texts that would become part of the corpus were selected and the AleuskaPhraseo corpus was built up.

German source texts, Basque target texts and intermediary versions were included in this corpus. These were included as a preliminary analysis showed that the texts were indirectly translated through an intermediary version (predominantly through a Spanish translation). In general terms, the corpus consists of 34 assumed direct and 14 indirect translations. A diversity of authors is ensured and it consists of around 3.5 million words. It is not a large corpus, but with regard to corpus size, we absolutely agree with Anthony when he argues that "[t]he value of a corpus is clearly dependent not on its size but on what kind of information we can extract from it" (Anthony 2013: 146).

After following the steps described in Section 3.1 (digitising, cleaning, and aligning the 110 texts to the corpus) the texts were uploaded to a database (with the option *base datos* from Figure 11). The MySQL database management system was used to create the database. This meant that a SQL file containing all aligned corpus texts had to be created.



Utilice los botones para modificar la alineación ofrecida

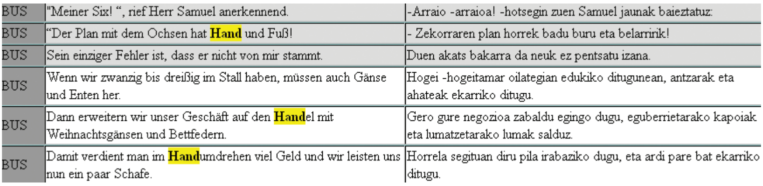
Die Verwandlung (Unidades: 1297)	Metamorfosis (Unidades: 1297)	Die Verwandlung (Unidades: 1297)
1#: Die Verwandlung.	1#: La Metamorfosis .	1#: Metaformosia.
2#: Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt.	2#: Cuando Gregor Samsa se despertó una mañana después de un sueño intranquilo, se encontró sobre su cama convertido en un monstruoso insecto".	2#: Goiz batean, loaldi aztoragarriaren ostean, Gregorio Samsa zomorro zatar bihurturik esnatu zen.
3#: Er lag auf seinem panzerartig harten Rücken und sah, wenn er den Kopf ein wenig hob, seinen gewölbten, braunen, von bogenförmigen Versteifungen geteilten Bauch, auf dessen Höhe sich die Bettdecke, zum gänzlichen Niedergleiten bereit, kaum noch erhalten konnte.	3#: Estaba tumbado sobre su espalda dura, y en forma de caparazón y, al levantar un poco la cabeza, veía un vientre abombado, parduzco, dividido por partes duras en forma de arco, sobre cuya protuberancia apenas podía mantenerse el cobertor, a punto ya de resbalar al suelo.	3#: Oskol gogorraren kontra atzeraka etzanik zegoen eta, burua altxatzean, bere sabel borobil eta iluna ikusi zuen, zimurdura kiribilduez beteriko irtenune azaltsuez estalita. Haien gainean labain egin eta, bere tokian egon ezinik, lurrera erortzeko zorian zegoen oheko tapakia.
4#: Seine vielen, im Vergleich zu seinem sonstigen Umfang kläglich dünnen Beine flimmerten ihm hilflos vor den Augen.	4#: Sus muchas patas, ridículamente pequeñas en comparación con el resto de su tamaño, le vibraban desamparadas ante los ojos.	4#: Hanken lodiera normalarekin konparatuta harrigarriro iharrak ziren mila zango zerabilzkion dantzan, inolako harmoniarik gabe.

Figure 11. Aligning screen of TRACEAligner 2.0 with the option of uploading the texts to a database

Once all texts were uploaded, searches for extracting PUs from the corpus were conducted using TRACEAligner 2.0's query engine. This was not integrated within the tool. Instead, the user had to open a link in a browser. Another disadvantage of this version (which has been overcome in the latest version of the program), was the length of time needed to receive query results. Depending on corpus size and type of query, it could take several minutes for the user to obtain results.

Phraseological units containing the word *Hand* in German and *esku* in Basque were first searched for in the corpus. Many researchers have analysed this type of

PU (somatic PUs containing a constituent that refers to a human body part), and it has been shown that they are prolific across many languages. A key-word-based extraction was conducted to obtain the PUs; i.e. the words *Hand* in German and *esku* in Basque (together with their variants) were searched for in the corpus. This first search created a lot of "noise", therefore results which did not represent a PU were manually excluded. As can be seen in Figure 12, the first and last results both represent PUs in German (*mit Hand und Fuß* and *im Handumdrehen*), but the one in the middle is just a word that begins with *Hand* (*Handel*).



BUS	"Meiner Six! ", rief Herr Samuel anerkennend.	-Arraio -arraioa! -hotsegin zuen Samuel jaunak baieztatuz:
BUS	"Der Plan mit dem Ochsen hat **Hand** und Fuß!	- Zekorraren plan horrek badu buru eta belarririk!
BUS	Sein einziger Fehler ist, dass er nicht von mir stammt.	Duen akats bakarra da neuk ez pentsatu izana.
BUS	Wenn wir zwanzig bis dreißig im Stall haben, müssen auch Gänse und Enten her.	Hogei -hogeitamar oilategian edukiko ditugunean, antzarak eta ahateak ekarriko ditugu.
BUS	Dann erweitern wir unser Geschäft auf den **Hand**el mit Weihnachtsgänsen und Bettfedern.	Gero gure negozioa zabaldu egingo dugu, eguberrietarako kapoiak eta lumatzetarako lumak salduz.
BUS	Damit verdient man im **Hand**umdrehen viel Geld und wir leisten uns nun ein paar Schafe.	Horrela segituan diru pila irabaziko dugu, eta ardi pare bat ekarriko ditugu.

Figure 12. A query result in TRACEAligner 2.0

Two further external tools were used to extract other phraseological units: AntConc and Foma. The former was used to extract binomials, a special type of PU that consists of two elements that belong to the same grammatical category. These two units are joined by a preposition or a conjunction (for instance, *safe and sound, back and forth*). The Cluster/N-Grams option was employed to extract this type of PU. The second NLP tool was useful in extracting predefined patterns, as thoroughly explained in Sanz-Villar (2019). First, the corpus was lemmatised and tagged at PoS level, and then a code to extract predefined patterns in Basque was written and processed using Foma. Once the PUs to be analysed were extracted with these external tools, they were searched for from the Query screen of TRACEAligner 2.0, together with the correspondent target text(s).

4.2 Theatre corpora

TAligner 3.0 was used to build a parallel corpus of theatre texts as part of an ongoing doctoral thesis. The corpus, called TEATRAD, is aligned at utterance level. This specific type of alignment draws on the consideration of the utterance (*réplica*) as the most appropriate unit for description and comparison of dramatic texts as proposed by Merino-Álvarez (1992: 285, 1994a: 397).

As with the AleuskaPhraseo corpus, TEATRAD corpus was also compiled following TRACE projects methodology, which continued to be applied within the TRALIMA/ITZULIK group (Merino-Álvarez 2017: 149). Thus, text selection derives from analysis of a catalogue of play-texts translations. It was created primarily for the analysis of certain cases, although it has the potential to be expanded and used for further purposes in the future.

The corpus is currently composed of original plays in English from the 1950s and 1960s and available (re)translations generated in Spanish in the 20th and 21st centuries. These translations include unpublished scripts used in theatre productions or published acting or reading editions. Titles by Arthur Miller, Tennessee Williams, Samuel Beckett and Harold Pinter comprise the corpus. Regarding size, it contains around 500,000 words.

A range of steps were undertaken to build this parallel corpus of theatre translations: Texts were digitised to TXT format, cleaned, tagged (divided into utterances, as well as character names, dialogues, and stage directions), aligned, and linked to their original or target texts. Since this process is mostly manual, it is time-consuming but the alignment would be more precise than if it was performed automatically, especially taking into account that some translations omit of several utterances of the original text.

Figure 13 shows the alignment of a source play and its translations at utterance level. In the editing window the tags which structurally mark up each text part (speakers, stage directions and dialogue) can be observed. This type of alignment and tags are specific for the compilation of corpora of dramatic texts, therefore they only operate if we select the corresponding text type at the beginning of the corpus building process (see Section 3.1).

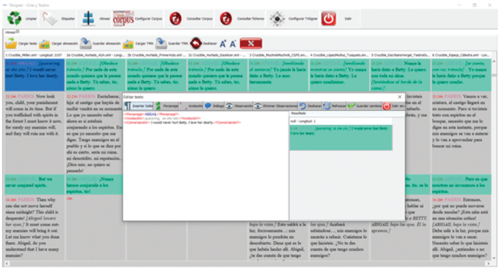


Figure 13. Tags and alignment of theatre texts in TAligner

The corpus thus created may be exploited in different ways within the program. For example, in the tool's query section, users may search words, parts of words or expressions from the text. The possibility of retrieving occurrences from an original text aligned with more than one target text is useful in this corpus which includes different retranslations, as it assists with the comparison of not only an original and a target text, but also of one translation with other translations of the same text. Another possibility is to use the aligning screen as a visualiser to compare utterances and make manual annotations. These will also be automatically retrieved and counted if we search for them in the query section of the tool.

One aspect which is being explored in the corpus is the translation of orality markers, characteristic features of dramatic texts. By way of example, Figure 14 shows the search of a vocative in Arthur Miller's play *The Crucible*, along with the results. The utterances which contain occurrences appear aligned with the corresponding ones from the translations. Since simultaneous alignment of more than two texts is possible in this tool, longer textual chains may be queried. The occurrences provide data for a comparative analysis of how this vocative was transferred in the different target texts.

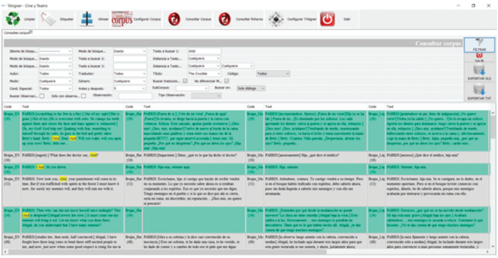


Figure 14. Occurrences of a query in TAligner

This query attempts to exemplify the possibility which TAligner 3.0 offers to facilitate descriptive analyses of translations, in relation not only to their original text but also to other translations.

5. Conclusions

TAligner 3.0 is a user-friendly tool to create and query parallel corpora. An effort was made to enhance the program in response to researchers' needs. This process benefited the evolution of the program as a whole and provided the software with some useful features for any user who wishes to conduct a study with it.

TAligner's main advantage, which sets it apart from most corpus management programs, is that it allows both corpus alignment and analysis within one tool. The main advantages of this integrated approach are, first, that it makes the process of creating and analysing corpora more user-friendly and flexible, reducing unnecessary technical complexities such as format conversions. Second, it guarantees that the parallel corpus is supported for analysis, since each specific option for corpus building is reflected in the analysis interface. Moreover, some options for corpus building are peculiar to this tool and they are linked with analysis options that are not found in other tools, such as structural mark-up and analysis filters. As Sanjurjo-González (2018: 25) points out, other programs which process parallel corpora, such as AntPConc (Anthony 2014), ParaConc (Barlow 1995), Sketch Engine (Kilgarriff et al. 2014) or CQPweb (Hardie 2012) require alignment to be performed externally and subsequently included in the tool, necessitating additional technical knowledge to carry out this process. On the other hand, other aligner tools do not include the possibility to make queries, such as TCA2 (Hofland and Johansson 1998) or LF Aligner (Farkas 2010). The integration of corpus alignment and query functions within the same program, without bringing up format conversion or other issues, is a distinct advantage for users. TAligner not only adds to the relatively small number of tools developed to analyse parallel corpora, but it also provides this helpful and uncommon option. As a side benefit, this opens the range of potential users from researchers to students, which might have implications for translation training environments.

In addition, the tool includes some distinctive features in the process of creating and querying parallel corpora. Firstly, it carries out segmentation and alignment in different structural units for each text type; i.e. it is genre-oriented. Prose texts are aligned at sentence level whereas theatrical texts take the utterance as the alignment unit. This alignment option adapts better to each text type structure and is supported by previous studies (e.g. Merino-Álvarez 1994a, 1994b; Pérez 2004; Bandín 2007). The tool also adds specific mark-ups for other structural parts of the texts, such as dialogues or stage directions, which allow for specific queries in those text sections. A second characteristic feature of corpus creation in TAligner is the possibility of aligning several texts simultaneously. This is especially useful for the study of indirect translations or retranslations of the same source text. To our knowledge, ACM (Sanjurjo-González 2017) is the only tool that allows the alignment of more than two texts (up to four). TAligner also allows manual alignment, which can be as rigorous as the user wishes. Finally, the program includes the option to design and use a set of observations as a form of tailor-made annotation. The program does not yet include some frequent types of annotation such as POS or semantic. This flexible option may be of interest to certain studies. In relation to corpus analysis, TAligner's main particularity is that its query screen is adapted to the previous characteristics. Searches take into account different structural mark-ups in each text, as well as the alignment of various texts. In addition, annotations can be queried and retrieved. Although the tool has mainly been used with English, Basque, Spanish and German texts, it is language independent, a characteristic which may also be considered an advantage for its general usage.

These options were found to be useful in the management of corpora within TRALIMA/ITZULIK. Indeed the use of TAligner 3.0 has facilitated the building and analysis of parallel and bi-/multilingual corpora in a range of studies, as exemplified through the narrative and theatre corpora presented. In addition, since no previous technical knowledge is required to apply the tool, it can be helpful in projects on a variety of scales and including various users. In this sense, as well as the aforementioned research studies, TAligner has been used in translator training lessons, as well as translation degree students' dissertations (e.g. Riobello Leiva 2016; Manso-Osma 2018).

Moreover, TAligner's usability transcends the immediate context of the research group, as this freeware tool is currently available at CorpusNet website as well as the digital archive (ADDI) of the University of the Basque Country for the whole research community. Other academics or students interested in compiling and searching parallel corpora may benefit from this easy-to-use program. As mentioned above, the combination of an aligner and a query section within the same tool is an advantage, and other distinctive characteristics of TAligner, such as the recognition of different structural and alignment units or the possibility of aligning more than two texts simultaneously, may also be regarded as helpful for certain studies. This would seem natural as shared research needs influenced the program's design.

However, the tool is still under development and some issues yet need to be solved. One possible area of improvement would be to overcome the program's limitations regarding annotation and statistics. In this sense it would be useful to include POS tagging and lemmatisation (which currently must be performed outside the tool), as well as further analysis options such as collocations, N-grams or keyword lists. Future work would also be directed towards adapting the interface and user manual to other languages so that the tool can be used in a wider variety of contexts. All in all, the intention is to continue to implement new functions, as well as apply the tool to further descriptive translation studies.

Acknowledgements

Research group TRALIMA/ITZULIK, GIU 16/48, University of the Basque Country, UPV/EHU, Basque Government consolidated research group IT1209-19; PIF17/46.

Funding

Red de Excelencia CorpusNet, funded by MINECO project FFI2016-81934-RED.

References

Anthony, Laurence. 2013. "A critical look at software tools in corpus linguistics." *Linguistic Research* 30 (2): 141–161. https://doi.org/10.17250/khisli.30.2.201308.001

Anthony, Laurence. 2014. AntPConc [Computer software].

Anthony, Laurence. 2017. "Introducing corpora and corpus tools into the technical writing classroom through Data-Driven Learning (DDL)." In *Discipline-Specific Writing: Theory into practice*, ed. by John Flowerdew, and Tracey Costley, 162–180. Abingdon: Routledge.

Arrula, Garazi. 2018. Autoitzulpenaren teoria eta praktika Euskal Herrian / Theory and practice of self-translation in the Basque Country. Doctoral dissertation. http://hdl.handle.net/10810/27983

Bandín, Elena. 2007. Traducción, recepción y censura de teatro clásico inglés en la España de Franco. Estudio descriptivo-comparativo del Corpus TRACEtci (1939–1985). León: Universidad de León. Doctoral dissertation. https://buleria.unileon.es/handle/10612/1885

Barambones, Josu. 2009. La traducción audiovisual en ETB-1: Estudio descriptivo de la programación infantil y juvenil. Doctoral dissertation. http://hdl.handle.net/10810/12182

Barlow, Michael. 1995. ParaConc [Computer software].

Cabanillas González, Candelas. 2016. La traducción audiovisual en ETB2: Estudio descriptivo del género Western. Vitoria-Gasteiz: Universidad del País Vasco UPV/EHU. Doctoral dissertation. http://hdl.handle.net/10810/18287

Farkas, Andreas. 2010. LF Aligner [Computer software]

Gutiérrez Lanza, Camino, Elena Bandín Fuertes, José Enrique García González, and Sergio Lobejón Santos. 2015. "Desarrollo de software de etiquetado y alineación textual: TRACE Corpus Tagger/Aligner 1.0©." *II Congreso Internacional de Humanidades Digitales Hispánicas: Innovación, globalización e impacto*. Madrid: UNED. http://hdh2015.linhd.es/ebook/hdh15-gutierrezlanza.xhtml

Hardie, Andrew. 2012. CQPweb [Computer software].

Hofland, Knut and Johansson, Stig. 1998. "The Translation Corpus Aligner: A program for automatic alignment of parallel texts". In *Corpora and Cross-linguistic research. Theory, Method, and Case Studies*, eds. Stig Johansson and Signe Oksefjell, 87–100. Amsterdam; Atlanta: Rodopi.

Kenny, Dorothy. 2000. "Translators at a play: exploitations of collocational norms in German-English translations." In *Working with German corpora*, ed. by Bill Dodd, 143–160. Birmingham: The University of Birmingham Press.

Kenny, Dorothy. 2001. *Lexis and Creativity in Translation. A corpus-based Approach*. Manchester: St. Jerome.

Kilgarriff, Adam et al. 2014. Sketch Engine. [Computer software].

Lobejón-Santos, Sergio. 2013. Traducción inglés-español y censura de textos poéticos (1939–1978): TRACEpi (1939–1978). León: Universidad de León. Doctoral dissertation. http://hdl.handle.net/10612/6133

Manso-Osma, Aritz. 2018. Análisis de la traducción de compuestos alemanes del lenguaje científico-técnico mediante corpus paralelo de aprendices. BA dissertation http://hdl.handle.net/10810/30031

Manterola, Elizabete. 2011. Euskal literatura beste hizkuntza batzuetara itzulia. Bernardo Atxagaren lanen itzulpen moten arteko alderaketa. Doctoral dissertation. https://addi.ehu.es/handle/10810/12382

McEnery, Tony, and Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Merino-Álvarez, Raquel. 1992. "Rewriting for the Spanish stage." *KOINÉ. Annali della Scuola Superiore per Interpreti e Traduttori San Pellegrino* 2 (1–2): 283–289.

Merino-Álvarez, Raquel. 1994a. "La réplica como unidad de descripción y comparación de textos dramáticos traducidos." In *IV Encuentros Complutenses en torno a la traducción*, ed. by Margit Raders, and Rafael Martín-Gaitero, 397–404. Madrid: Editorial Complutense.

Merino-Álvarez, Raquel. 1994b. *Traducción, tradición y manipulación. Teatro inglés en España 1950–1990*. León: Universidad de León / Lejona: Universidad del País Vasco.

Merino-Álvarez, Raquel. 2017. "Traducción y censura: investigaciones sobre la cultura traducida inglés-español (1938–1985)." *Represura. Revista de Historia Contemporánea española en torno a la represión y la censura aplicadas al libro*, 2: 139–163. http://www.represura.es/represura_2_nueva_epoca_2017.pdf

Merino-Álvarez, Raquel, and Olaia Andaluz-Pinedo. 2017. "Peter Shaffer en la cultura española." *Creneida. Anuario de Literaturas Hispánicas*, 5: 239–278. http://www.creneida.com/revista/creneida-5-2017/

Merino-Álvarez, Raquel, and Olaia Andaluz-Pinedo. 2020. "60 años de Beckett en España: Esperando a Godot, de la censura a la audiodescripción." In *Samuel Beckett: Literatura y Traducción; Literature and Translation; Littérature et Traduction*, ed. by Bernardo Santano, 37–57. Bern: Peter Lang.

Pérez, M. 2004. *Traducciones censuradas de teatro norteamericano en la España de Franco (1939–1963)*. Bilbao: Universidad del País Vasco.

Riobello Leiva, Janire. 2016. Análisis de la traducción alemán-inglés y alemán-español de las partículas modales en obras de Kafka. BA dissertation. https://addi.ehu.es/handle/10810/21195?locale-attribute=en

Sanjurjo-González, Hugo. 2017. ACTRES Corpus Manager. [Computer software].

Sanjurjo-González, Hugo. 2018. *Creación de un framework para el tratamiento de corpus lingüísticos*. León: Universidad de León, Área de Publicaciones.

Sanz-Villar, Zuriñe. 2015. Unitate fraseologikoen itzulpena: alemana-euskara. Literatur testuen corpusean oinarritutako analisia. Doctoral dissertation. http://hdl.handle.net/10810/15128

Sanz-Villar, Zuriñe. 2018. "Diseño, descripción y análisis de un corpus multilingüe (alemán-español-euskera)." *TRANS* 22: 133–148. https://doi.org/10.24310/TRANS.2018.v0i22.2964

Sanz-Villar, Zuriñe. 2019. "An Overview of Basque Corpora and the Extraction of Certain Multi-Word Expressions from a Translational Corpus." In *Parallel corpora for contrastive and translation studies: New resources and applications*, ed. by Irene Doval Reixa, and M. T. Sánchez Nieto, 233–247. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.90.14san

Scott, Mike. 2004. WordSmith Tools. [Computer software].

Toury, Gideon. 2004. *Estudios descriptivos de traducción y más allá*. Madrid: Cátedra.

Toury, Gideon. 2012. *Descriptive Translation Studies and Beyond*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/btl.100

Uribarri, Ibon. 2016. "Taligner: itzulpen corpus eleaniztunak sortzeko tresna." *Senez*, 47: 251–265. https://eizie.eus/eu/argitalpenak/senez/20161103/17uribarri

Xiao, Richard, and Ming Yue. 2009. "Using corpora in translation studies: The state of the art". In *Contemporary Corpus Linguistics*, ed. by Paul Baker, 237–261. London; New York: Continuum.

Zubillaga, Naroa. 2013. Alemanetik euskaratutako haur- eta gazte-literatura: zuzeneko nahiz zeharkako itzulpenen azterketa corpus baten bidez. Doctoral dissertation. http://hdl.handle.net/10810/12431

Zubillaga, Naroa, Zuriñe Sanz-Villar, and Ibon Uribarri. 2015. "Building a trilingual parallel corpus to analyse literary translations from German into Basque." In *New directions in corpus-based translation studies*, ed. by Claudio Fantinuoli, and Federico Zanettin, 71–92. Berlin: Language Science Press. http://langsci-press.org/catalog/book/76

Developing a corpus-informed tool for Spanish professionals writing specialised texts in English

María Pérez Blanco and Marlén Izquierdo
ULE / UPV/EHU

This chapter describes the development and use of Promociona-TÉ, a tool for Spanish professionals who need to write herbal tea promotional texts (HTPTs) in English. The tool was developed from a comparable-corpus approach, given the absence of parallel corpora for under-researched domains (Skadiņa et al. 2010) such as HTPTs. It is the result of applying the insights of contrastive linguistic research in the workplace towards overcoming problems of communication (Rabadán 2019). Two analyses were conducted at the rhetorical and lexicogrammatical levels, providing the tool with a prototypical macrostructure of HTPTs together with a pool of drafting lines and a domain-specific bilingual glossary. We conclude by describing three advantages of this text generator over the use of translation, either human or machine.

Keywords: Promociona-TÉ, corpus-informed, writing tool, specialised discourse, knowledge transfer, comparable corpus, ACTRES

1. Introduction

Scientific research, in the so-called soft sciences as much as the hard sciences, can be said to be applied in nature when it seeks to satisfy an existing need in the real world. English, now unquestionably the world's *lingua franca* (ELF), has become the means of cross-cultural communication that guarantees communicative effectiveness, irrespective of varying surface-level features (Jenkins 2009). Culture itself needs to be understood in two ways, as big culture and small culture (Atkinson 2004). The former relates to national culture; the latter may be seen in a variety of spheres, such as the second-language learning classroom (Colombo 2012), a given discipline (Hyland 2000), or a professional domain (Bhatia 2008). Globalisation, together with technological innovation, has broadened not only our communication

https://doi.org/10.1075/btl.158.06per

needs but also the different ways in which communication takes place, with new text types emerging in different contexts in response to newly arising needs of use. Consequently, text types that have not thus far been seen as meriting investigation are now beginning to attract the interest of those engaged in applied linguistic research. Such is the case with online food descriptions, a promotional sub-genre that many local and regional companies seek to use as a means of breaking into global markets. When such commercial expansion relies on the use of ELF, successful communication will depend not only on the accurate transmission of relevant subject-specific information within the professional domain, but also on compliance with cultural conventions, both at the big and the small culture levels. To this end, acceptable language use, plus an awareness of genre conventions, are paramount.

In the humanities, the ACTRES research group has successfully transferred knowledge from academia to industry,[1] an alliance demanded by language users of different kinds (Anthony 2016). Indeed, ACTRES has demonstrated the usefulness and usability of contrastive research from a corpus approach in the development of a variety of language applications for different purposes: translation quality assessment (Rabadán et al. 2014), English language learning and teaching (Rabadán & Izquierdo 2012) and professional writing aids (cf. Footnote 1). This paper deals with one example of the latter, namely, Promociona-TÉ, a generator of herbal tea promotional texts. This text type is representative of the abovementioned sub-genre of online food descriptions.

Applied linguistics, and especially contrastive research, profit enormously from the use of (bi/multilingual) corpora (Hunston 2002; Johansson 2007). Although different in purpose and in defining characteristics, both parallel and comparable corpora have prominent roles in translation and contrastive studies, as attested by the increasing number of practical applications in both fields (Doval & Sánchez Nieto 2019). In particular, comparable corpora offer an endless source of authentic and reliable data that allow researchers to observe linguistic features by comparing original texts that are not conditioned by translation effects. Taking advantage of an *ad hoc* English-Spanish comparable corpus (cf. 4.2), we have developed the corpus-informed writing tool described in this paper.

[1] ACTRES, *Análisis Contrastivo y Traducción Inglés-Español* (English-Spanish Contrastive Analysis and Translation) is a research group that has been led by Prof. Rabadán (ULE) since the mid-1990s, and includes researchers from various universities. See its website for a detailed description of the various applications developed: (https://actres.unileon.es/wordpress/?page_id=434&lang=es).

2. Parallel corpora: Applications in cross-linguistic research

The rapid development of parallel corpora in recent decades has revolutionised linguistic studies. Almost twenty years ago, Borin (2002) noted that parallel corpus linguistics seemed to have emerged as "a distinct field of research within corpus linguistics, itself a fairly young discipline" (p. 1), given its continuous expansion as a result of the proliferation of projects on the creation, annotation and processing of parallel corpora. In fact, a wide variety of parallel corpora have been built in recent years for different languages and purposes (Doval & Sánchez Nieto 2019).

The potential applications of parallel corpora have grown over time. First, they play a major role in cross-linguistic studies, in that they provide a sound basis for contrastive analysis. Parallel corpora, which contain collections of original texts and their translations into one or more other languages, allow us to shed light on "how the same content is expressed in two languages" (Aijmer & Altenberg 1996: 13). Translation paradigms have been seen as a means of establishing cross-linguistic correspondences (James 1980), and the linguistic choices that translators make "inadvertently supply evidence of the meanings of the forms they are receiving and producing" (Noël 2003: 757). As a consequence, parallel corpora are of direct benefit to the fields of translation practice and training, lexicography, and foreign language teaching. In particular, they are of considerable help to human translators in improving the precision of their work in areas such as terminology and phraseology, since they provide "greater certainty as to the equivalence of particular expressions" (Aston 1999: 303).

More recently, parallel corpora have become a unique source for Machine Translation (MT) and bi/multilingual Natural Language Processing (NLP). Both Example-Based Machine Translation (EBMT) and Statistical Machine Translation (SMT) benefit from access to large databases of already-translated texts that provide essential training data for SMT or EBMT models. Likewise, they allow for lexical and terminological extraction. Parallel corpora have also been used to develop Computer-Assisted Translation (CAT) tools such as translation memories, bilingual concordances and translation-oriented word processors.

On the other hand, parallel corpora require alignment tools, desirably annotation – as in any other upgraded monolingual corpus-, and a corpus query system that browses two datasets simultaneously, which makes them time-consuming to develop (Mitkov 2018). In addition, such corpora may not be feasible for all types of texts. Therefore, the availability of translations is a key factor that determines not only the compilation of parallel corpora but also language applications based on them. With regard to the former, i.e., the digitisation of a corpus, successful alignment programs have been developed. On the basis of the algorithm developed for the Canadian Hansards (Gale & Church 1993), Johansson & Hofland built the Translation Corpus Aligner (TCA) for the English-Norwegian Parallel Corpus (1998), one of the first automatic tools used for aligning an original-translation textual pair at the sentence level. Upgraded versions have subsequently been used successfully with other language pairs like English-Portuguese (Santos & Oksefjell 2000), English-Spanish (Izquierdo et al. 2008), and Norwegian-Spanish (Hareide & Hofland 2012). More recently, the TRALIMA[2] research group has developed TAligner (Sanz-Villar 2019) for the simultaneous alignment of an original textual unit and two or more translations. Most importantly, TAligner has been used with various languages, such as German and Basque, in addition to English and Spanish. This is an interesting feature because most programs of this kind rely, among other things, on anchor word lists that users would feed the program with. Such lists will necessarily differ from one language pair to another. In addition, aligners may also work on a sentence-length basis, which might prove to be challenging if translation takes place between languages that are typologically different, such as English and Russian, or Spanish and Basque. Corpus querying is not without its difficulties either. The fact that most systems are developed for monolingual corpora adds a further technical challenge in parallel corpus compilation. P-ACTRES 2.0 (Sanjurjo-González & Izquierdo 2019) is browsed using an adaptation of Corpus Web Bench (CWB), originally developed for monolingual corpus exploitation.

On the other hand, leaving aside the costs involved (Zanettin 2012) and the unavoidable problem of *translationese* (Baker 1993; McEnery & Xiao 2008), the task of building a large parallel corpus is complex in that it demands the availability of translations. Corpus size has been acknowledged as one of the traditional drawbacks of parallel corpora (Tiedemann 2011; Zanettin 2012) due to the scant availability of translated texts. As Johansson (2007) observes, the range of translated texts is usually more restricted than original texts. On the same lines, it would (also) be very desirable to have access to as many versions of a translation of the same source texts as possible, rather than a one-to-one parallel corpus, so that correspondences do not represent just one individual's perspective (Malmkjær 1998). However, the reality is that most texts (except literary works) are translated only once (McEnery & Xiao 2008) and some types of texts are never translated, as is the case with, for example, almost all newspaper opinion articles (Pérez Blanco 2018). Indeed, despite the technological advances and the development of corpus linguistics and applied linguistics over the last two decades, Doval and Sánchez Nieto note that there are still contexts where "no translation evidence is readily accessible" (2019: 9). That is the scenario for under-resourced and minority languages, but the same is true

[2] TRALIMA, *Traducción, Literatura y Medios Audiovisuales* (Translation, Literature and Media), is a research group (GIU 16/48) based in the University of the Basque Country (UPV/EHU), in Spain.

especially for texts that belong to a specific, narrow domain (Skadiņa et al. 2010), research on which has thus far not been necessary.

According to Skadiņa et al. (2010), the availability of corpora dramatically affects the quality of current data-driven MT systems; whereas results are quite good for language pairs with large corpora available (e.g. English and French), MT systems are barely usable for under-resourced languages and domains. In such research contexts, comparable corpora might be regarded as an alternative and a more promising approach (Mitkov 2018) (cf. 4.2.).

3. Comparable corpora: Applications in cross-linguistic research

Although translators undoubtedly welcome parallel corpora as a rich source of real translation solutions, we should also bear in mind that "the earliest translation-oriented corpora were not parallel but comparable" (Marco 2019: 40). Even today, the value of comparable corpora in translation research should not be underestimated. First, it is easier to compile collections of original texts in two or more languages which belong to the same genre, domain, and sampling period, and between which a translation commission might be undertaken. Second, comparable corpora provide data on authentic language use that is supplemented with data from parallel corpora. In this sense, their value for contrastive analysis is unquestionable, as they avoid the effect of *translationese*. Third, the combination of both types of corpora allows for data triangulation (Marco 2019) where the explanatory potential of parallel corpora can account for findings of comparable corpus research. Most importantly, a comparable corpus might be the only way of resolving the problems that arise from the absence of a parallel corpus, which may certainly be the case for some under-researched languages and specialised domains.

In fact, specialised comparable corpora have proven to be particularly helpful for highly domain-specific translation tasks, given that technical language here might be as 'foreign' for the translator as any other, unknown foreign language (McEnery and Xiao 2008). Corpus-aided translations are of higher quality with respect to the understanding of a particular field (Bowker 1998), and translators with access to comparable corpora are thus able to improve their productivity and reduce the number of errors they make (McEnery & Xiao 2008).

Comparable corpora might also compensate for the absence of a parallel corpus in contexts where the latter is neither available nor is it feasible to build one. Borin's (2002) distinction between parallel corpora as a tool for exploring linguistic phenomena, on the one hand, and as a valuable source of data in computational linguistics, on the other, can certainly be applied to comparable corpora. In this regard, an alliance between computational linguists and corpus linguists may prove to be fruitful for parallel corpora. (Corpus) linguists may provide high quality material to be exploited (by computational linguists) with the assistance of machines. The writing tool described in this paper illustrates the collaboration between corpus linguistics and computational linguistics, yet represents an intermediate stage towards further synergies in the near future for the development of a (domain specific) controlled natural language for a restricted promotional sub-genre, namely, online food descriptions.

These corpus-informed applications attest to the potential value of comparable corpora for cross-linguistic communication. Capitalising on previous descriptive work, a number of text generators have been built by the ACTRES research group to assist Spanish professionals in writing technical and specialised texts in English in different fields, starting with electronic product descriptions, and wine tasting notes for online promotional descriptions of various kinds of food and drink (cf. Footnote 1). These text generators, which implement rhetorical and lexicogrammatical data extracted from English-Spanish comparable corpora, are conceived of as professional writing aids for contexts where professional translation is not always a possibility due to costly translation services and/or the genre and/or domain specificity of the translation commission. Given that the end users are not linguists, the writing tool supports authors by guiding them through a user-friendly interface which offers suggestions both in terms of prototypical text structure and text type-specific drafting lines (cf. 4.3).

The development of these "content-rich applications", built to satisfy a real communication need in the workplace, also represents a step forward in the "automation for restricted domains and genres" (Rabadán 2019: 59). Nevertheless, a thorough discussion of the contribution of current research to NLP is beyond the scope of this paper. Instead, we will focus on the description of Promociona-TÉ, a corpus-informed writing tool customised for small businesses as an alternative to translation.

4. Promociona-TÉ: A comparable-corpus-informed writing tool

4.1 Writing in English for specific purposes

One effect of globalisation is the increasing need for non-native speakers (NNS) to skilfully communicate in English. Writing is perhaps the most difficult language skill to master fully; it is the most laborious of the skills, the slowest to develop, and the last to be learned (Sheerin 2008). In addition, its communicative effect is not immediate, which requires that there be a high degree of clarity, efficiency and a lack of ambiguity for successful communication when the message is decoded

(Van Geyte 2013). (English) writing is fundamental for a range of users and in a variety of contexts, from students in higher education to professionals in the workplace. Indeed, more and more NNS of English need to use this language for various purposes, not only for general tasks but, more challengingly, for specific, work-related ones. The linguistic competence of NNS of English, which outnumber native speakers (NS) (Ethnologue 2018), may be insufficient when it comes to using a specialised variety of the language. This limitation in communicative competence may be due to the fact that the teaching of English for Specific Purposes (ESP) has often been carried out in isolation, ignoring authentic professional practice (Bhatia 2008). Consequently, when embarked on real business communication, many professionals realise that their English is too general to successfully produce specialised texts in the written mode. Whereas a solution to this problem might be for ESP teachers to simulate real workplace writing practices in the classroom, this would only solve part of the problem. Professionals who are no longer students also need assistance in their everyday, specialised writing tasks. It is in the workplace that they face the challenge of written communication, and thus it is here that solutions are required. This is why we have developed Promociona-TÉ, an ESP writing tool for the workplace.

The application has been created for NNS with a B2 level of English that work for small to medium-sized local companies within the tea industry.[3] By local we refer to businesses located in the region of León, as the underlying project is funded by the local government of this region of Spain.[4] With an interest in expanding their operations outside Spain, these companies need to write promotional texts of their products in English. The choice of this language reflects a business plan of achieving a presence in English-speaking markets. In addition, promotional texts might also be written in English to facilitate subsequent, indirect translations between languages with a smaller international projection. In other words, instead of translating from Spanish into, say, Arabic, English would become an intermediary code that makes translation possible.

Promociona-TÉ is the result of an agreement between the ACTRES research group and Pharmadus Botanicals, S.L., a tea manufacturer with operations in the United States and the United Arab Emirates, amongst others.[5] The tool generates herbal tea promotional texts (HTPTs), a subgenre of so-called online food descriptions.

[3] According to the Common European Framework of Reference for Languages.

[4] Project LE227U13, Castilla y León.

[5] https://www.pharmadus.com/

Because this study arises from a recently identified real-world need, HTPTs have not previously been of scholarly interest, nor are they commonly commissioned as translations. Hence they constitute an under-researched, narrow domain. Skadina et al. (2010) note the difficulty, even impossibility, of having translations of textual practices representative of these kinds of narrow domains. Consequently, there are no human-produced parallel texts of HTPTs that make possible machine translations of these texts for professional use. Nevertheless, even if this were a possibility, professionals for which this tool was developed could not normally afford to have the HTPTs translated by a human translator. Likewise, they could not meet the expenses of machine translation and post-editing. This is why the tool, which aims to be an alternative to translation, was by necessity built using a comparable corpus approach.

4.2 The corpus: ACTEaS_Promo

ACTEaS_Promo is a purpose-built comparable corpus of online tea descriptions written in British English (EN) and European Spanish (ES). At the time the writing tool was developed, the corpus contained 150 texts per language, with an overall size of 36,266 words, these being quite evenly distributed: 17,673 words in EN and 18,593 words in ES.[6] All the texts were downloaded from the website of 24 different tea brands and tea shops; we randomly chose 12 sources per language. Together with linguistic variety, another criterion was that a good selection of teas be represented, including not only different types such as black, green, and blue tea, among others, but also herbal and fruit teas. We therefore discarded different descriptions of the same tea type. That is, we could have included different brand descriptions of, for example, 'mint tea' in the corpus, yet we decided not to do this, on the assumption that a greater variety of teas would provide us with a wider array of register-specific lexical units. This approach enabled us to equip the generator with a built-in glossary composed of food-related words and multi-word expressions (cf. 4.3.2). Most importantly, we aimed for corpus representativeness with a careful choice of texts. The size of the corpus was also considered in this regard. ACTEaS_Promo was compiled with clear research aims in mind. As such, the corpus should be tagged at the rhetorical level in order to identify the prototypical macrostructure of a specific genre. Hence, a large corpus is neither necessary (Biber & Conrad 2009) nor manageable (López Arroyo & Roberts 2014). To account for corpus representativeness in terms of size, we applied the statistical formula $\varepsilon = \frac{z_{\alpha/2} \cdot \sqrt{p \cdot q}}{\sqrt{n}}$ to calculate the margin of error (ε), setting the

[6] By July 2019, the corpus was larger (145,601 words) and is still under construction.

confidence level at 95%. $z_{\alpha/2} = 1.96$ is the value of the z standard normal distribution, and α is the complementary value to the confidence level, in this case, 0.05; p is the estimated proportion of the population and q equals 1-p, and the value for both is 0.5; and n stands for the actual population per language. Accordingly, it was estimated that 150 texts per language, with an average of 121 words, would guarantee representativeness for the study. In fact, the actual margin of error is 0.72 for the EN data and 0.74 for ES, both well above 95%.

ACTEaS_Promo was used in a multilevel corpus-driven analysis using ad-hoc tools that enabled the tagging of all the texts as well as the browsing of register-specific recurrent patterns. The following section shows how all the linguistic-descriptive work was useful and usable in the implementation of Promociona-TÉ.

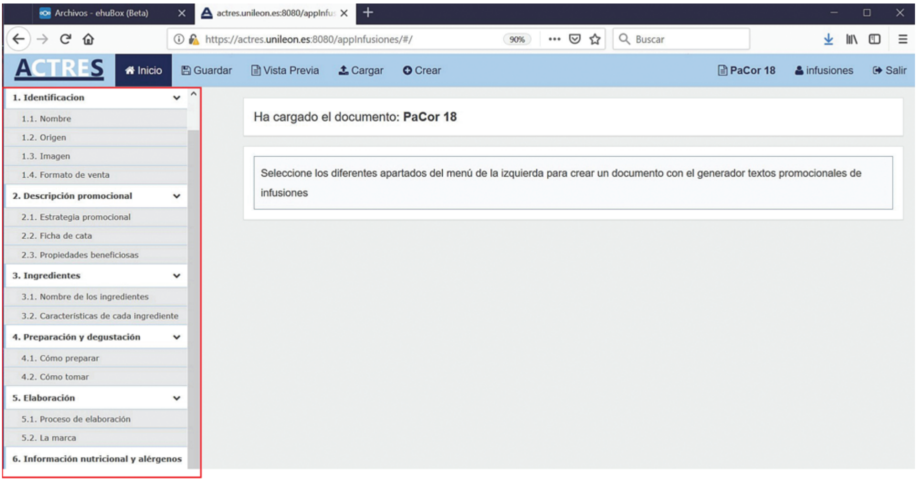
4.3 Generating herbal tea promotional texts

HTPTs encode informational-persuasive discourse (Biber & Zhang 2018), whereby a given tea is described and evaluated in a positive way for persuading potential customers to try the product. HTPTs are multimodal texts; they contain written text, pictures, multimedia resources like videos, plus hyperlinks that let the web user, potentially a tea consumer, interact dynamically with the text (Lam 2013). All these semiotic devices, the textual ones, multimedia resources, and even the overall web layout, have communicative functions. Yet, we believe that it is the purely linguistic elements that are most likely to pose a clear problem of communication for English NNS professionals. The scope of this study, therefore, is the linguistic composition of the HTPT, examined from a top-down approach.

To equip the tool with all the linguistic elements needed for writing acceptable, accurate and conventional HTPTs in English, we carried out two complementary analyses on the corpus. First, we conducted a contrastive rhetorical analysis to identify and tag functional chunks that recur from text to text. At this stage, the widely accepted move analysis methods by Swales (1990) and Bhatia (2004) were most useful. Second, we carried out a register analysis (Biber & Conrad 2009) of prototypical moves and steps to observe pervasive language features at the terminological and lexicogrammatical levels.

4.3.1 *Prototypical macrostructure*

Tagging at the rhetorical level was done semi-automatically using the ACTRES Tagger (Izquierdo & Pérez Blanco 2020). This process yielded a prototypical macrostructure of HTPTs that is common to EN and ES texts. When we begin writing, the tool makes the user aware of such a structure by referring to a display menu on the left hand side of the interface; its six building moves are shown, and the user can click on each to display their steps (see Figure 1).


Archivos - ehuBox (Beta)
actres.unileon.es:8080/appInfu
https://actres.unileon.es:8080/appInfusiones/#/
90%
Buscar
ACTRES
Inicio
Guardar
Vista Previa
Cargar
Crear
PaCor 18
infusiones
Salir
1. Identificacion
1.1. Nombre
1.2. Origen
1.3. Imagen
1.4. Formato de venta
2. Descripción promocional
2.1. Estrategia promocional
2.2. Ficha de cata
2.3. Propiedades beneficiosas
3. Ingredientes
3.1. Nombre de los ingredientes
3.2. Características de cada ingrediente
4. Preparación y degustación
4.1. Cómo preparar
4.2. Cómo tomar
5. Elaboración
5.1. Proceso de elaboración
5.2. La marca
6. Información nutricional y alérgenos

Ha cargado el documento: **PaCor 18**

Seleccione los diferentes apartados del menú de la izquierda para crear un documento con el generador textos promocionales de infusiones

Figure 1. HTPT prototypical macrostructure and writing menu displayed

The rhetorical analysis carried out in this prior descriptive stage was helpful for our tool in two ways. First, it provided us with information about the actual rhetorical realisation in each language sample, making it possible to identify cross-linguistic differences. Second, juxtaposition brought to light potential difficulties for the end user, namely, Spanish professionals.

In particular, it was observed that not all moves are equally necessary; also, they are not always all present, and the information specified is not always essential. This is partly due to the fact that the content provided in one given segment, for example *Step 3.2.Características de cada ingrediente* (Ingredient's features) within M3-Ingredients, may be referred to in another chunk, in this case within M2 in *Step 2.2. Ficha de cata* (Tasting note). Therefore, the end user may not find it necessary to fill in all the moves and steps shown in the display menu to produce an acceptable HTPT. That said, it is very much recommended that at least M1-Identification and M2-Promotional Description be filled in, because their occurrence rate in the corpus is 100%. Most importantly, M6-Nutrition & Allergies is obligatory according to EU regulation 1169/2011.[7] By contrast, fewer promotional texts in the EN sample contain suggestions on how to consume tea (step 4.2.) than in the ES sample.

[7] EU regulation 1169/2011, chapter I, article I (2a) p. 25.

Table 1. Realisation of HTPT moves and steps per language

Move Step	EN realisation rate	ES realisation rate
M1 – IDENTIFICATION	**100%**	**100%**
1.1 NAME	100	100
1.2 ORIGIN	24.7	16.7
1.3 IMAGE	100	100
1.4 PACKAGE & PRICE	12.7	8
M2 – PROMOTIONAL DESCRIPTION	**100%**	**100%**
2.1 MARKETING STATEMENT	83.3	62.7
2.2 TASTING NOTE	81.3	55.3
2.3 (HEALTH) PROPERTIES	22	20.7
M3 – INGREDIENTS	**72%**	**74.7%**
3.1 NAME	71.3	73.3
3.2 CHARACTERISTICS	14	32
M4 – SUGGESTIONS	**56.7%**	**77.3%**
4.1 HOW TO MAKE	53.3	66.7
4.2 HOW TO TAKE	12	34
M5 – PROCESSING	**32%**	**18%**
5.1 THE PROCESS	12.7	8
5.2 THE BRAND	19.3	10
M6 – NUTRITION & ALLERGIES	**40%**	**6%**

While the order suggested in the macrostructure resembles the preferred pattern in our corpus, end users can alter it in the output file if the need arises. In fact, the tool is flexible with regard to the writing of the different chunks. Thus we can write texts in a sequential way, starting with M1 and its steps, then M2 and its corresponding steps, and so on until M6; however, we might start with M3, for example, then continue with any other chunk. Whatever the order, the tool does not require the user to fill in a given section entirely to continue writing. The text generator is, rather, an aid that guides professionals in their writing, and it is ultimately they who decide how to do this.

4.3.2 *Model lines*

The usefulness of rhetorical annotation lies in the possibility of further browsing the corpus per functional text segment, i.e., move and step. Using the ACTRES Browser, a purpose-built tool, we first gathered all chunks tagged with the same rhetorical label in every language. We observed phraseological and lexicogrammatical patterns that, even though formally diverging, were functionally equivalent. On the basis of this perceived similarity, we examined all instances of each move and step in English to compile a pool of style guidelines that could subsequently be integrated into the tool. This examination was manually handled paying attention to the surrounding co-text of keywords whose content matched the communicative function of the chunk it was found in. For example, for the step of tasting note, words like 'colour', 'taste', 'ingredient', or 'variety', amongst others, helped us pin down a variety of recurrent patterns. To avoid duplication, we conceived of our model lines as extended p-frames to fill in with content-specific words, which urged the need for a specialised glossary (cf. 4.3.3). The tool is meant to help create a **prototypical** HTPT, therefore we grouped together all sentences according to one criterion: although differing in their lexical content (paradigmatic choices), they seemed to have the same or similar syntagmatic skeleton, our drafting/model line. Similarity was controlled for thanks to three different types of drafting lines in terms of obligatoriness and/or optionality as will be explained later. We discarded any model of which there was only one occurrence in our corpus. The number of writing suggestions per step differed, ranging from three to eleven model lines. For example, for *Step 2.2. Ficha de cata* (Tasting note) we identified nine different model lines, which are shown in (1) to (9):

(1) The {(VARIEDAD DE TÉ) / (INGREDIENTE)} is known for its (SABOR)^ flavour with a [{strong/bold/mild}] taste of (INGREDIENTE)^ and (SABOR) undertones.
e.g. The green leaf Tulsi is known for its fresh, mellow flavor with a strong taste of cloves and earthy undertones.

(2) This [caffeine-free] {(VARIEDAD DE TÉ) / (INGREDIENTE)} is [(VALORACIÓN POSITIVA)] (SABOR) and (ACCIÓN) (COLOR) in color and the liquor is (SABOR), with (MEDIDA) of (SABOR) (INGREDIENTE) flavors, and a [{strongly/decidedly/mildly/lightly/slightly}] (SABOR) note.
e.g. This caffeine-free herbal is wonderfully aromatic and steeps ruby red in color and the liquor is full-bodied, with tons of sweet fruit/berry flavors, and a decidedly tart note.

(3) The (INGREDIENTE)^ (ACCIÓN) to {make/produce} a [{strongly/decidedly/mildly/lightly/slightly}] [(AROMA)^] and [{strongly/decidedly/mildly/lightly/slightly}] [(SABOR)^] cup.
e.g. The lemongrass, ginger, lemon peel, licorice, and sprinkling of peppermint blend to make a fragrant, zesty, and lightly spiced cup.

(4) It has a (SABOR)^ {flavour/taste/scent} [and {(SABOR)/ (INGREDIENTE)} undertones / {to give/with} a (TEXTURA) (SABOR) finish / with (SABOR)^ notes/ that lingers on the tongue].
e.g. It has a bright, fresh flavor, to give a soothing, sweet finish with warming and coooling notes. / It has a lemony scent, and ginger undertones.

(5) [It] (ACCCIÓN) a (VALORACIÓN POSITIVA) (COLOR)^ [colour] with a [(PROPIEDAD)] {(INGREDIENTE)/ (SABOR)} flavor.
e.g. Brews a pleasant yellow-green with a cooling menthol flavor.

(6) A [{strongly/decidedly/mildly/lightly/slightly}] (SABOR)^ [herbal] {tea/blend/infusion/tisane} with the [(VALORACIÓN POSITIVA)] [and] (SABOR)^ taste of (INGREDIENTE).
e.g. A full-flavoured, deeply fruity infusion with the distinctive and delicious aniseed-like taste of star anise.

(7) (ACCIÓN) something (SABOR) and (TEXTURA) with this [caffeine-free] [herbal] {tea/blend/infusion/tisane} of (SABOR)^ taste and the (AROMA) aroma of (INGREDIENTE).
e.g. Sip something sweet and soothing with this caffeine-free blend of juicy, tropical pineapple taste and the fragrant aroma of chamomile flowers.

(8) It is (SABOR), [{strongly/decidedly/mildly/lightly/slightly}] (SABOR) and very (TEXTURA), almost (TEXTURA).
e.g. It's sweet, slightly spicy and very smooth, almost silky.

(9) (SABOR)^ [flavours], {with plenty of (TEXTURA)^ and (SABOR) finishing flavours of (INGREDIENTE) / hinting at (INGREDIENTE)^ with a (SABOR) finish}.
e.g. Statuesque and broad, with plenty of grip and structure and rich finishing flavours of malt. / Statuesque and broad flavours, hinting at celery and fennel with a sappy finish.

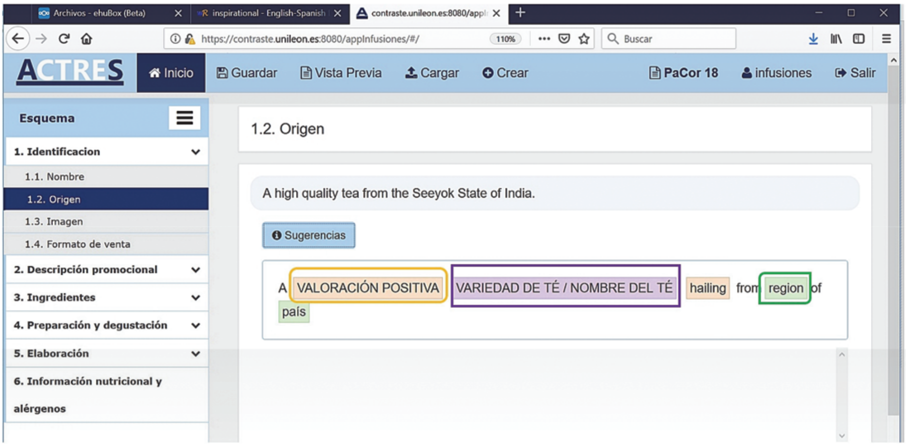
In total, the tool provides the user with 64 model lines, whose prototypicality would guarantee acceptability in English as a medium of communication.

In the tool, such model lines function as controlled language choices for the Spanish-speaking user to consider during their writing. In other words, they behave as drafting lines that can be adapted to different communication needs by filling in a few words to specify different elements of content like flavour, texture, ingredients, amongst others, in a string of pre-written language (cf. 4.3.3). The information needed to complete the drafting lines is written in Spanish to allow full understanding on the part of the users, and thus to make their writing process easier and more natural.

As illustrated in Figure 2, we distinguish three types of drafting lines, differentiated in the tool with colours.[8] While they all are open to management by the user, the suggestions shaded in green are obligatory. This means that they must always be filled in, this because of the demands of idiomaticity, grammatical accuracy

[8] Colours correspond to the different types of brackets in Examples (1)–(9). A colour coding was deemed more user-friendly in the interface than parenthetical distinctions.

and genre acceptability. If the drafting line is shaded in orange, that information is optional and the user is free to decide whether he/she wants to edit it or simply delete it. Finally, purple shading indicates that the user is provided with two or more choices to select from. In this case the line is necessary, and its content delimited to the options given.



1.2. Origen

A high quality tea from the Seeyok State of India.

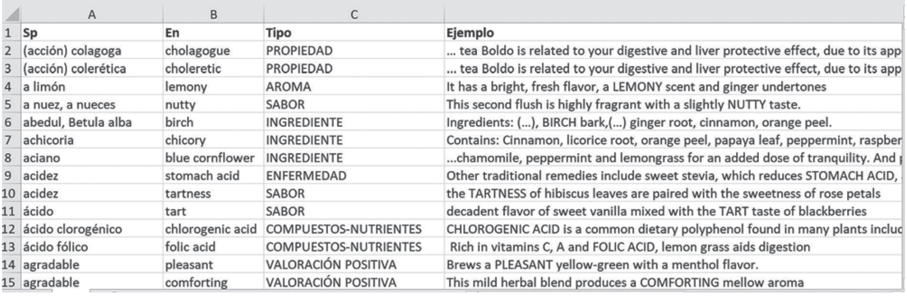
A VALORACIÓN POSITIVA VARIEDAD DE TÉ / NOMBRE DEL TÉ hailing from region of país

Figure 2. Types of drafting lines

An exception to the procedure described above is found in *Step 2.1. Estrategia promocional*. As the primarily informational-persuasive chunk of the text type, this text segment is so eclectic that a prototypical wording was difficult to pin down. There might be various reasons for this: text length, which seems to be highly brand-dependent; information load, with purely descriptive as well as persuasive resources intertwined; creativity in discourse, with very many metaphors used in both the description of the teas and that of the effect that tea drinking might have on the customer. Alternatively, the user is informed of six prototypical promotional strategies that have been derived from a corpus-based analysis of register features at the move/step level (Izquierdo & Pérez Blanco 2020). Delimiting drafting lines per strategy was fraught with difficulty, so the user is instead informed of the type of content conveyed by the strategies, as expected in this functional chunk. The strategy could be to sell tea drinking as an enjoyable experience; to praise the benefits of the blend; to appeal to the exclusiveness and to the aesthetics of the product and/or to show that tea drinking is a popular habit, not only in society generally but with well-known people in particular. In addition to helping the user decide what kind of effect to trigger in the customer through a given strategy, these recommendations include some language tips for the user to consider, such as the use of imperatives, the personal pronoun 'you', positive evaluative resources and intensifiers, amongst others.

4.3.3 *Built-in glossary*

To further assist the user in completing the drafting lines, the text generator is equipped with a built-in glossary. Using AntConc 3.5.2 (Anthony 2018) we extracted word lists from each language sample in the corpus. Then, we manually identified food-related lexis. To a certain extent, the retrieval of these domain-specific items was based on intuition. Therefore, we double checked their presence and function in the various moves and steps previously identified. In this way, we gathered almost 600 entries, which provide corpus-based functional equivalent terms in EN and ES, as shown in Figure 3.



	A	B	C	
1	**Sp**	**En**	**Tipo**	**Ejemplo**
2	(acción) colagoga	cholagogue	PROPIEDAD	… tea Boldo is related to your digestive and liver protective effect, due to its app
3	(acción) colerética	choleretic	PROPIEDAD	… tea Boldo is related to your digestive and liver protective effect, due to its app
4	a limón	lemony	AROMA	It has a bright, fresh flavor, a LEMONY scent and ginger undertones
5	a nuez, a nueces	nutty	SABOR	This second flush is highly fragrant with a slightly NUTTY taste.
6	abedul, Betula alba	birch	INGREDIENTE	Ingredients: (…), BIRCH bark,(…) ginger root, cinnamon, orange peel.
7	achicoria	chicory	INGREDIENTE	Contains: Cinnamon, licorice root, orange peel, papaya leaf, peppermint, raspber
8	aciano	blue cornflower	INGREDIENTE	…chamomile, peppermint and lemongrass for an added dose of tranquility. And
9	acidez	stomach acid	ENFERMEDAD	Other traditional remedies include sweet stevia, which reduces STOMACH ACID,
10	acidez	tartness	SABOR	the TARTNESS of hibiscus leaves are paired with the sweetness of rose petals
11	ácido	tart	SABOR	decadent flavor of sweet vanilla mixed with the TART taste of blackberries
12	ácido clorogénico	chlorogenic acid	COMPUESTOS-NUTRIENTES	CHLOROGENIC ACID is a common dietary polyphenol found in many plants incluc
13	ácido fólico	folic acid	COMPUESTOS-NUTRIENTES	Rich in vitamins C, A and FOLIC ACID, lemon grass aids digestion
14	agradable	pleasant	VALORACIÓN POSITIVA	Brews a PLEASANT yellow-green with a menthol flavor.
15	agradable	comforting	VALORACIÓN POSITIVA	This mild herbal blend produces a COMFORTING mellow aroma

Figure 3. Built-in glossary entries

The third column in the spreadsheet in Figure 3 shows a variety of semantic categories that we established as key, register-specific language choices in the drafting lines. Besides the specific terminology used to describe the attributes of a herbal tea such as 'taste', 'colour', 'aroma', or 'texture', and also the properties of the herbal tea, the glossary contains other semantic categories: 'ingredients'; 'illnesses' or 'ailments' that we commonly treat by drinking herbal teas; actions involved in tea drinking, among others. Such categories, in fact, enable the drafting of the model lines as previously shown and help the user to recognise the type of information required. The glossary appears whenever we have to fill in a green-shaded suggestion in our drafting line. Because the tool is addressed to users with a B2 level of English, we would not find the kind of simpler vocabulary that is expected to be known at this level, the verb 'relax', for example.

4.4 Writing assistance

Promociona-TÉ is dynamic software with a user-friendly interface that guides the writer through the HTPT writing process. When we log onto the tool, we have two options: writing a text from scratch, or returning to an existing text.

When we click on a given step we are presented with a text box, which is to be filled in using a number of suggestions that the user gets by clicking on the button *sugerencias* ('suggestions'). Let us take M1-Identification, *Step 1.1. Name* to demonstrate this process. The user is provided with a number of typical patterns to choose from, such as the model lines A to C in (10) here:

(10) A. Property (of the infusion/ingredient) + Ingredient
B. (Main) Ingredient + herbal tea
C. Property (of the infusion) + tea

First, the user should click on *añadir* ('add') to select one of the suggestions to be completed. On doing this, a real example taken from the corpus pops up so that s/he can see what the text would look like. For example, for drafting line A. Property + Ingredient, we get example 'Restoring Echinacea and Raspberry', as shown in Figure 4.

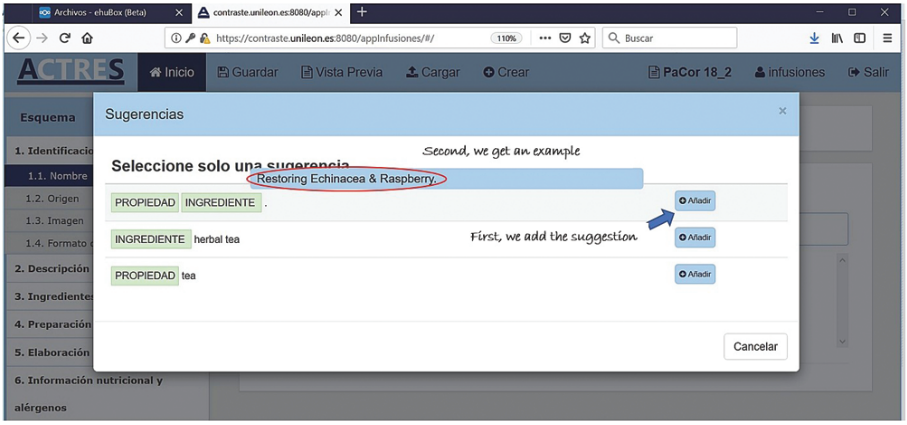
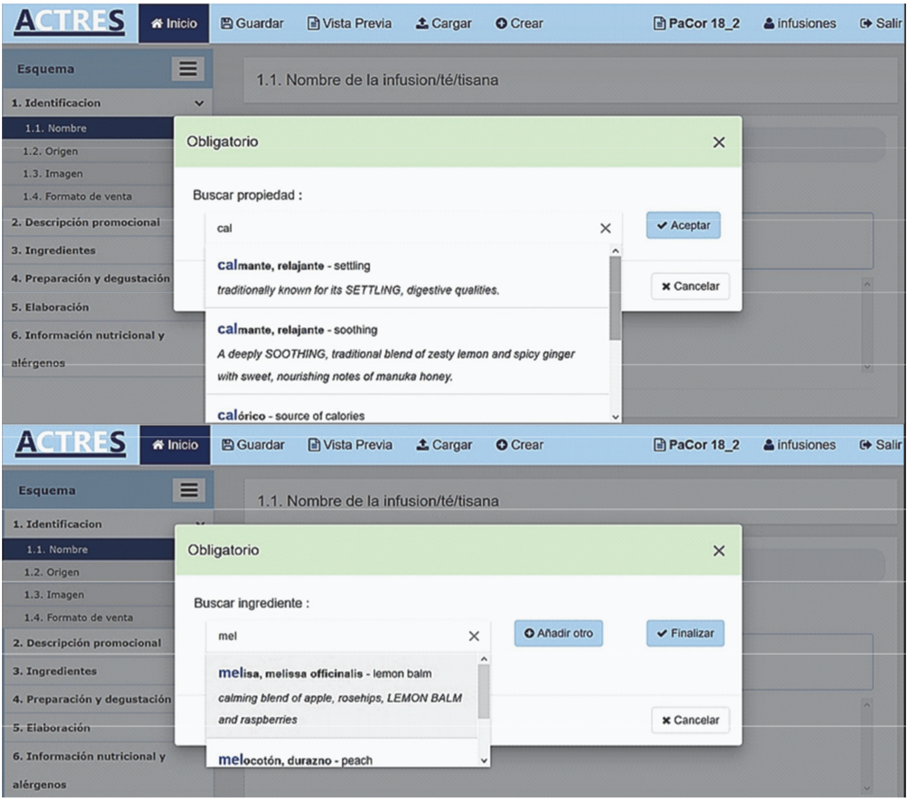


Figure 4. Selection of model line with ensuing pop-up example

Once we add a given drafting line to our text box we are in a position to write specific content for our HTPT. It is at this moment that the writing assistance begins, thanks to the colour coding, which indicates to the user what is obligatory and what is not. Most importantly, the built-in glossary is activated every time we have to fill in a green-shaded model line. To visualise the way in which the glossary assists the user, consider the suggestion 'Property + Ingredient' above and imagine that the infusion is meant to ease nerve pain. All we have to do is type in the Spanish word *calmante* ('calming'/ 'soothing') to get the English equivalent. The tool detects the first letters of the word and retrieves all possible matches from the built-in glossary.

We then simply select the most appropriate one. Sometimes there is more than one functional equivalent in English. Therefore, every entry in the glossary offers an example of the term in use so that the user can judge which is the best choice. When we click on our preferred equivalent, the text appears in the box and we can click on *aceptar* ('accept') to keep it. If we need to mention other properties of the tea, or duplicate one of the categories marked in colour, the tool lets us add as many others as necessary. Figure 5 shows these steps.



cal

calmante, relajante - settling
traditionally known for its SETTLING, digestive qualities.

calmante, relajante - soothing
A deeply SOOTHING, traditional blend of zesty lemon and spicy ginger with sweet, nourishing notes of manuka honey.

calórico - source of calories

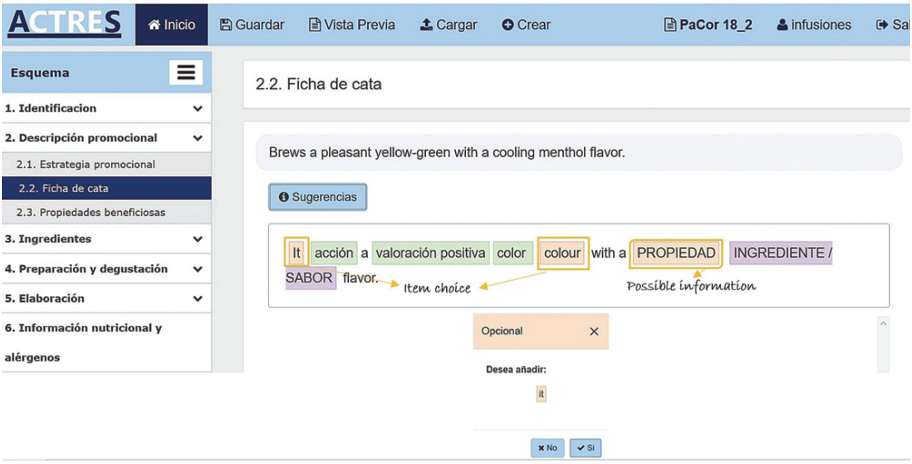
mel

melisa, melissa officinalis - lemon balm
calming blend of apple, rosehips, LEMON BALM and raspberries

melocotón, durazno - peach

Figure 5. Glossary activation

The writing of orange-shaded, optional drafting lines is slightly different. First, categories shaded in orange can be indicated either in lowercase, as illustrated by 'It/colour', or in capitals, as in 'PROPERTIES' in Figure 6.



2.2. Ficha de cata

Brews a pleasant yellow-green with a cooling menthol flavor.

It acción a valoración positiva color colour with a PROPIEDAD INGREDIENTE / SABOR flavor.

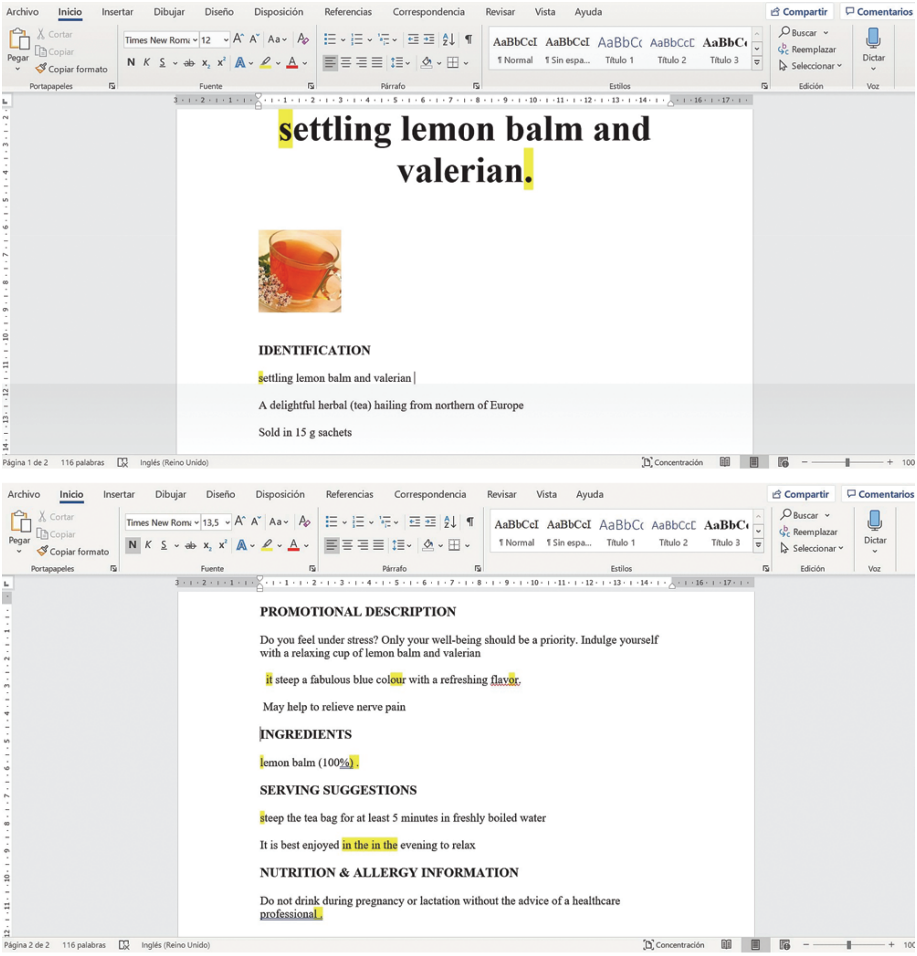
Figure 6. Writing up of an optional drafting line

Lowercase font type indicates that the user has an option: s/he can insert the given word or not. On the other hand, if a drafting chunk shaded in orange features capitals, the user should introduce the type of information specified. In the case of Figure 6, a property of the herbal tea is required in the model line. If we want to include it, we must click within the line, which will turn into a green (obligatory) drafting chunk that is linked to the glossary.

On the other hand, purple shading provides the user with a selection of information or items that are necessary for the chunk to work idiomatically. The editing of this kind of line is the same as in the case of optional information in orange. On the one hand, we might select a given item from an existing list visually marked in lowercase; on the other, we might opt to fill in the line with an expected semantic category or type of information, indicated in capitals, that, on being selected, becomes a green (obligatory) line, thus triggering the glossary when we start writing.

Once we have finished the process of filling in all the moves and steps that we want, we can click *vista previa* ('preview') to see the output text. As shown in Figure 7, the text may need formatting with regard to font size, misplaced/unnecessary punctuation or even certain grammar issues that as yet the machine is not able to resolve. We can now download a .doc version of the file and modify as necessary.

Once corrected, the Word document is saved as an .mht file that opens by default in Internet Explorer or as a Word document. In the latter case, the text will be tidy and well ordered. As indicated, this tool assists the user in the writing of the specialised, textual part of an HTPT. Therefore, multimodal elements typical of this online promotional sub-genre are not dealt with, except for the image in M1. The user may download the output text and manipulate it as necessary to insert and format any other feature of their web interface. All the write-up stages described



settling lemon balm and valerian.

IDENTIFICATION

settling lemon balm and valerian

A delightful herbal (tea) hailing from northern of Europe

Sold in 15 g sachets

PROMOTIONAL DESCRIPTION

Do you feel under stress? Only your well-being should be a priority. Indulge yourself with a relaxing cup of lemon balm and valerian

it steep a fabulous blue colour with a refreshing flavor.

May help to relieve nerve pain

INGREDIENTS

lemon balm (100%) .

SERVING SUGGESTIONS

steep the tea bag for at least 5 minutes in freshly boiled water

It is best enjoyed in the in the evening to relax

NUTRITION & ALLERGY INFORMATION

Do not drink during pregnancy or lactation without the advice of a healthcare professional .

Figure 7. Text output downloaded as a word document for correction

above are shown in a demo video of how the generator works, which can be found on the ACTRES website.[9]

4.4.1 *Feedback from users*

As indicated in 4.1., Promociona-TÉ was developed as part of an agreement between ACTRES and Pharmadus Botanicals S.L. In personal communication with the department in charge of testing the tool, we learned that "the tool is quick to use", "the information contained in the tool mirrors reality", "the vocabulary is rich and fits the tea description aptly", and "the model lines are very useful. In fact, they

[9] https://actres.unileon.es/demos/generadores/applications.html#generatorsSection

do reproduce some of the expressions that we frequently use in our tea descriptions". Their evaluation also noted a further advantage of the tool, in that it may also be used for assessing the quality of texts previously written and stored. This was not the aim of the tool as developed, yet it does underline its broad usefulness and usability in the context of use. On the other hand, a specific criticism of the tool was that the label for *Step 2.1. Estrategia promocional* was found to be inaccurate.[10]

4.5 Tool programming and architecture

Promociona-TÉ is a good example of interdisciplinary research. The descriptive work at the language level has been transferred to the computational level thanks to the collaborative efforts of linguists and engineers. The former provides the latter with reliable data for content-rich applications that meet a real need of communication in a specific professional realm.

A variety of programming languages were used in the implementation of the software. HTML5 is a mark-up language used in the design of the graphical user interface. Accordingly, three cascading style sheets (CSS3) were used (a) to achieve a modern-looking interface that adapts to screen size, using the Bootstrap toolkit, and (b) to personalise general views and the sidebar menu with *main.css* and *sidebar-menu.css* respectively.

AngularJS (1.5.0) is a JavaScript-based front-end web framework for client-side model-view control. This program is useful to develop single-page applications. MongoDB (2.6.11) is a NoSQL document-oriented database that stores JSON-like documents.[11] Finally, Node.js (4.2.6) is a server-side JavaScript environment that has been used to query the MongoDB database program, and to create API REST, which is a protocol for data exchange on the Internet. We have used it to retrieve query hits in a JSON format. It also enables image storage. Figure 8 shows the architecture of the application.

The writing aid has been developed according to a workflow divided into several stages, as shown in Figure 9. Once logged onto the home page, the end-user starts writing a file. They will get suggestions for drafting that they will have to complete, either with the help of a glossary or by themselves. As the user progresses, the writing is saved, which enables a preview of the output at any time, whether or not the text is fully edited. The file saved may be closed and reopened later, hence the circularity illustrated in the workflow diagram.

[10] Feedback provided by L. Martínez, from the Marketing Department of Pharmadus Botanicals S.L., in November 2018. With regard to the texts Pharmadus S.L. have generated with the tool, we do not have access to this material, in compliance with the agreement signed with the company.

[11] JSON (JavaScript Object Notation) is a language-independent file format.

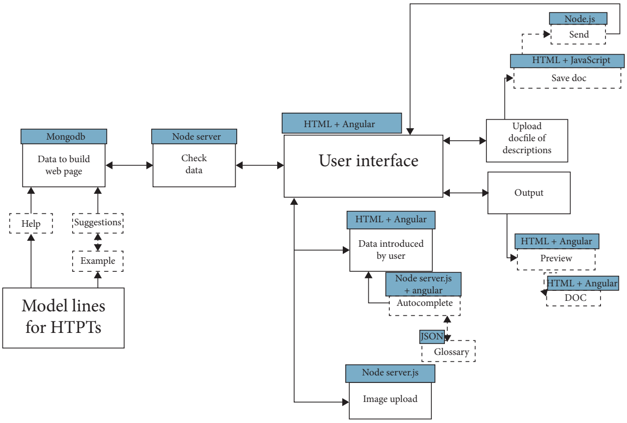


Figure 8. Application architecture

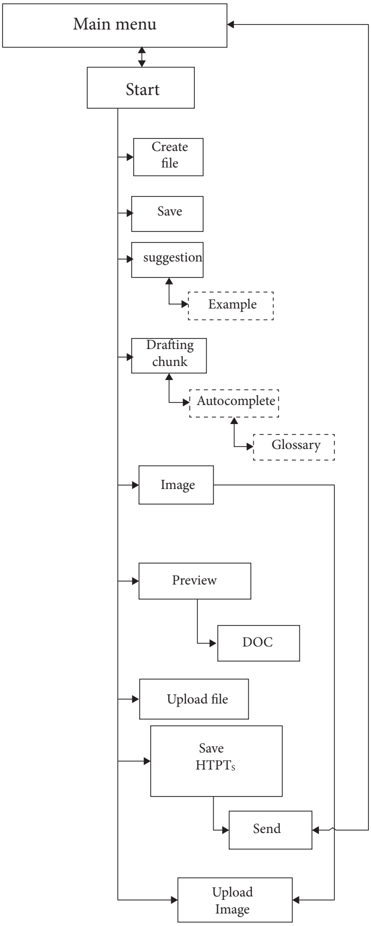


Figure 9. Software workflow

4.6 Advantages of the tool

Promociona-TÉ is conceived of as a writing tool that meets the communicative needs of a professional community, as an alternative to unfeasible or unaffordable translation. In fact, if compared with human or machine translation (MT), this tool exhibits a number of advantages. First, as an alternative to human translation, this DIY tool allows potential users, such as small local businesses, to dispense with the costs of buying in translation services. On the other hand, it fills in the gap created by the limitations of using MT, which often results in poor quality translations when dealing with certain text types or narrow (under-resourced) domains (Skadiņa et al. 2010). In this regard, being corpus-informed, the model text lines that the tool provides are accurate, as well as representative, which makes them a better option than costly post-editing of MT. The comparable approach we have followed has offered unique insights into a very specific, under-researched knowledge domain, and this compensates for the translator's or a machine's need to have specialised knowledge or subject-field understanding (Bowker 1998), guaranteeing effective and accurate output texts in the target language (i.e. English).

A further benefit of the application is the glossary of functional equivalents built using an "in-domain corpus", in that MT engineers regard the storage of domain key terms to be an essential requirement towards ensuring domain adaptation (Skadiņs 2011). Finally, despite advances in the field of Neural Machine Translation (NMT) to overcome many of the weaknesses of traditional MT, such translation has thus far not proved robust enough, especially with rare words (Wu et al. 2016). This is clearly a particularly serious shortcoming when we are dealing with specialised translation, such as the text type under study in this paper. Also, NMT systems are computationally very expensive, perhaps even prohibitively so when a high volume of data is required (ibid). It goes without saying that such technology is beyond the reach of the kind of small businesses that our text generator is designed to help.

In addition, from a strictly linguistic point of view, the current research paves the way for the development of a Controlled Natural Language (CNL) for a specific/restricted domain, this with the aim of pursuing further linguistic and professional alliances. The contrastive work carried out has provided useful rhetorical, semantic and lexicogrammatical information that could form the basis of the next stages in the construction of a CNL. The linguistic knowledge gained through the present research, including its underlying components, is a step prior to the necessary parametrization of text features at the levels of text structure, syntax, semantics and pragmatics. Thus, for example, the identification and extraction of the key terms (i.e. glossary) found in our texts is a first step towards the unsupervised semantic tagging of our corpus (Sanjurjo-González et al. 2019). On the other hand, the move-analysis conducted, which is based on the underlying pragmatic function(s) of the text segments (i.e. moves and steps), makes it possible to distinguish a text's illocutionary forces (persuade, inform, instruct), while the register analysis has provided a pool of lexico-grammatical resources.

5. Conclusions

The aim of this chapter has been to explain the development and functioning of an English specialised writing tool for Spanish professionals. In particular, the tool is customised to meet the needs of a very specific group of users in the industrial setting of (herbal) tea manufacture and promotion. Our scholarly interest in HTPTs, very much an under-researched text type, has been motivated by a university-industry alliance between the ACTRES research group and a number of local food businesses (tea, wine, cheese manufacturers) towards meeting their needs in writing online food descriptions. The tool described here, then, is the ultimate goal of an application-oriented research project.

As noted in the introduction, the absence of parallel texts in the herbal tea domain led to the compilation of an English-Spanish comparable corpus, on the basis of which the writing tool presented here was developed. This corpus-informed writing tool is a valuable result of linguistic research and has an immediate practical application, in that it satisfies a real communication need in a professional realm. Indeed, Promociona-TÉ is a clear example of what is known as knowledge transfer, going as it does from linguistic description to professional application, something which is increasingly demanded in our technology-driven society. Such transfer, indeed, foregrounds the usefulness and usability of the tool. It is useful because it contains specific vocabulary, style guidelines and common structures required in specialised documents, which are unlikely to be taught on a course in ESP. Because traditional teaching, even ESP, is unlikely (or unable) to cover such specialised areas of knowledge, professionals working in related fields are likely to lack the ability to produce this type of writing. A tool for the workplace, then, is exactly what is required to fill this gap. More importantly, being corpus-informed, this text generator guarantees the acceptability and correctness of the output text, which in turn preserves effective communication and, in the long run, successful business practice. In short, usability is a defining criterion, since this time-saving writing tool is meant to ease actual practice in the workplace.

As a writing aid, the tool is limited in its commercial applications, in that it is envisaged that potential users will use it to create online textual descriptions appropriate for use on the web. An upgraded version of the tool, possibly extending its scope, would require another project to examine the multimodality of online food descriptions. This indeed is an interesting niche for future research. Another current limitation of the tool is the prototypicality of *Step 2.1. Estrategia promocional*, which requires improved drafting.

Acknowledgements

We would like to thank the anonymous reviewers that made comments on a previous draft of this paper. Their suggestions and advice have helped us to improve it considerably. Any errors remain our own.

Funding

Research for this paper has been funded through Project FFI2016-75672-R of the Spanish Ministry for Economy, Industry and Competitiveness.

References

Aijmer, Karin, and Bengt Altenberg. 1996. "Introduction". In *Languages in Contrast. Papers from a symposium on text-based cross-linguistic studies, Lund 4–5 March 1994* [Lund Studies in English] ed. by Karin Aijmer, Bengt Altenberg and Mats Johansson, 11–16. Lund: Lund University Press.

Anthony, Laurence. 2016. "Tailoring Corpus Tools for Academia and the Language Industry: A Developer's Perspective". *Plenary Conference given at VIII CILC, University of Málaga*, Spain, 2–4 March 2016.

Anthony, Laurence. 2018. *AntConc 3.5.2* [Computer Software] Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software.html

Aston, Guy. 1999. "Corpus Use and Learning to Translate". *Textus* 12: 289–314. Available online at: https://www.sslmit.unibo.it/~guy/textus.htm (accessed on 18 July 2019).

Atkinson, Dwight. 2004. "Contrasting Rhetorics/Contrasting Cultures: Why Contrastive Rhetoric Needs a Better Conceptualization of Culture" [Special issue on Contrastive Rhetoric]. *Journal of English for Academic Purposes* 3 (4): 277–289. https://doi.org/10.1016/j.jeap.2004.07.002

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies". In *Text and Technology: In Honour of John Sinclair*, ed by Gill Francis and Elena Tognini-Bonelli, 233–252. Amsterdam: John Benjamins. https://doi.org/10.1075/z.64.15bak

Bhatia, Vijay K. 2004. *Worlds of Written Discourse*. London/New York: Continuum International Publishing.

Bhatia, Vijay K. 2008. "Genre Analysis, ESP and Professional Practice". *English for Specific Purposes* 27:161–174. https://doi.org/10.1016/j.esp.2007.07.005

Biber, Douglas, and Susan, Conrad. 2009. *Register, Genre, and Style*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511814358

Biber, Douglas and Meixiu Zhang. 2018. “Expressing Evaluation without Grammatical Stance: Informational Persuasion on the Web”. *Corpora* 13 (1): 97–123. https://doi.org/10.3366/cor.2018.0137

Borin, Lars. 2002. “…and never the twain shall meet?”. In *Parallel Corpora, Parallel Worlds: Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999*, ed. by Lars Borin, 1–43. Amsterdam: Rodopi. https://doi.org/10.1163/9789004334298_002

Bowker, Lynn. 1998. “Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study”. *Meta* 43 (4): 631–651. https://doi.org/10.7202/002134ar

Colombo, Laura M. 2012. “English Language Teaching Education and Contrastive Rhetoric”. *Elted* 15: 1–6.

Doval, Irene and Maria T. Sánchez Nieto (eds). 2019. *Parallel Corpora for Translation Studies: New Resources and Applications*. Studies in Corpus Linguistics. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.90

Ethnologue 2018. *Ethnologue, 21st Edition*. Last accessed on 16 April 2020 at https://lemongrad.com/english-language-statistics/

Gale, William A. and Kenneth Church. 1993. A Programme for Aligning Sentences in Bilingual Corpora. *Computational LInguistics* 19 (1): 75–102.

Hareide, Lidun and Knut Hofland. 2012. “Compiling a Norwegian-Spanish Parallel Corpus: methods and challenges”. In *Quantitative Methods in Corpus Based Translation Studies*, ed. by Michael P. Oakes and Meng Ji, 75–114. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.51.04har

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Hyland, Ken. 2000. *Disciplinary Discourses: Social Interactions in Academic Writing*. Harlow, UK: Pearson.

Izquierdo, Marlén and María Pérez Blanco. 2020. A Multi-level Contrastive Analysis of Promotional Strategies in Specialised Discourse. *English for Specific Purposes* 58: 43–57. https://doi.org/10.1016/j.esp.2019.12.00

Izquierdo, Marlén, Knut Hofland, and Oystein Reigem. 2008. “The ACTRES Parallel Corpus: an English-Spanish Translation Corpus”. *Corpora* 3: 31–41. https://doi.org/10.3366/E1749503208000051

James, Carl. 1980. *Contrastive Analysis*. London: Longman.

Jenkins, Jennifer. 2009. “English as a Lingua Franca: Interpretations and Attitudes. *World Englishes* 28 (2): 200–207. https://doi.org/10.1111/j.1467-971X.2009.01582.x

Johansson, Stig. 2007. *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.26

Johansson, Stig and Knut Hofland. 1998. “The Translation Corpus Aligner: A Program for Automatic Alignment of Parallel Corpus”. In *Corpora and Cross-linguistic Research: Theory, Method and Case Studies*, ed. by Stig Johansson and Signe Oksefjell-Ebeling, 87–100. Amsterdam: Rodopi.

Lam, Phoenix. 2013. “Interdiscursivity, Hypertextuality, Multimodality: A Corpus- Based Multimodal Move Analysis of Internet Group Buying Deals”. *Journal of Pragmatics* 51: 13–39. https://doi.org/10.1016/j.pragma.2013.02.006

López Arroyo, Belén and Roda P. Roberts. 2014. “English and Spanish Descriptors in Wine Tasting Terminology”. *Terminology* 20 (1): 25–49. https://doi.org/10.1075/term.20.1.02lop

Malmkjær, Kirsten. 1998. "Love thy neighbour: Will Parallel Corpora endear Linguists to Translators?" *Meta* 43 (4): 534–541. https://doi.org/10.7202/003545ar

Marco, Josep. 2019. "Living with Parallel Corpora. The Potentials and Limitations of their Use in Translation Research". In *Parallel Corpora for Translation Studies: New Resources and Applications*, ed. by Irene Doval and María T. Sánchez Nieto, 39–56. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.90.03mar

McEnery, Tony and Richard Xiao. 2008. "Parallel and Comparable Corpora: What is Happening?" In *Incorporating Corpora: The Linguist and the Translator*, ed. by Gunilla M. Anderman and Margaret Rogers, 18–31. Clevedon: Multilingual Matters.

Mitkov, Ruslan. 2018. "The Name of the Game is Comparable Corpora". *Plenary Conference given at International Symposium PaCor 2018, Universidad Complutense de Madrid*, Spain, 5–7 November 2018.

Noël, Dirk. 2003. "Translations as Evidence for Semantics: An Illustration". *Linguistics* 41: 757–785. https://doi.org/10.1515/ling.2003.024

Pérez Blanco, María. 2018. "The Discourse Functions of Certainly and its Spanish Counterparts in Journalistic Opinion Discourse". *RESLA* 31 (2): 520–549. https://doi.org/10.1075/resla.16026.per

Rabadán, Rosa. 2019. "Working with parallel corpora: Usefulness and usability". In *Parallel Corpora for Contrastive and Translation Studies. New resources and applications*, ed. by Irene Doval and Maria T. Sánchez Nieto, 57–77. Amsterdam/Philadelphia: John Benjamins.

Rabadán, Rosa and Marlén Izquierdo, M. 2012. "Designing Writing Materials for the Business English Language Class". In *Intercultural Inspirations for Language Education. Spaces for Understanding*, ed. by Ilona Semrádová, 46–55. Hrádec Kralové: Hrádec Kralové University.

Rabadán, Rosa, Héctor Alaiz-Moretón, Ramón-Ángel Fernández, Ana García-Gallego, Camino Gutiérrez-Lanza, Belén Labrador, Noelia Ramón, and Hugo Sanjurjo-González. 2014. *Procedimiento de evaluación de la calidad gramatical de las traducciones al español de textos en lengua inglesa (PETRA 1.0)*. Available online at http://actres.unileon.es/?page_id=50&lang=en

Sanjurjo-González, Hugo and Marlén Izquierdo. 2019. "P-ACTRES 2.0: A Parallel Corpus for Cross-linguistic Research". In *Parallel Corpora for Contrastive and Translation Studies. New resources and applications*, ed. by Irene Doval and Maria T. Sánchez Nieto, 215–232. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.90.13san

Sanjurjo-González, Hugo, Rosa Rabadán, and Camino Gutiérrez-Lanza. 2019. "Creating a Dataset for Domain Bilingual Semantic Annotation based on the USAS Framework". *Paper presented at X CILC Conference*, Valencia (Spain), May 2019.

Santos, Diana and Signe Oksefjell-Ebelling. 2000. "An evaluation of the Translation Corpus Aligner, with Special Reference to the Language Pair English-Portuguese". In *NODALIDA'99, Proceedings from the 12th "Nordisk datalingvistikkdager, Trondheim, 9–10 December 1999*, ed. by Torbjørn Nordgård, 191–205. Trondheim: Department of Linguistics, NTNU.

Sanz-Villar, Zuriñe. 2019. "An Overview of Basque Corpora and the Extraction of certain Multi-word Expressions from a Translational Corpus". In *Parallel Corpora for Contrastive and Translation Studies. New resources and applications*, ed. by Irene Doval and Maria T. Sánchez Nieto, 233–247. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.90.14san

Sheerin, Patrick. 2008. "From Word to Essay: Common Mistakes in Students' Written Compositions" In *Estudios de metodología de la lengua inglesa (IV)*, ed. by Leonor Pérez Ruiz, Isabel Pizarro Sánchez & Elena González-Cascos Jiménez, 345–355. Valladolid: Centro Buendía.

Skadiņa, Inguna, Andrejs Vasiļjevs, Raivis Skadiņš, Robert Gaizauskas, Dan Tufiş, and Tatiana Gornostay. 2010. "Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation". In *Proceedings of 3rd Workshop on Building and Using Comparable Corpora. LREC 2010*, Valletta, Malta.

Skadiņš, Raivis. 2011. "Combined use of Rule-Based and Corpus-based Methods in Machine Translation". Unpublished PhD diss., University of Latvia.

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

Tiedemann, Jörg. 2011. *Bitext Alignment*. San Rafael CA: Morgan & Claypool Publishers. https://doi.org/10.2200/S00367ED1V01Y201106HLT014

Van Geyte, Els. 2013. *Writing: Learn to Write better Academic Essays*. London: HarperCollins.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc VLe, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprintarXiv:1609.08144*.

Zanettin, Federico. 2012. *Translation-driven Corpora*. London: Routledge.

PART II

Corpus-based studies and explorations

English and Spanish discourse markers in translation

Corpus analysis and annotation

Julia Lavid-López
Instituto Universitario de Lenguas Modernas y Traductores,
Universidad Complutense de Madrid

The study and annotation of discourse markers (DMs) in the context of translation is a much needed and challenging task not only for descriptive translation studies, but also for Natural Language Processing (NLP) applications. Their various meanings are difficult to identify and annotate, even for trained human experts. In this chapter, a methodology for the analysis and annotation of DMs is proposed, using three highly frequent DMs in English *-in fact, actually* and *really-* and their translations into Spanish as a case study. The methodology consists of an initial corpus analysis phase followed by a corpus annotation phase. The corpus analysis provides qualitative and quantitative information on the meanings of these DMs by looking at their translations in large parallel corpora. The corpus annotation phase specifies the annotation procedure, which can be generalized to other DMs and to other language pairs, and form the basis for large-scale cross-linguistic annotation of DMs.

Keywords: discourse markers, translation, English, Spanish, parallel corpus, corpus annotation

1. Introduction

As explained by Lavid (2019), the study of discourse markers (DMs) in the context of translation is crucial due to the idiomatic nature of these structures (Aijmer 2007, Beeching 2013), their frequent polysemy and their significant role in the coherence and readability of a text. However, their study has been usually carried out from a monolingual perspective or comparing English and other languages, and, to date, there are no systematic studies which address their cross-language behavior in

https://doi.org/10.1075/btl.158.07lav

the context of translation between English and Spanish.[1] However, as shown by previous works (Aijmer 2007, Lavid 2019, Lavid & Avilés 2019), the translation of DMs into one or more languages is an optimal empirical tool for investigating their meanings and uses, allowing to get a better picture of their correspondences and complementing the "circumstantial evidence provided by monolingual corpora" (Noël 2003: 156). In addition, annotated parallel datasets of DMs in different languages are scarce and very much needed in the Natural Language Processing (NLP) community for improving Machine Translation (MT) and Multilingual Text Generation (MTG) systems, which often fail to handle discourse markers properly.

The current study aims at contributing to current research efforts in this direction by describing a methodology which is applied to a specific case study -the English DMs *in fact, actually* and *really*, and their Spanish back translations-, but which can be generalized to other DMs and to other language pairs.

The specific research questions addressed in this study are the following: What does corpus analysis reveal about the meanings and uses of these DMs when looking at their translations in parallel corpora? Is it possible to identify core and peripheral meanings within the lexico-semantic field constructed for these DMs? How can the results of the corpus analysis be fruitfully used for annotation purposes?

The paper is organized as follows. Section 2 presents the data used in this study and the methodology proposed. Section 3 describes the corpus analysis phase, focusing first on the translations of *in fact, actually* and *really* into Spanish (3.1), and afterwards on the back translations into English (3.2). Section 4 presents an annotation model and procedure for the investigated DMs on the basis of the meanings identified in the corpus analysis phase. Finally, Section 5 summarises the work reported and provides some pointers for the future.

2. Data and methodology

Finding aligned translated data for immediate use in corpus analysis research without computational processing has not been usual so far. However, a number of aligned parallel datasets in different languages have become available to the research community in recent years, either through web-based platforms or by request to developers. This is the case of the SketchEngine platform, which has made available an important number of aligned parallel corpora in different languages (Kilgarriff et al. 2014).[2] One of the largest ones included in Sketch Engine is Opus2, a freely available collection of parallel corpora covering over 90 languages and including

[1] One exception is (Carretero 2012).

[2] The platform can be consulted at www.sketchengine.eu

data from several domains (Tiedemann 2009). A list of the subcopora included in Opus2 is provided in Appendix A.

Given the availability of these parallel corpora through the Sketch Engine interface, and their overall standard use in NLP contexts, the parallel English-Spanish and Spanish-English subsets of the OPUS2 corpus were selected. Both datasets are aligned and consist of 550.4 million and 548.9 million tokens, respectively.

The methodology proposed consists of two phases: a corpus analysis phase and a corpus annotation phase.

In the corpus analysis phase the translation correspondences of the three English DMs –*in fact, actually and really*- and their back translations are empirically investigated by looking at their frequency distribution in the English-Spanish and the Spanish-English parallel datasets of the Opus2 corpus. Their selection is based on their high frequency in the language, their frequent polysemy and their semantic relatedness. The aim is to identify their various meanings and pragmatic uses, and to construct a lexico-semantic field which can serve as a *tertium comparationis* for an in-depth cross-linguistic analysis.

In the corpus annotation phase, an annotation model for DMs is proposed, consisting of several steps: preparing a training suite, designing an annotation scheme and guidelines, annotating some fragment of the training corpus, and measuring the agreement between annotators.

The corpus analysis phase is described in Section 3, followed by the corpus annotation phase in Section 4.

3. Corpus analysis phase

The corpus analysis carried out in this study is inspired in Dyvik's method of 'semantic mirrors' (Dyvik 2004, 2005), which implies that the meaning of words becomes visible in translation and that translation mirrors the meaning of a word. In his own words: "The anatomy of meaning emerges in the translational tension between languages" (Dyvik 2005: 7). He proposes the following steps in the translation analysis: first, a set of translations of a word from the source language is identified in a parallel corpus. This set is called a t-image. Afterwards, the words from that t-image are translated back into the original language, and this second set is called an inverted t-image. Finally, by observing the distribution of words and overlapping translations of two images, it is possible to identify their different senses.

This method is particularly interesting in the case of discourse markers, which are usually polyfunctional and have multiple translations, as confirmed by some previous cross-linguistic studies (see Lavid 2019, Aijmer & Simon-Vandenbergen 2004, Bazzanella and Morra 2000, *inter alia*).

As it is often the case, a particular marker in one language is translated by a set of lexical items showing different frequencies. The more frequent translations are then considered to be more prototypical or core, while the less frequent are peripheral. Also, when we observe the second set of back translations, i.e., the inverted t-image, which also form a set of lexical items, we can compare the two images and get a description of their paired lexico-semantic fields.

In this study the following steps were taken during the corpus analysis phase:

1. First, some previous studies describing some of the meanings and functions of the English and the Spanish markers from a monolingual perspective were consulted. This was done to establish a baseline for comparison with other possible ones which could emerge from the translation analysis.
2. The Spanish translations of *in fact, actually* and *really* were identified in the parallel samples, their frequencies were counted and their meanings were identified (first mirroring).
3. Afterwards, the back translations of these DMS were identified in the parallel samples, their frequencies were counted and their meanings were also identified (second mirroring).
4. A lexico-semantic field or semantic map was constructed on the basis of the translation analysis carried out in (2) and (3).

The following subsections present the range of translations of the English DMs into Spanish (Section 3.1) and their back translations (Section 3.2). The translation analysis is preceded by a brief overview of some previous relevant work on some of the investigated DMs. Section 3.3. presents the lexico-semantic field which was constructed on the basis of the translation analysis.

3.1 Some previous work on English DMs

In this section only a brief summary of some previous monolingual studies on *in fact* and *actually* is presented, given the focus of this paper on translation data.

Previous monolingual studies on *in fact* and *actually* have shown that they are very close to one another, both semantically and pragmatically (see Smith and Jucker 2000, Oh 2000), although some scholars have also pointed out some differences. Thus, Oh (2000) has pointed out that *actually* has two separate uses in discourse: (i) a use with local scope as an 'emphasizer' which intensifies the truth value of the clause in which it occurs; (ii) a use with global scope frequently found in the context of contradiction and disagreement, whereas *in fact* tends to mark an increase in the strength of a previous assertion.

The DM *actually* has received extensive attention in the literature (see Oh 2000, Taglicht 2001, Clift 2001, Tognini-Bonelli 1993, and Aijmer 2013, 2015, *inter alia*). In her analysis of the formal and functional features of *actually* in different English varieties of spoken discourse, two core meanings emerge: one referring to contrast or revision, and another having to do with surprise and novelty (Aijmer 2015). Her analysis also shows that initial (left periphery) or final position (right periphery) in the utterance are associated with specific functions:: in the right periphery it is used for correcting a preceding claim and it can also imply surprise and novelty. In the left periphery it is used to introduce/emphasize a speaker's perspective on what is talked about, being potentially argumentative or confrontational.

Schwenter and Traugott (2000) have remarked that *in fact* is a highly polysemic item, and identified three main meanings: (1) *in fact$_1$*, and adverbial at the VP level meaning of 'in practice, as far as can be told from evidence, in actuality'; (2) *in fact$_2$*, an adversative adverb with primarily epistemic modal meaning. In this use it often collocates with *but* and combines the meaning of epistemic adverbs like *certainly* and adversative adverbs like *however;* (3) *in fact$_3$*, an adverb that signals that what follows is a stronger argument than what precedes, with respect to the speaker's rhetorical purpose at that point in the discourse. This *in fact* often collocates with *and* and it is to be situated in the same semantic field as *what's more* and *indeed*.

With respect to these meanings, Aijmer and Simon-Vandenbergen highlight that, from a synchronic point of view, they are not different polysemies but "pragmatic implicatures which are conventionalised to a greater or lesser extent" (2004, 1788). Furthermore, on the basis of translation data of *in fact* into Dutch, they add another use of *in fact*, which they refer to as *for in fact*. Their proposal is that the different uses of *in fact* have in common the meaning of 'in truth'. In other words, the core meaning of *in fact* is "to signal the speaker's attitude that something is in reality or in truth the case". But there are at least three pragmatic uses which emerge from the Dutch translations of *in fact*: those equivalents which express a contrast (i.e., equivalents of the ' but' type of in fact), those which express a stronger reformulation (i.e., equivalents of the ' and' type), and those which express a reason (i.e., equivalents of the ' for' type). The latter use was not considered by Schwenter and Traugott in their monolingual data but was discovered when analysing the translations into Dutch. The DM *really* has received much less attention than *in fact* and *actually* in the literature (see, Stenström 1986, Simon-Vandenbergen 1988), so it will not be discussed here, but will rather be studied from the translation perspective.

A detailed translation analysis of each of these DMs is presented in the following subsections.

3.2 Translation analysis of *in fact*

The English DM *in fact* is translated in more than half of the cases by *de hecho* in Spanish (56,4%), followed at a distance by other translations such as *en realidad* (16,8%), *en efecto* (7,5%), *efectivamente* (4,1%), and *incluso* (2,7%), as shown in Table 1:

Table 1. *In fact* in Spanish translations

English original	Spanish translation	Frequency	Percentage
In fact	De hecho	12600	56,47185371
	En realidad	3765	16,87432772
	En efecto	1692	7,583363213
	efectivamente	924	4,141269272
	incluso	623	2,792219433
	precisamente	306	1,371459304
	efectivamente	924	4,141269272
	realmente	424	1,900322696
	Además	447	2,003406239
	Más bien	147	0,658838293
	Más aún	30	0,134456795
	Es más	430	1,927214055
TOTAL		**22312**	

These frequencies indicate a strong translation equivalence between *in fact* and *de hecho*, which is confirmed when looking at the English translations. Here *in fact* occurs in 70% of the cases as the translation of *de hecho*, while it only occurs in smaller proportions as the translation of other markers such as *en efecto* (8,9%), *en realidad* (6,2%), *efectivamente* (4,9%), as shown in Table 2.

Table 2. *In fact* in English translations: Spanish sources

Spanish original	English translation	Frequency	Percentage
De hecho	In fact	13181	70,04463811
En efecto		1689	8,975449038
En realidad		1175	6,244021681
efectivamente		928	4,931448613
Además		450	2,391327452
realmente		439	2,332872781
Es más		431	2,290360293
Precisamente		310	1,647358912
La verdad		149	0,79179509
De verdad		36	0,191306196
Más aún		30	0,15942183
TOTAL		**18818**	

When *in fact* is translated by *de hecho* it is used in its core epistemic meaning to express a high degree of certainty, i.e., "to signal the speaker's attitude that something is in reality or in truth the case" (Aijmer and Simon-Vandenbergen 2004), as illustrated by (1), where the DM is in bold:

(1) a. *With regard to other factors, the deficit ratio has not exceeded the ratio of public investment to GDP since 1997.* ***In fact****, there have been budget surpluses since 1998.* (ECB)

b. *En cuanto a otros factores, la ratio de déficit no ha superado a la inversión pública en porcentaje del PIB desde 1997.* ***De hecho****, se han producido superávit presupuestarios desde 1998.*

The analysis of the less frequent translations of *in fact* shows that this marker can be used with other meanings, some of which have been identified in monolingual studies (Falk, 2006); for example, the translation by *en efecto* has a clear "confirmatory" meaning in Spanish, highly related but not completely interchangeable with the core epistemic one, as illustrated by (2):

(2) a. *Next, as a matter of urgency, we muss pass legislation, firstly, regarding identification of the exact characteristics of goods transported.* ***In fact****, according to the experts, the oil the Erika was carrying was supposed to sink to the bottom and should never have reached the coast.* (EUROPARL3)

b. *Después, debemos legislar urgentemente, y ante todo, para que se conozcan las características exactas de las mercancías transportadas.* ***En efecto****, según los expertos, el petróleo del Erika debía manar por el fondo y nunca llegar a las costas.*

This use had been previously detected in monolingual studies (Llopis Cardona 2012) and is also expressed by other close markers such as *efectivamente*, which also has a "confirmatory" meaning in Spanish, as illustrated by (3):

(3) a. *The water issue will definitely not be settled today.* ***In fact****, the problems of enlargement and climate change will offer new prospects.* (EUROPARL3)

b. *De todos modos, hoy no se va a concluir el expediente del agua.* ***Efectivamente****, la cuestión de la ampliación y las evoluciones climáticas abren nuevas perspectivas.*

The translation by *en realidad* points to a "corrective / counterargumentative" meaning, illustrated by (4):

(4) a. *Looking at it from the point of view of social and employment policy we certainly do not have a legislative programme before us.* ***In fact*** *we have not had a legislative programme for several years now.* (EUROPARL3)

b. *Si enfocamos la cuestión desde el punto de vista de la política social y de empleo es evidente que no tenemos ante nosotros un programa legislativo.* ***En realidad*** *no hemos tenido un programa legislativo desde hace varios años*

The "additive/elaborative" meaning of *in fact*, previously identified in other studies (Aijmer & Simon-Vandenbergen 2004), is captured by the Spanish translations *además, es más*, and *más aún*, as illustrated by (5), (6) and (7):

(5) a. ***In fact****, the word seems to have been banished from the Commission's vocabulary, whereas liberal and American references abound.* (EUROPARL3)

b. ***Además****, la palabra "social" parece estar proscrita del vocabulario de la Comisión, mientras que abundan las referencias liberales y norteamericanas.*

(6) a. *You could hardly expect this Commissioner to agree with a tax which would, in fact, impede international economic movement.* ***In fact****, we were shaken by it.* (EUROPARL3)

b. *Así, difícilmente pueden ustedes esperar de este comisario que esté a favor de un impuesto que precisamente supone un obstáculo al tráfico económico internacional.* ***Es más****, nos sentimos completamente estremecidos.*

(7) a. *In many Islamic countries, girls are genitally maimed, although this is not mentioned anywhere in the Koran.* ***In fact****, the Koran prohibits this practice.* (EUROPARL3)

b. *En numerosos países islámicos se les practica la ablación del clítoris a las mujeres. Todo ello sin que el Corán diga nada al respecto.* ***Más aún****: el Corán lo prohíbe.*

Another use which has been overlooked in monolingual studies is the *specifying* one, captured by the Spanish translation *precisamente*, as in (8) and (9):

(8) a. *We must seize this opportunity and* ***in fact*** *this is a chance for us to win support for our principles there.* (EUROPARL3)

b. *Hemos de aprovechar esta posibilidad y* ***precisamente*** *tendremos también la ocasión y la posibilidad de lograr que presten oídos allí a nuestros principios.*

(9) a. *I am taking Kingsley to Red' s old environment.* ***In fact****, to the Midtown Hotel which Cannon used as his hideout.* (EUROPARL3)

b. *Llevo a Kingsley a el antiguo ambiente de Red. Más* ***precisamente****, al Hotel-Midtown, el que Cannon usó como escondite.*

A summary of the Spanish translations of *in fact* and the identified meanings is presented in Table 3:

Table 3. Spanish translations of *in fact* and its meanings

English DM	Spanish translation	Meaning
In fact	De hecho	Epistemic
	En efecto/efectivamente	Confirmatory
	En realidad	Counterargumentative/corrective
	Incluso, además, es más, más aún	Additive
	Precisamente	Specifying

3.3 Translation analysis of *actually*

The most frequent translations of *actually* into Spanish are *realmente* (29,6%), and *en realidad* (23%). These are followed by other less frequent markers such as *de hecho* (12,5%), *efectivamente* (11%), *la verdad* (6,5%), and *de verdad* (3,8%), *incluso* (3%), *en verdad* (2,9%), *en efecto* (1,6%), as shown in Table 4:

Table 4. *Actually* in Spanish translations

English original	Spanish translation	Frequency	Percentage
actually	realmente	5509	29,65175736
	en realidad	4451	23,95715593
	de hecho	2332	12,5518058
	efectivamente	2053	11,05011034
	la verdad	1226	6,598848162
	de verdad	708	3,810754077
	incluso	573	3,08412724
	en verdad	547	2,944184294
	en efecto	303	1,630873567
	verdaderamente	262	1,410194305
	es más	219	1,178750202
	además	207	1,11416115
	precisamente	189	1,017277571
TOTAL		**18579**	

These frequencies indicate a strong translation equivalence between *actually*, on the one hand, and *realmente* and *en realidad*, on the other, while the other translations seem to be less central. This tendency can also be observed in the English translations, where *actually* occurs as the translation of *en realidad*, and *realmente* with similar frequencies to the Spanish translations (20,1% and 18,9% respectively). However, it also occurs as the translation of *de hecho* in 35,4% of the cases, as shown in Table 5:

Table 5. *Actually* in English translations: Spanish sources

Spanish original	English translation	Frequency	Percentage
De hecho	actually	5072	35,4734928
En realidad		2885	20,17764722
Realmente		2710	18,95369982
La verdad		1226	8,574625822
De verdad		702	4,909777591
En verdad		543	3,797733949
Además		455	3,182263254
En efecto		297	2,077213596
Es más		218	1,524688768
Precisamente		190	1,328857183
TOTAL		**14298**	

The translation equivalents of *actually* point to different meaning aspects of this DM. Thus, when *actually* is translated by *de hecho*, it is used as an emphasizer, intensifying the truth value of the clause in which it occurs, as in (10) below:

(10) a. *His Charter is really misnamed because it will* ***actually*** *reduce these rights.* (EUROPARL3)
b. *Señalo bien, "sin razón" ya que,* ***de hecho****, va a reducir esos derechos.*

Other translations such as *realmente* or *de verdad* are used with a strong epistemic meaning, identifying the truth value of the proposition, as illustrated by (11), (12) and (13) below:

(11) a. *The Commission must track down the illegal aid and the aid which* ***actually*** *hinders the internal market.* (EUROPARL3)
b. *La Comisión debe perseguir las ayudas ilegales y aquellas que* ***realmente*** *ponen cortapisas al mercado interior.*

(12) a. *It is of course difficult to determine how long the Serbian population will need to continue fighting before Milosevic will* ***actually*** *disappear off the scene.* (EUROPARL3)
b. *Por supuesto, es difícil determinar ahora cuánto tiempo tendrá que seguir el pueblo serbio con las luchas hasta que Milosevic haya desaparecido* ***de verdad****.*

(13) a. *Is he confident now that these substances are* ***actually*** *safe for children?* (EMEA)
b. *¿Está seguro de que esas substancias son* ***de verdad*** *inocuas para los niños?*

The translation by *la verdad* points to two possible uses: (a) a counter-argumentative or corrective one, as in (14); and (b) an evaluative one, as in (15):

(14) a. *I want a child from you.* ***Actually****, I want two or three kids!* (OPENSUBTITLES2011)

b. *Quiero un hijo tuyo.* ***La verdad*** *es que quiero dos o tres*

(15) a. *I do not* ***actually*** *believe that this is appropriate in the European Parliament.* (EUROPARL3)

b. ***La verdad*** *es que no me parece procedente en el Parlamento Europeo.*

While the former use often occurs in first-initial position, the latter is more frequent in medial position. The counter-argumentative or corrective function found in (14) is very similar to the one expressed by *en realidad*, which is used in the context of a contradiction and/or disagreement, as in (16):

(16) a. *He sends a check every week to his sweet, gray-haired mother.* ***Actually****, she' s silver- haired.* (OPENSUBTITLES2011)

b. *¿ Manda un cheque cada semana a su dulce mamá de pelo blanco?* ***En realidad****, su pelo es gris.*

While this corrective function has been identified in monolingual studies, the evaluative one has emerged in the translation analysis.

Another function which has not been identified in monolingual studies is the confirmatory one, as shown by the Spanish translations *en efecto* and *efectivamente*, as illustrated by (17) and (18):

(17) a. *The European Union should* ***actually*** *place social issues at the heart of the building of Europe.* (EUROPARL3)

b. ***En efecto****, la construcción de la Unión Europea debería hacerse en torno a lo social.*

(18) a. *Having set it up, I hope the Commission* ***actually*** *makes use of it and therefore does not just allow it to sit around in the corridors and gather dust.* (EUROPARL3)

b. *Habiéndolo establecido espero que la Comisión Europea haga* ***efectivamente*** *uso del mismo y, por tanto, no permita que languidezca en los pasillos y quede cubierto de polvo.*

This confirmatory function is shared with *in fact*, as we saw before.

Finally, the corpus analysis also allows to identify an additive/elaborative function of *actually*, as revealed by translations such as *es más*, as in (19):

(19) a. *Leopardi, the greatest Italian poet. Actually, THE Italian poet.* (OPENSUBTITLES2011)

b. *Leopardi, el mejor poeta italiano. Es más, El Poeta.*

The translation correspondences of *actually* and their identified meanings can be summarised and grouped, as shown in Table 6:

Table 6. Spanish translations of *actually* and its meanings

English DM	Spanish translation	Function / Meaning
Actually	De hecho	Emphasizer
	En efecto/efectivamente	Confirmatory
	Realmente/de verdad	Epistemic (truth identifier)
	La verdad	Counterargumentative/corrective
		Evaluative
	Es más	Additive / Elaborative

3.4 Translation analysis of *really*

In the Spanish translations, the English DM *really* is mostly translated by *realmente* (79,4%), followed at a distance by *en realidad* (14,7%) and *de hecho* (1,2%) as shown in Table 7:

Table 7. *Really* in Spanish translations

English original	Spanish translation	Frequency	Percentage
really	realmente	34916	79,45205479
	en realidad	6462	14,70440996
	la verdad	1649	3,752332408
	de hecho	529	1,203750057
	efectivamente	248	0,564328949
	en efecto	142	0,323123834
	TOTAL	**43946**	

This tendency also occurs in the English translations, with almost the same proportions as in the Spanish translations, with *really* occurring in 81,7% of the cases as the translation of *realmente*, but only in 15,1% of the cases as the translation of *en realidad*, and in 1,2% as the translation of *de hecho*, as shown in Table 8:

Table 8. *Really* in English translations

Spanish original	English translation	Frequency	Percentage
realmente	really	34363	81,74659815
En realidad		6379	15,17508802
De hecho		515	1,225140356
Además		390	0,927776192
Es más		239	0,568560282
efectivamente		249	0,592349415
En efecto		140	0,333047864
TOTAL		**42036**	**81,74659815**

These frequencies indicate a strong translation equivalence between *really* and *realmente*.

As to the possible meanings/functions of *really*, its core pragmatic meaning is to identify the truth value of the proposition, as illustrated in (20), where it is translated by *realmente*:

(20) a. *It is advisable to test your blood sugar immediately after taking glucose to check that you* ***really*** *have hypoglycaemia.* (EMEA)
b. *Es recomendable analizar su nivel de azúcar en sangre inmediatamente después de la ingestión de glucosa para confirmar que padece* ***realmente*** *hipoglucemia.*

However, as pointed out by previous monolingual studies, *really* can also be used as an intensifier or emphasizer, as in (21):

(21) a. *I* ***really*** *do wonder whether this is an approach of which the green Environment Ministers in my country are aware.* (EUROPARL3)
b. *Me pregunto* ***realmente*** *si esto se ha hecho con la aquiescencia de los ministros ecologistas de medio ambiente de mi país.*

When translated by *en realidad*, a highly polysemic item in Spanish, it can be used with very different meanings, such as the adversative one, especially when preceded by "pero" ("but"), as in (22):

(22) a. *In the case of swaps or FRAs, the outstanding amount is notional but it is* ***really*** *used for calculating the amounts of interest effectively exchanged between two parties.* (ECB)
b. *En el caso de swaps o de acuerdos de tipos de interés futuros, el saldo vivo es ficticio, pero* ***en realidad*** *se utiliza para calcular el importe correspondiente a intereses intercambiado efectivamente entre dos partes.*

When translated by *de hecho*, it is used as an emphasizer, as in (23):

(23) a. *Why delay this proposed legislation, which* ***really*** *ought already to have been in place?* (EUROPARL3)
b. *¿Por qué retrasar esta propuesta que,* ***de hecho****, ya debería estar vigente?*

When translated by *la verdad*, it has an evaluative meaning, as in (24):

(24) a. *For over a year the old Commission could have done some work on this. We* ***really*** *have not progressed very far.* (EUROPARL3)
b. *Durante más de un año la antigua Comisión podía haber trabajado algo en el tema.* ***La verdad*** *es que no hemos avanzado mucho.*

Other less frequent translations by *en efecto* or *efectivamente* point to meanings of confirmation, as in (25):

(25) a. *And that is what we are discussing at this late hour. This* ***really*** *is a subject that is close to the citizen, so it is a pity it is being debated at such a late hour.* (EUROPARL3)

b. *Así, es lógico que se plantee la pregunta de qué hace la UE para proteger el euro de falsificaciones. De ello estamos hablando en hora tan tardía.* ***En efecto****, es un tema que preocupa al ciudadano.*

The translation equivalents of *really* and the identified meanings are summarised in Table 9:

Table 9. Spanish translations of *really* and its meanings

English DM	Spanish translation	Meaning
really	realmente	truth identifier
	en realidad	adversative
	de hecho	emphasizer
	efectivamente, en efecto	confirmation
	la verdad	evaluative

3.5 Analysis of the back translations

This section focuses on the analysis of the back translations of the English DMs studied above, i.e, the English translations of the most frequent Spanish equivalents of *in fact, actually* and *really*, i.e., the markers *de hecho, en realidad*, and *realmente*.

Many of these Spanish DMs have been studied from a monolingual perspective in reference works on Spanish discourse markers (see Martín Zorraquino and Portolés 1999, Portolés 2007, Loureda Lamas and Acín Villa 2010), but also in other more specific studies, often combining the diachronic with the synchronic analysis, thus providing an interesting developmental perspective on their paradigmatic relationships, their semantics and their pragmatics.

The marker *de hecho* has been studied by several researchers (Fuentes Rodriguez 1994, 1996; García Negroni 2011; Fanego 2010 *inter alia*). Fanego (2010) provides a very thorough study of this marker from the perspective of grammaticalization. She distinguishes several uses of *de hecho* in Present-day Spanish which have much in common with the functions reported in the literature for its English, French and Italian cognates, since they all go back to the Latin noun *factum* 'deed, action'. Apart from its use as an adverbial of manner (which she refers to as de hecho$_1$), with the meaning "in practice, de facto", the three main uses identified in her study are the confirmatory (*de hecho*$_{2Conf}$), the epistemic-adversative (*de hecho*$_{2Adv}$) and the elaborative (de hecho$_3$) one. She concludes that "these three uses are better interpreted as generalized conversational implicatures, that is, default inferences and conventions of use in language-specific communities that can be exploited to imply/insinuate certain meanings, but may however be cancelled." (Fanego 2010, 19).

For González Manzano (2013), the markers introduced by the preposition "en" in Spanish kept their original etymological meaning referring to a metaphorical and hypothetical space of reality, certainty or truth ("in reality", "in truth"), opposed to the space of irreality or falseness, which explains the contrastive value which they can present. In addition, markers such as *en realidad* and the closely related *en verdad* conventionalized counterargumentative functions derived from their original locative meaning (Canes Nápoles & Delbecque 2017).

A detailed translation analysis of each of these DMs is presented in the following subsections.

3.5.1 *English translations of* de hecho

In the English back translations, the Spanish DM *de hecho* is mostly translated by *in fact* (63,2% of the cases), followed by *actually* (24,3%) and by others at a distance, such as *indeed* (8,3%) and *really* (2,4%), as shown in Table 10:

Table 10. *De hecho* and its English translations

	English translation	**Frequency**	**Percentage**
De hecho	in fact	13181	63,26070263
	actually	5072	24,34248416
	indeed	1735	8,326934152
	really	515	2,471683624
	in reality	66	0,316759455
	moreover	148	0,71030908
	furthermore	119	0,571126896
	TOTAL	**20836**	

The high number of translations of *de hecho* by *in fact* indicates that these two items share a high number of uses in both languages and confirms the strength of their equivalence, which was already observed in the Spanish translation of *in fact* by *de hecho*, as described in Section 3.2.

The most frequent use in these cases is the confirmatory one, as illustrated in (26) and (27):

(26) a. *Señora Presidenta, considero que tiene mucha razón nuestra colega Maij-Weggen cuando califica la situación de catastrófica.* ***De hecho****, es una situación complicada, con muchos muertos.* (EUROPARL3)

b. *Madam President, I think that Mrs Maij-Weggen is quite right to call the situation catastrophic. It is,* ***in fact****, a complex situation, in which many people have died.*

(27) a. *Como ha dicho la gente, la situación allí es extremadamente precaria.* ***De hecho****, existe el riesgo de un golpe militar en el futuro.*

b. *As people have said, the situation there is extremely volatile. There is,* ***in fact****, a risk of a military coup in the future.*

It is also possible to find additive/elaborative uses, captured by the translation *indeed*, as in (28):

(28) a. *En primer lugar, este asunto está en proyecto desde 1993, por tanto no ha sido ninguna sorpresa para ninguno de los otros productores.* ***De hecho****, cualquiera de estos solicitantes puede aún presentar una solicitud para su producto en particular.* (EUROPARL3)

b. *First of all, this matter has been in the pipeline since 1993, so it has not really been sprung on any other commodity producers.* ***Indeed*** *any such applicants can still apply in respect of their particular product.*

In addition, the adversative use of *de hecho* is frequently translated by *in reality*, as in (29):

(29) a. *Señor Presidente, el Presidente Chirac ha banalizado, calificándolas de irreales y no pragmáticas, las reacciones negativas de muchos diputados ante el resultado de Niza.* ***De hecho****, el irrealista es el Presidente si piensa que con un Tratado así será posible ampliar la Unión.* (EUROPARL3)

b. *Mr President, President Chirac has played down the negative reactions of many of the Members to the outcome of Nice, describing them as unrealistic and not pragmatic.* ***In reality****, it is the President who is being unrealistic if he thinks that it will be possible to enlarge the Union with such a Treaty.*

Besides these three main uses, which have been previously identified in monolingual studies, the translation analysis revealed other related pragmatic ones: for example, the second most frequent translation of *de hecho* is *actually*, with a clear emphatic use, intensifying the truth of the proposition, as illustrated in (30):

(30) a. *Para mí, en cuanto al gasto público, el año crucial es siempre el año uno – tienes suerte si llegas al año dos o tres. El año uno es,* ***de hecho****, el gran año de gasto para nosotros.* (EUROPARL3)

b. *For me, in public-spending terms, the crucial year is always year one – you are lucky if you get to year two or three. Year one is* ***actually*** *the big year of expenditure for us.*

The English translations of *de hecho* and their identified meanings can be grouped as shown in Table 11:

Table 11. Meanings of *de hecho* in English back translations

Spanish DM	English translation	Meaning
De hecho	in fact	confirmation
	actually	emphasis
	indeed	addition/elaboration
	in reality	adversative

3.5.2 *English translations of* en realidad

In the English back translations, the Spanish DM *en realidad* is mostly translated by *actually* (25% of the cases), and by *in fact* (22,1% of the cases). Other translations are less frequent, as shown in Table 12:

Table 12. *En realidad* and its English translations

Spanish original	English translation	Frequency	Percentage
En realidad	actually	674	25,0371471
	really	335	12,44427935
	in fact	596	22,13967311
	indeed	241	8,952451709
	in reality	451	16,75334324
	in effect	156	5,794947994
	moreover	62	2,303120357
	furthermore	46	1,708766716
	truly	131	4,866270431
	TOTAL	**2692**	

All these different translations reflect the highly polysemic nature of *en realidad*, which has also been pointed out in monolingual studies (Canes Napoles & Delbecque 2017). It also indicates that *en realidad* is much closer semantically to *actually* and to *in fact* than to other items such as *indeed* or *furthermore*.

When translated by *actually*, the most frequent meaning is to express the speaker's high certainty in the truth of the proposition, as in (31):

(31) a. *El primero es un informe sobre el sistema de subvenciones de las leguminosas de grano y el segundo es una propuesta para mejorar la tramitación administrativa del sistema. ¿De qué tipo de plantas se trata* ***en realidad****?* (EUROPARL3)

b. *The first is a report on the aid scheme in place for grain legumes, and the second a proposal for improved administration of the system. What vegetables are we* ***actually*** *talking about?*

When translated by *really*, its main use is as emphasizer, as in (32):

(32) a. *¿En que situación se halla la política de asilo europea? Es difícil decirlo, porque* ***en realidad*** *no hay una política europea de asilo.* (EUROPARL3)

b. *What shape is the European asylum policy in? That is hard to say, as there is not* ***really*** *a European asylum policy to speak of.*

When translated by *in fact*, a highly polysemic DM as well, it has a corrective function, as in (33):

(33) a. *Para finalizar, quiero decir que nos encontramos aún muy al principio.* ***En realidad*** *estamos al principio del principio.*

b. *I should like to close by saying that we are still just right at the beginning. We are* ***in fact*** *at the beginning of the beginning.*

But it can also have an adversative meaning preceded by “pero” (“but”), as in (34):

(34) a. *Digo bosques tropicales, pero,* ***en realidad****, un nuevo aspecto de este reglamento, en comparación con el anterior, consiste en que el ámbito de aplicación de el reglamento comprende ahora todos los países en desarrollo, y, por tanto, incluye países como Sudáfrica, China y las regiones de el Mediterráneo.* (EUROPARL3)

b. *I say tropical forests, but* ***in fact*** *a new feature of this regulation compared to the previous one is that the scope of the regulation now covers all developing countries and hence now includes countries such as South Africa, China and the Mediterranean and Middle East regions.*

When *en realidad* is translated by *indeed*, it tends to be used with an elaborative meaning, as in (35):

(35) a. *Sobre todas esas cuestiones, la delegación de el Parlamento presentó un sólido alegato ante el Consejo en las negociaciones y logró muchas concesiones.* ***En realidad****, desde la mayoría de los puntos de vista, se podría considerar que las negociaciones tuvieron éxito.* (EUROPARL3)

b. *On all these issues, Parliament ’s delegation made strong representations to the Council in the negotiations and won many concessions.* ***Indeed****, from most points of view, it could and would be assumed that the negotiations were a success.*

When preceded by "pero" ("but") and translated by *in reality*, the meaning expressed is adversative, as in (36):

(36) a. *La Cumbre de Tampere trató de la lucha contra la delincuencia. Todo el mundo estaba como muy satisfecho, pero* ***en realidad*** *se ha visto poco progreso.* (EUROPARL3)

b. *The Tampere Summit was about combating crime. Everyone was supposedly pleased about it but,* ***in reality****, we have made little progress.*

The English translations of *en realidad* and their identified meanings can be grouped as shown in Table 13:

Table 13. Meanings of *en realidad* in English back translations

Spanish DM	English translation	Meaning
En realidad	Actually	truth value
	in fact	corrective
	really	emphasis
	indeed	addition/elaboration
	In reality	adversative

3.5.3 *English translations of* realmente

In the English back translations, the Spanish DM *realmente* is translated by *really* in the majority of the cases (72,4%), followed at quite a distance by *actually* (13.3%), *truly* (8%), *indeed* (3%) and *in fact* (1.8%), as shown in Table 14:

Table 14. *Realmente* and its English translations

	English translation	Frequency	Percentage
Realmente	really	30026	72,48280024
	actually	5533	13,35666868
	truly	3383	8,166566083
	indeed	1596	3,852745926
	in fact	776	1,873264937
	in reality	84	0,202776101
	in effect	27	0,065178033
TOTAL	**Total**	**41425**	

The high number of translations of *realmente* by *really* confirms the strength of their equivalence, which was already observed in the first mirroring when looking at the Spanish translation of the English DM *really* by *realmente*.

The most frequent use is the epistemic one, expressing the speaker's certainty in the truth value of the proposition, as in (37):

(37) a. *Para nosotros, en cambio, es importante que la UE se concentre en pocos sectores, en aquellos en los que* ***realmente*** *pueda ser útil.* (EUROPARL3)
b. *For us, it is important that the EU should instead concentrate on just a few areas in which it can* ***really*** *be of use.*

When translated by *actually*, it is used as an emphasizer, as in (38):

(38) a. *Esto quiere decir que si nosotros, los europeos, nos esforzamos a tiempo, seremos también competitivos en este ámbito de las nuevas tecnologías. ¿En dónde están* ***realmente*** *los déficit de Europa?* (EUROPARL3)
b. *This means that if we in Europe get our act together in good time then we will also be competitive in this field of new technologies. Where then do Europe's shortcomings* ***actually*** *lie?*

When translated by *truly*, it also emphasizes the truth value of the proposition, as (39) and (40):

(39) a. *Señor_Presidente en ejercicio de el Consejo,* ***realmente*** *lamento su respuesta.* (EUROPARL3)
b. *Mr President-in-Office of the Council, I* ***truly*** *regret your reply.*

(40) a. *La sede de la agencia no debería estar en Barcelona, Helsinki o Italia, sino en Internet, en el ciberespacio. Así sería eficiente, transparente y accesible. Sería* ***realmente*** *una solución del siglo XXI.* (EMEA)
b. *It would be efficient, transparent and accessible. This would be a* ***truly*** *twenty-first century solution.*

When translated by *indeed* it has a highly emphatic function, as in (41):

(41) a. *Usted ha dicho esta mañana que va a apoyar la enmienda nº 19 cuando se vote aquí sobre la misma. Solamente quiero tener la confirmación de que va a recoger* ***realmente*** *esta enmienda nº 19.* (EUROPARL3)
b. *You said this morning that you would endorse Amendment No 19 at today 's vote. I just wanted to confirm that you will* ***indeed*** *accept Amendment No 19.*

The rather infrequent translations by *in fact* point to a strong epistemic meaning, as in (42) and (43):

(42) a. *Todavía se culpa a menudo a esas mujeres en muchos países y,* ***realmente,*** *son las víctimas.* (EUROPARL3)
b. *All too often in many countries, those women are criminalised whereas,* ***in fact****, they are the victims.*

(43) a. *Esta última constituye en mi opinión el corazón de la problemática de mi informe y determinará la orientación de nuestra votación. Esta pregunta es: ¿Qué significa* ***realmente*** *la ciudadanía europea?* (EUROPARL3)

b. *This is in my view the crux of the problem described in my report and will determine the way we vote. This question is: what* ***in fact*** *does European citizenship really mean?*

The English translations of *realmente* and their identified meanings can be summarised as shown in Table 15:

Table 15. Meanings of *realmente* in English translations

Spanish DM	English translation	Meaning
Realmente	really	epistemic/truth value epistemic/emphasis
	actually	emphasis
	indeed	emphasis
	in reality	adversative
	in fact	epistemic

3.6 Lexico-semantic field construction

As shown in the previous subsections, the analysis of the translations and back translations of the three English DMs investigated in this study has revealed interesting translation preferences which capture different meanings and pragmatic functions of these markers, some of which had not been identified in previous monolingual analysis. Moreover, as pointed out by Aijmer & Simon-Vandenbergen (2004), it should be possible to use translations as tools to construct lexico-semantic fields, which could then serve as a *tertium comparationis* for an in-depth cross-linguistic analysis.

Figure 1 shows the lexico-semantic field of the three English DMs *in fact, actually*, and *really*, as it emerges from their translations. The three items are related to each other since they share translations and sources in Spanish, but *in fact* and *actually* are more multifunctional and more central to the field since they have a higher number of links. Also, while some Spanish equivalents are linked to all three English words, others are linked only to two (e.g. *precisamente*) or to one only (e.g. *incluso*).

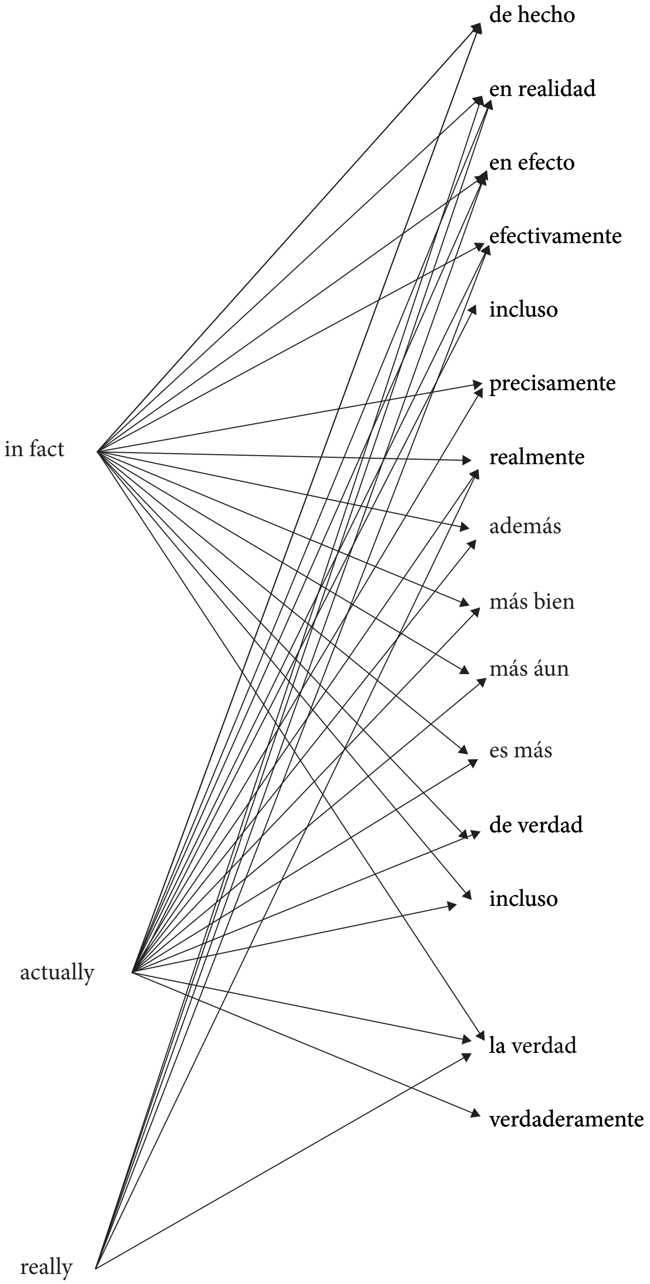


Figure 1. Lexico-semantic field of *in fact, actually and really*

Figure 2 presents the second translation image where the back translations add items to the initial ones, e.g.: *in reality, moreover, furthermore, in effect, truly*. The items with higher number of links are more multifunctional and more central in the field. In Spanish, it is *de hecho, en realidad* and *realmente* which seem to be most multifunctional, and this is why they are selected as sources in Figure 2.

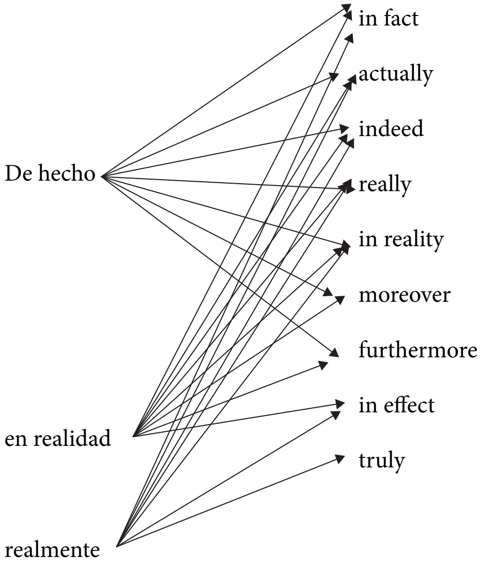


Figure 2. The second translation image: English translations of the most frequent Spanish translations of *in fact, in reality and really*

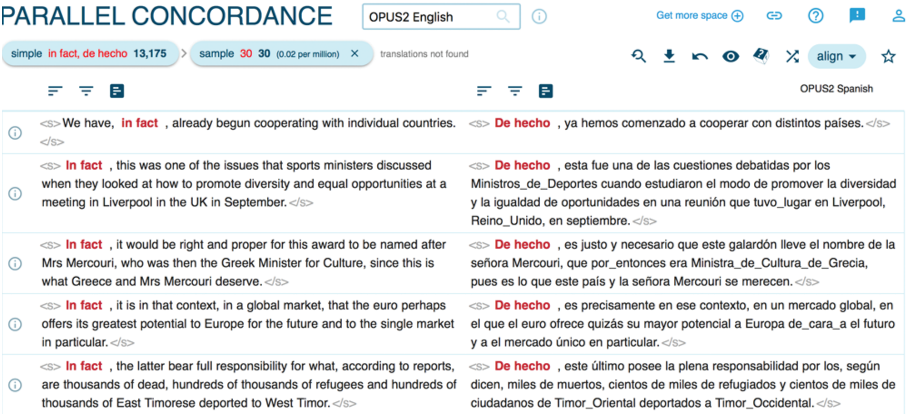
An item with three lines is an equivalent of all three Spanish DMs and belongs to the core of the lexico-semantic field; this is the case of *in fact, actually* and *really*; an item with two lines is less central to the field, such as *moreover* and *furthermore*; an item with only one line, such as *truly* is a translation of only one item –*realmente*- is more marginal, but not uninteresting, since it belongs to a subfield related to the core items.

4. Corpus annotation phase

On the basis of the results of the corpus analysis phase, where the meanings and uses of the English and the Spanish DMs were identified and grouped into a lexico-semantic field, the second phase of this study presents an annotation model for DMs consisting of several steps:[3]

[3] The annotation model proposed here is inspired in the annotation methodology described by Hovy & Lavid (2010).

1. The first step is to identify and prepare a training corpus as starting material. This is a selection of representative texts on which annotations will be performed. In the case of the investigated DMs, the training corpus can be prepared by downloading from the Sketch Engine platform a random sample of bitexts from the different subcorpora of the Opus2 Corpus, as shown in Figure 3:



OPUS2 English	OPUS2 Spanish
<s>We have, **in fact** , already begun cooperating with individual countries. </s>	<s> **De hecho** , ya hemos comenzado a cooperar con distintos países. </s>
<s> **In fact** , this was one of the issues that sports ministers discussed when they looked at how to promote diversity and equal opportunities at a meeting in Liverpool in the UK in September. </s>	<s> **De hecho** , esta fue una de las cuestiones debatidas por los Ministros_de_Deportes cuando estudiaron el modo de promover la diversidad y la igualdad de oportunidades en una reunión que tuvo_lugar en Liverpool, Reino_Unido, en septiembre. </s>
<s> **In fact** , it would be right and proper for this award to be named after Mrs Mercouri, who was then the Greek Minister for Culture, since this is what Greece and Mrs Mercouri deserve. </s>	<s> **De hecho** , es justo y necesario que este galardón lleve el nombre de la señora Mercouri, que por_entonces era Ministra_de_Cultura_de_Grecia, pues es lo que este país y la señora Mercouri se merecen. </s>
<s> **In fact** , it is in that context, in a global market, that the euro perhaps offers its greatest potential to Europe for the future and to the single market in particular. </s>	<s> **De hecho** , es precisamente en ese contexto, en un mercado global, en el que el euro ofrece quizás su mayor potencial a Europa de_cara_a el futuro y a el mercado único en particular. </s>
<s> **In fact** , the latter bear full responsibility for what, according to reports, are thousands of dead, hundreds of thousands of refugees and hundreds of thousands of East Timorese deported to West Timor. </s>	<s> **De hecho** , este último posee la plena responsabilidad por los, según dicen, miles de muertos, cientos de miles de refugiados y cientos de miles de ciudadanos de Timor_Oriental deportados a Timor_Occidental. </s>

Figure 3. English-to-Spanish parallel concordance in Sketch Engine

2. The second step is to design an annotation scheme and guidelines on the basis of the empirical findings emerging from the translation analysis. Annotation schemes usually consist of a core tagset, with the coarser annotation tags, and an extended tagset, where finer-grained tags are specified. In the case of the DMs investigated in this study, the core tagset consists of two main tags: *anti-oriented* [AO] and *co-oriented* [CO], respectively.[4] The former refers to meanings of contrast, opposition or contrary to expectation, as illustrated by the uses of *en realidad* as in (33) above, repeated here for convenience as (44):

(44) a. *Digo bosques tropicales, pero, **en realidad**, un nuevo aspecto de este reglamento, en comparación con el anterior, consiste en que el ámbito de aplicación del reglamento comprende ahora todos los países en desarrollo, y, por tanto, incluye países como Sudáfrica, China y las regiones de el Mediterráneo.* (EUROPARL3)

b. *say tropical forests, but **in fact** a new feature of this regulation compared to the previous one is that the scope of the regulation now covers all developing countries and hence now includes countries such as South Africa, China and the Mediterranean and Middle East regions.*

[4] The labels 'co-oriented' and 'anti-oriented' have been proposed by A. Canes Napoles and N. Delbecque in their study on the polysemy of the Spanish marker "en realidad" (2017: 179). Here they have been adapted to separate the clusters of meanings specified in the annotation scheme.

Co-oriented tags refer to meanings which modulate or elaborate the information of the main proposition in several possible ways. This can be information related to truth identification [TI], correction [COR], evaluation [EV], reformulation [RF], specification [SPE], recapitulation [RC], intensification [INT], addition [ADD] and confirmation [CONF].
For example, the meaning of truth identification [TI] is illustrated by *in fact* and its translation into Spanish by *de hecho* in (1) above, repeated below as (45):

(45) a. *With regard to other factors, the deficit ratio has not exceeded the ratio of public investment to GDP since 1997.* ***In fact****, there have been budget surpluses since 1998.* (EUROPARL3)

b. *En cuanto_a otros factores, la ratio de déficit no ha superado a la inversión pública en porcentaje del PIB desde 1997.* ***De hecho****, se han producido superávit presupuestarios desde 1998.*

The meaning of correction [COR] is exemplified by *en realidad* and its translation by *in fact*, as in (46):

(46) a. *Para finalizar, quiero decir que nos encontramos aún muy al principio.* ***En realidad*** *estamos al principio del principio.*

b. *I should like to close by saying that we are still just right at the beginning. We are* ***in fact*** *at the beginning of the beginning.*

The meaning of intensification is illustrated by *actually* and its Spanish translation *de hecho* in (47):

(47) a. *His Charter is really misnamed because it will* ***actually*** *reduce these rights.* (EUROPARL3)

b. *Señalo bien, " sin razón" ya que,* ***de hecho****, va a reducir esos derechos.*

Addition is illustrated by *in fact* and its translation by *además* in (48), while the meaning of confirmation is illustrated by *in fact* and its translation *en efecto* in (49):

(48) a. ***In fact****, the word seems to have been banished from the Commission' s vocabulary, whereas liberal and American references abound.* (EUROPARL3)

b. ***Además****, la palabra " social " parece estar proscrita del vocabulario de la Comisión, mientras que abundan las referencias liberales y norteamericanas.*

(49) a. *Next, as a matter of urgency, we muss pass legislation, firstly, regarding identification of the exact characteristics of goods transported.* ***In fact****, according to the experts, the oil the Erika was carrying was supposed to sink to the bottom and should never have reached the coast.* (EUROPARL3)

b. *Después, debemos legislar urgentemente, y ante todo, para que se conozcan las características exactas de las mercancías transportadas.* ***En efecto****, según los expertos, el petróleo del Erika debía manar por el fondo y nunca llegar a las costas.*

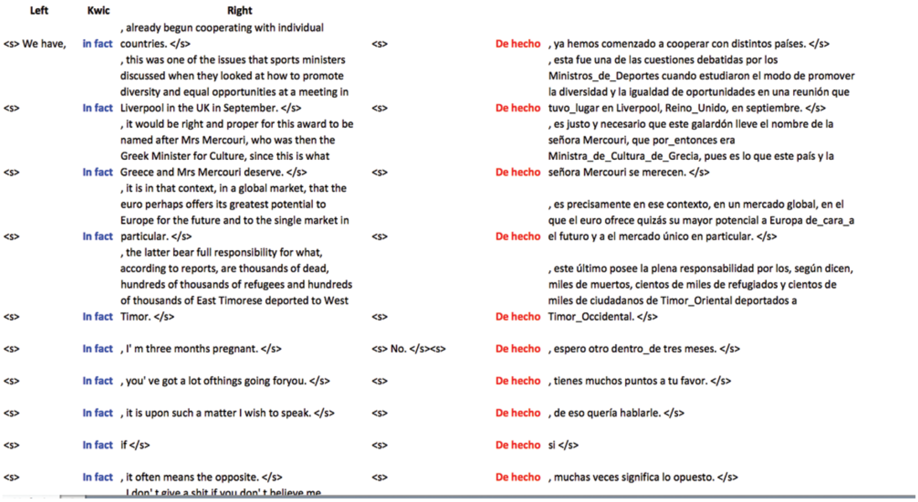
The core and the extended tagsets of the proposed annotation scheme are presented in Table 16:

Table 16. Annotation tagsets for studied DMs

Core tagset	Extended tagset
Anti-oriented [AO] Co-oriented [CO]	Adversative [ADV]
	Truth identifier [TI]
	Corrective [COR]
	Evaluator [EV]
	Reformulating [RF]
	Specifying [SPE]
	Recapitulation [RC]
	Intensifier [INT]
	Additive/elaborative [ADD]
	Confirmatory [CONF]

In addition to the annotation scheme, it is necessary to write annotation instructions or guidelines (the 'annotation manual' or 'codebook) to guide annotators in the annotation process. In the case of DM, annotators must receive precise instructions on which tags to choose and how to deal with difficult cases. For example, they will first choose one of the general tags from the core tagset, i.e., *co-oriented* or *anti-oriented*, and then will proceed to choose the more specific ones from the extended tagset. If in doubt about which fine-grained tag to choose or in cases of disagreement, annotators can keep only one of tags from the core tagset, or assign double labels. For example, if in doubt between a *corrective* meaning and a *reformulating* one, annotators can simply choose the more general tag *co-oriented*; also, if one annotator uses a *truth identifier* tag [TI] for an ocurrence of one the DMs and the other annotator uses the *intensifying* one [INT], this disagreement can be resolved by assigning a double label [TI]- [INT].

3. The third step is to annotate some fragment of the training corpus, in order to determine the feasibility of the annotation scheme and guidelines. In the case of the DMs, this can be perfomed by saving the bitexts in Excel format and manually annotate them with the tags of the annotation schema, as shown in Figure 4 for *in fact*:



Left	Kwic	Right			
<s> We have,	in fact	, already begun cooperating with individual countries. </s>	<s>	De hecho	, ya hemos comenzado a cooperar con distintos países. </s>
<s>	In fact	, this was one of the issues that sports ministers discussed when they looked at how to promote diversity and equal opportunities at a meeting in Liverpool in the UK in September. </s>	<s>	De hecho	, esta fue una de las cuestiones debatidas por los Ministros_de_Deportes cuando estudiaron el modo de promover la diversidad y la igualdad de oportunidades en una reunión que tuvo_lugar en Liverpool, Reino_Unido, en septiembre. </s>
<s>	In fact	, it would be right and proper for this award to be named after Mrs Mercouri, who was then the Greek Minister for Culture, since this is what Greece and Mrs Mercouri deserve. </s>	<s>	De hecho	, es justo y necesario que este galardón lleve el nombre de la señora Mercouri, que por_entonces era Ministra_de_Cultura_de_Grecia, pues es lo que este país y la señora Mercouri se merecen. </s>
<s>	In fact	, it is in that context, in a global market, that the euro perhaps offers its greatest potential to Europe for the future and to the single market in particular. </s>	<s>	De hecho	, es precisamente en ese contexto, en un mercado global, en el que el euro ofrece quizás su mayor potencial a Europa de_cara_a el futuro y a el mercado único en particular. </s>
<s>	In fact	, the latter bear full responsibility for what, according to reports, are thousands of dead, hundreds of thousands of refugees and hundreds of thousands of East Timorese deported to West Timor. </s>	<s>	De hecho	, este último posee la plena responsabilidad por los, según dicen, miles de muertos, cientos de miles de refugiados y cientos de miles de ciudadanos de Timor_Oriental deportados a Timor_Occidental. </s>
<s>	In fact	, I' m three months pregnant. </s>	<s> No. </s><s>	De hecho	, espero otro dentro_de tres meses. </s>
<s>	In fact	, you' ve got a lot ofthings going foryou. </s>	<s>	De hecho	, tienes muchos puntos a tu favor. </s>
<s>	In fact	, it is upon such a matter I wish to speak. </s>	<s>	De hecho	, de eso quería hablarle. </s>
<s>	In fact	if </s>	<s>	De hecho	si </s>
<s>	In fact	, it often means the opposite. </s>	<s>	De hecho	, muchas veces significa lo opuesto. </s>
		I don' t give a shit if you don' t believe me			

Figure 4. Annotation file for *in fact* (in blue) and its Spanish translations (in red)

The annotation should be ideally performed by two expert annotators working independently on the training corpus of bitexts described above.

4. The fourth step is to measure the results of the annotation, i.e., comparing the annotators' decisions, deciding which measures are appropriate and how they should be applied, as well as the level of agreement is satisfactory. However, as explained by Hovy & Lavid, a high agreement –at 90%– is simply unreachable for many complex problems, and it is the intended use of the annotated corpus which will determine the level of agreement which will be considered satisfactory: for training machine translation systems there should be enough annotated data at high enough agreement, but for the identification of relevant linguistic phenomena and validation of linguistic theories, "perhaps it doesn't matter what the agreement level is, as long as poor agreements are seriously investigated". (Hovy & Lavid 2010).

   The final step is to decide on the size of the corpus to be annotated. This will depend on the intended use of the annotations: if the goal is to train a machine translation system, a large portion of the corpus should be annotated, possibly over several months or years, with many intermediate checks, improvements, etc. If the goal is to identify meanings and pragmatic uses of the DMs through their translation correspondences, a smaller corpus would probably be useful for investigating those meanings.

5. Summary and concluding remarks

This paper has proposed a methodology for the study of DMS in translation which can be useful not only for descriptive research in this area but also for the creation of annotated bitexts with DMs in different languages. Although applied to a specific case study, i.e. the English DMs *in fact, actually* and *really* and their Spanish translations, the methodology can be generalized to other DMs and to other language pairs, and form the basis for large-scale cross-linguistic annotation of DMs.

The corpus analysis phase has revealed that by looking at their translations and corresponding back translations it is posible to identify meanings which may have been overlooked in a purely monolingual analysis. It has also shown that this method allows to group core and peripheral meanings and to construct a lexico-semantic field where similar meanings are close together in the field and more distant meanings are further away from each other.

In the corpus annotation phase an annotation model for DMs has been proposed consisting of several steps to ensure that the human-coded annotations are reliable and consistent (gold standard annotations). The level of agreement which will be considered satisfactory and the size of the corpus to be annotated will depend on the goal of the annotations: if the goal is to train a computational system so that it can learn from the annotated data, it will be necessary to annotate a larger corpus at high enough agreement, but if the goal is to investigate translation correspondences of DMs, a small annotated corpus is probably enough.

Future work will focus on performing annotation experiments of the DMs studied in this paper, and on applying the proposed methodology to other related DMs within the semantic field of elaboration, as recently initiated with English and Spanish DMs expressing digression (see Lavid 2019). The results of these empirical studies will form the basis for future work on large-scale cross-linguistic annotation of DMs for NLP applications such as MT and MTG.

References

Aijmer, Karin and Anne Marie Simon-Vandenbergen. 2004. A model and a methodology for the study of pragmatic markers: the semantic field of expectation. *Journal of Pragmatics* 36: 1781–1805. https://doi.org/10.1016/j.pragma.2004.05.005

Aijmer, Karin. 2007. "Translating discourse particles: a case of complextranslation." In *Incorporating Corpora: The Linguist and the Translator*, 95–116. Clevedon: Multilingual Matters. https://doi.org/10.21832/9781853599873-009

Aijmer, Karin. 2013. Understanding pragmatic particles: a variational pragmatic approach. Edinburgh: Edinburgh University Press, 2013. Pp. ix + 162.

Aijmer, Karin. 2015. Analysing discourse markers in spoken corpora: actually as a case study. In *Corpora and discourse studies*, 88–109. London: Palgrave Macmillan. https://doi.org/10.1057/9781137431738_5

Bazzanella, C. & Morra, L. 2000. "Discourse markers and the indeterminacy of translation." In *Argomenti per una linguistica della traduzione. On linguistic aspects of translation. Notes pour une linguistique de la traduction*, edited by I. Korzen & C. Marello, 149–157). Alessandria: Edizioni dell'Orso.

Beeching, Kate. 2013. "A parallel corpus approach to investigating semantic change." In *Advances in Corpus-Based Contrastive Linguistics: Studies in honour of Stig Johansson*, 103–125. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.54.07bee

Canes Nápoles, Amalia and Nicole Delbecque. 2017. "En realidad, polisemia y polifuncionalidad de un marcador discursivo." *Revista Internacional de lingüística iberoamericana* 29, 173–205.

Carretero, M. 2012. "Los adverbios ingleses certainly, naturally y surely, y su traducción al español." In *Telar de traducción especializada*, 141–152. Madrid: Dykinson.

Clift, Rebecca. 2001. "Meaning in interaction: the case of *actually*." *Language* 77(2): 245–491. https://doi.org/10.1353/lan.2001.0074

Dyvik, Helge. 2004. "Translations as Semantic Mirrors: from Parallel Corpus to WorldNet." In *Language and Computers, Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, 311–326. Amsterdam: Rodopi. https://doi.org/10.1163/9789004333710_019

Dyvik, Helge. 2005. "Translations as a Semantic Knowledge Source." In *Proceedings of the second Baltic Conference on Human Language Technologies*, 27–38. Tallin: Estland.

Falk, Johan. 2006. "En efecto es su cumpleaños mañana: observaciones sobre el marcador del discurso en efecto." In *Discurso, interacción e identidad: homenaje a Lars Fant*, edited by Johan Falk, Johan Gille and Fernando Wachtmeister Bermúdez, 37–63. Stockholm: Stockholm University.

Fanego, Teresa. 2010. "Paths in the development of elaborative discourse markers: Evidence from Spanish." In *Subjectification, Intersubjectification and grammaticalization*, edited by Kristin Davidse, Lieven Vandelanotte and Hubert Cuyckens, 197–237. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110226102.2.197

Fuentes Rodriguez, Catalina. 1994. "Usos discursivos y orientación argumentativa: de hecho, en efecto, efectivamente." *Español Actual* 62: 5–18.

Fuentes Rodríguez, Catalina. 1996. *La sintaxis de los relacionantes supraoracionales*. Madrid: Arco Libros.

García Negroni, María Marta. 2011. "En efecto, efectivamente y de hecho: confirmación, acuerdo y prueba en el discurso científico escrito en español." In *Los discursos del saber: prácticas discursivas y enunciación académica*, edited by María Marta García Negroni, 23–40. Buenos Aires: Editoras del Calderón.

González Manzano, Mónica. 2013. "Gramaticalización de los marcadores epistémicos en español." PhD diss., Universitat de Barcelona.

Hovy, Eduard and Julia Lavid. 2010. "Towards a science of corpus annotation: a new methodological challenge for Corpus Linguistics." *International Journal of Translation* https://www.cs.cmu.edu/~hovy/

Kilgarriff, Adam, Baisa, Vit, Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychlý, Pavel, and Vít Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography*, 1: 7–36. https://doi.org/10.1007/s40607-014-0009-9

Lavid, Julia. 2019. "Translation correspondences of Digressive Discourse Markers in English and Spanish: A Corpus-Based Study." In *Computational and Corpus-Based Phraseology*, edited by Gloria Corpas Pastor and Ruslan Mitkov, 239–252. Switzerland: Springer Lecture Notes in Artificial Intelligence 11755. https://doi.org/10.1007/978-3-030-30135-4_18

Lavid, Julia and Estefanía Avilés. 2019. "La anotación textual de los marcadores elaborativos en inglés y en español: un estudio de corpus." In *Contrastes inglés-español y traducción: estudios basados en corpus*, edited by Julia Lavid López: Madrid: Escolar y Mayo, 39–59.

Llopis Cardona. Ana. 2012. "Entre la modalidad y la conexión: la confirmación. El caso de en efecto." *RILCE* 31(2): 405–34.

Loureda Lamas and Acín Villa. 2010. *Los estudios sobre marcadores del discurso en español, hoy*. Madrid: Arco/Libros.

Martín Zorraquino, M. Antonia, Portolés, José. 1999. "Los marcadores del discurso". In *Gramática descriptiva de la lengua española*, edited by Ignacio bosque y Violeta Demonte, 4053–4200. Madrid: Espasa Calpe.

Noël, Dirk. 2003. "Translations as evidence of semantics: an illustration." *Linguistics* 41(4): 757–785. https://doi.org/10.1515/ling.2003.024

Oh, Sun-Young. 2000. "*Actually* and *in fact* in American English: a data-based analysis." *English Language and Linguistics* 4 (2), 243–268. https://doi.org/10.1017/S1360674300000241

Portolés, José. 2007. *Marcadores del discurso*. Barcelona: Ariel.

Simon-Vandenbergen, Anne-Marie. 1988. "What really really means in casual conversation and in political interviews." *Linguistica Antverpiensia* 22, 206–225.

Smith, Sara W., Jucker, Andreas H. 2000. "*Actually* and other markers of an apparent discrepancy between propositional attitudes of conversational partners." In *Pragmatic Markers and Propositional Attitudes*, edited by Andersen, G., and Fretheim, T., 207–237. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/pbns.79.10smi

Stenström, Anna-Brita. 1986. "What does really really do?" In *Strategies in Speech and Writing: English in Speech and Writing: A Symposium*, edited by Tottie, G., Bäcklund, I., 149–163. Stockholm: Almqvist & Wiksell.

Schwenter, Scott, Traugott, Elizabeth C. 2000. "Invoking scalarity: the development of in fact." *Journal of Historical Pragmatics* 1 (1), 7–25. https://doi.org/10.1075/jhp.1.1.04sch

Taglicht, Josef. 2001. "Actually, there's more to it than meets the eye." *English Language and Linguistics* 5(01): 1–16. https://doi.org/10.1017/S1360674301000119

Tiedemann, J. 2009. "News from OPUS – a collection of multilingual parallel corpora with tools and interfaces." In *Recent Advances in Natural Language Processing* (vol. V), edited by N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov, 237–248. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/cilt.309.19tie

Tognini-Bonelli, E. 1993. "Interpretative nodes in discourse: *actual* and *actually*." In *Text and technology: in honour of John Sinclair*, edited by M. Baker, G. Francis, and E. Tognini-Bonelli, 193–212. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/z.64.13tog

Appendix. List of subcorpora within Opus2 parallel corpus

- ECB – European Central Bank corpus (v.0.1)
- EMEA – European Medicines Agency documents (v.0.3)
- EUconst – The European constitution (v.0.1)
- EUROPARL – European Parliament Proceedings (v.3) [transcripts of spoken language]
- Opensubs – Open Subtitles corpus (v.2) [transcripts of spoken language]
- KDE4 – KDE4 localization files (v.2)
- KDEdoc – KDE manual corpus
- MultiUN – Translated UN documents
- OpenOffice – a collection of documents from http://www.openoffice.org/
- OpenOffice (v.3) – a collection of documents from http://www.openoffice.org/
- OpenSubtitles2011 – Open Subtitles corpus (2011 version) [transcripts of spoken language]
- RF – Regeringsförklaringen – Declarations of Government Policy by the Swedish Government
- SETIMES2 – A parallel corpus of the Balkan languages (v.2)
- SPC – Stockholm Parallel Corpora (v.1)
- TEP – The Tehran English-Persian subtitle corpus (v.0.1) [transcripts of spoken language]
- Tatoeba – a collection of translated sentences from Tatoeba
- TedTalks – transcription and translation of TED talks [transcripts of spoken language]
- UN – Translated UN documents
- hrenWaC – Croatian-English Parallel Web Corpus

The discourse markers *well* and *so* and their equivalents in the Portuguese and Turkish subparts of the TED-MDB corpus

Amália Mendes and Deniz Zeyrek
CLUL – Centre of Linguistics, School of Arts and Humanities, University of Lisbon / Graduate School of Informatics, Cognitive Science Department, Middle East Technical University, Ankara

The present research describes how the PDTB style of discourse annotation has been applied to transcribed TED Talks in the TED-MDB corpus, which include multifunctional elements such as *well* and *so*. The occurrence of these two markers is analyzed in English in comparison to the way they are conveyed in Portuguese and Turkish translations of TED-MDB. Their presence in discourse relations, as well as in question-response pairs found in TED Talks, are investigated. The implicitation of *well* and *so* in translation is shown to be related to their function in the context. Results also point to differences in the target languages regarding implicitation, especially when the discourse marker doesn't fulfil a connective function.

Keywords: translation, TED Talks genre, PDTB discourse annotation, discourse markers, question-response pairs, Portuguese, Turkish, English

Introduction

Discourse markers form a class of elements, such as "*well, but, oh, y'know*… that function in cognitive, expressive, social, and textual domains" (Schiffrin 2001: 54). They are multifunctional items typically found in speech, and operationally, they may be identified by several criteria; for example, they are short and phonologically reduced, restricted to sentence-initial position, have little or no propositional meaning, and are informal (Brinton, 1996). They have been of interest to contrastive linguistic research, translation studies, native versus non-native speech, and sociolinguistics (c.f. Aijmer & Simon-Vandenbergen 2011, and the references therein).

https://doi.org/10.1075/btl.158.08men

While discourse markers (DMs) are typical elements of conversational interaction, they can also be found in recently emerging genres, such as TED speeches.

TED Talks have the properties of both spoken and written discourse: they are prepared, possibly scripted monologues aimed to be informative, delivered in English to a live audience. They are transcribed according to the norms of written language with punctuation marks and later translated to other languages with time-stamped lines (sentences and groups of sentences) aligned with the original transcripts. The English transcripts eliminate some speech dysfluencies, such as extra-linguistic filled pauses and interrupted words, but keep DMs in the original language. It is of interest to discourse analysis in general and notably to cross-linguistic analyses to understand whether the emergent genre of TED talks translations keep DMs or eliminate them. From a linguistic perspective, this approach enables a better understanding of the nature of DMs in monolingual contexts. From a cognitive science perspective, the elimination strategy of the translator is interesting because the eliminated tokens, as opposed to the retained tokens, may reveal what the translator has focused on during the translation process.

The TED Talk speeches are especially useful for research on the multilingual analysis of DMs as the TED talk website provides translations of a large set of typologically varied languages; they also cover a wide range of topics, and they are open to the public, making them fully available for research.

Given the full accessibility of TED talks, and their characteristics common to written and spoken discourse, they offer a perfect data type for the analysis of DMs. We examine *well* and *so* in TED talks in the original language, English, and in the translations of two typologically different languages, Portuguese and Turkish. Jucker (1997) mentions four functions of *well*, including its frame-marker role, as well as its roles at the interpersonal level (operating as a face-threat mitigatory, qualifier, and pause filler). *So*, on the other hand, may have a propositional (semantic) meaning, and shares pragmatic properties with *well*, such as the structural function that indicates start, closing, pre-closing, continuity, topic change, or reformulation (Cuenca 2008, Buysse 2015, Crible et al. 2019). In the present study, we consider *well* and *so* to belong to a general class of elements with propositional, structural, and modal meanings (Cuenca and Marín 2009). Our use of the term DM covers these two markers with propositional, structural, and attitudinal meanings.

A recently released multilingual resource, namely TED-Multilingual Discourse Bank or TED-MDB, aims to engender research on the discourse structure of multiple languages to the extent that discourse relations are concerned (Zeyrek et al. 2018, 2019). This corpus includes the original English transcripts of six TED talks and their translation to German, Polish, Portuguese, Russian, Turkish as well as Lithuanian (Oleškevičienė et al. 2018) annotated at the discourse level with a coherence-based approach taking discourse connectives as anchors of coherence relations. It uses the PDTB (Penn Discourse TreeBank) style of annotation, which was initially designed for the annotation of the *Wall Street Journal* in English (Prasad et al. 2007). It has been subsequently applied to the written texts of several languages, such as Arabic (Al-Saif et al. 2011), Chinese (Zhou et al. 2015), Hindi (Oza et al. 2009), Turkish (Zeyrek et al. 2013) as well as conversational Italian (Tonelli et al. 2010).

The present study uses the English, Portuguese, and Turkish subparts of TED-MDB as the data. Parallel corpora that cover a large set of languages are scarce, even more so if we consider data annotated for discourse relations and discourse markers. TED-MDB is the first multilingual corpus that includes translated data in several languages, annotated in the same framework for discourse relations, and aligned. It provides a unique opportunity to inspect how discourse markers are treated in translated texts. The research is framed within the notion of implicitation in translation, which refers to the process of elimination of lexical elements in the target languages in the context of parallel corpora (Hoek et al. 2017; Crible et al. 2019). Our initial observations on various subparts of TED-MDB have been that DMs are not easy to translate literally, and the target languages differ in their translation strategies, particularly regarding the DMs that have a structural function or a modal meaning indicating the attitude, knowledge or stance of the speaker. Thus, in the present research, implicitation refers to the elimination of DMs in target languages.

We start with three working hypotheses:

- DMs with a structural function or modal meaning are more prone to implicitation than DMs with a propositional meaning. A previous study has proven that underspecified discourse markers, such as *and*, tend to be omitted in translations (Crible et al. 2019). Likewise, we would expect that DMs with a propositional meaning that contribute to the connectivity and interpretation of the sentences will tend to be translated. In contrast, DMs with non-propositional meaning will be understood as non-essential by the translator.
- Some languages are more prone to implicitation in translated texts. Given that we are observing two typologically different languages, Portuguese and Turkish, we might find differences in their treatment of DMs in translation.
- The discourse markers with a structural function or modal meaning that are kept in the translation will tend to be left out of the discourse relation annotation. Although the translator may have kept in the target language the DMs that do not fulfil a propositional meaning, we would expect the annotator to recognize that these are not discourse connectives of a discourse relation.

Research aims and method of analysis

Given the full accessibility of TED talks, and their characteristics common to written and spoken discourse, the present research aims to present a study of the translation of *so* and *well* in the source language (English) and Portuguese and Turkish translations of TED talks from TED-MDB within a cross-linguistic approach. It hopes to contribute to the current contrastive line of research on DMs that take parallel corpora as its input, and add to the existing knowledge on DMs that have not yet been central in the annotation of coherence relations of TED talks in TED-MDB.

As TED-MDB annotates discourse relations, *so* and *well* are examined to the extent that they appear in the annotated relations in the corpus. Our approach is data-driven, and we ask why *so* and *well* are omitted in some places and not others, proposing explanations about the data. The overall goal throughout the study is to reveal the pattern of use of *so* and *well* in Portuguese and Turkish translations.

Methodologically, all tokens of *well* and *so* in the English subpart and their equivalents in Portuguese and Turkish were manually specified in TED-MDB, extracted and transferred to spreadsheets. The use of *well* and *so* that fall out of the function of DM were left out of scope (e.g., *to do <u>well</u> financially; responsibility to do <u>so</u>*). The frequency of occurrence of the DMs kept or eliminated in the target texts were quantified and extended by qualitative assessments.

In the rest of the chapter, the TED-MDB corpus is presented. The view from a recent speech annotation study is overviewed, and the question-answer annotation scheme in various discourse banks are outlined, as these structures are frequently found in TED-MDB in relation to the two DMs under analysis. Data analysis is presented comparing the use of *well* to *so* in English and the translations. The conclusion discusses the findings of the research with the starting hypotheses.

The TED-MDB corpus

With a connective-based approach, TED-MDB annotates discourse relations, alternatively referred to as coherence relations or rhetorical relations, which are semantic relations that hold between text segments and carry labels such as temporal, contingency, comparison, expansion (Prasad et al. 2007). The corpus was compiled by extracting the transcripts of six talks in the seven languages from the WIT3 website (Cettolo et al. 2012). The texts were selected, making sure the author was a native speaker of a variety of English and that the talks represented a variety of topics. Texts which mostly relied on images and videos were avoided. The contents of TED-MDB are provided in Table 1, and the number of words per language is presented in Table 2.

Table 1. TED Talks annotated for 7 languages in TED-MDB

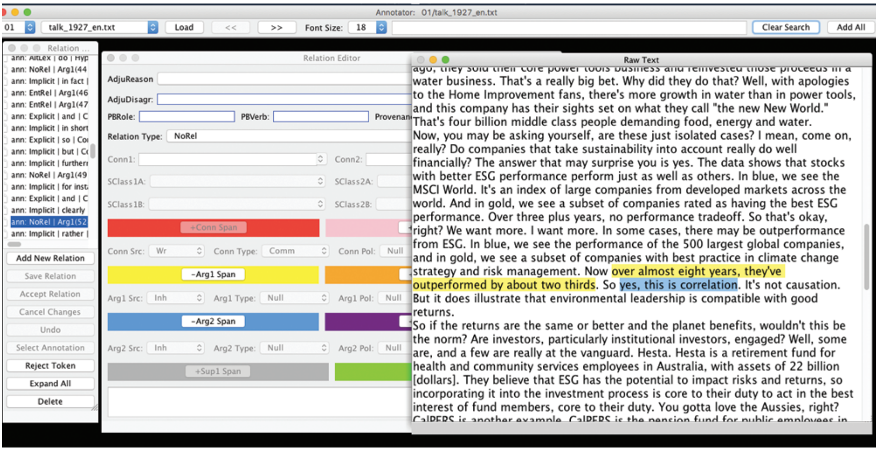
ID	Author	Title
1927	C. McKnett	The investment logic for sustainability
1971	D. Sengeh	The sore problem of prosthetic limbs
1976	J. Kasdin	The flower-shaped star-shade that might help us detect Earth-like planets
1978	S. Lewis	Embrace the near win
2009	K. Cahana	A glimpse of life on the road
2150	D. Troy	Social maps that reveal a city's intersections and separations

Table 2. Number of words per language in the corpus

Language	Number of words
English	7012
Portuguese	7166
Turkish	5164
Total	**19342**

TED-MDB has been created by expert native speaker annotators of discourse. Each mono-lingual team minimally consisted of a primary annotator, who was typically an experienced researcher, or the lead researcher of the team, and a secondary annotator. The inter-annotator agreement was calculated over 25% of the corpus. Across languages, the F-score value for discourse relation spotting was 0.70, and Cohen's κ ≥ 0.70 was reached for discourse relation type and top-level sense identification, which is considered a good standard for this task (Spooren and Degand, 2010). Detailed information on the annotation cycle and the inter-annotator agreement experiment is provided in Zeyrek et al. (2019).

Following the PDTB guidelines enriched with issues specific to TED talks, the annotators manually examine and annotate each talk in their language to avoid a possible bias from English annotations. Following the annotation scheme of the PDTB, TED-MDB annotates five sets of relations (Explicit, Implicit, Alternative Lexicalization/AltLex, Entity Relation/EntRel, No Relation/NoRel) and assigns a sense to them, where relevant, using the PDTB 3.0 sense hierarchy (Webber et al. 2019). The annotations are created by using the PDTB annotation tool, a Java-based tool developed specifically for the annotation of discourse relations in the scope of the PDTB project (Lee et al. 2016). Figure 1 illustrates the three sub-panels of the tool: the Relation List panel, on the left, contains all the annotation tokens of that file; the Relation Editor panel, on the centre, provides all the features for the annotation of a discourse relation; the Raw Text panel, on the right, shows the text. The annotations are stored in stand-off format, as pipe-delimited files, and indexed to the raw texts. Additionally, the tool doubles up as an adjudication tool to create gold files, and automatically creates an XML file (adjudication_groups.xml) with the set of annotations.



01 | talk_1927_en.txt | Font Size: 18

ann: AltLex | do | Hyp
ann: NoRel | Arg1(44
ann: Implicit | in fact |
ann: EntRel | Arg1(46
ann: EntRel | Arg1(47
ann: Explicit | and | C
ann: Implicit | in short
ann: Explicit | so | Co
ann: Implicit | but | Co
ann: Implicit | further
ann: NoRel | Arg1(49
ann: Implicit | for inst
ann: Explicit | and | C
ann: Implicit | clearly
ann: NoRel | Arg1(52
ann: Implicit | rather |

AdjuReason:
AdjuDisagr:
PBRole: PBVerb: Provenan
Relation Type: NoRel
Conn1: Conn2:
SClass1A: SClass2A:
SClass1B: SClass2B:
Conn Src: Wr Conn Type: Comm Conn Pol: Null
Arg1 Src: Inh Arg1 Type: Null Arg1 Pol: Null
Arg2 Src: Inh Arg2 Type: Null Arg2 Pol: Null

ago, they sold their core power tools business and reinvested those proceeds in a water business. That's a really big bet. Why did they do that? Well, with apologies to the Home Improvement fans, there's more growth in water than in power tools, and this company has their sights set on what they call "the new New World." That's four billion middle class people demanding food, energy and water.
Now, you may be asking yourself, are these just isolated cases? I mean, come on, really? Do companies that take sustainability into account really do well financially? The answer that may surprise you is yes. The data shows that stocks with better ESG performance perform just as well as others. In blue, we see the MSCI World. It's an index of large companies from developed markets across the world. And in gold, we see a subset of companies rated as having the best ESG performance. Over three plus years, no performance tradeoff. So that's okay, right? We want more. I want more. In some cases, there may be outperformance from ESG. In blue, we see the performance of the 500 largest global companies, and in gold, we see a subset of companies with best practice in climate change strategy and risk management. Now over almost eight years, they've outperformed by about two thirds. So yes, this is correlation. It's not causation. But it does illustrate that environmental leadership is compatible with good returns.
So if the returns are the same or better and the planet benefits, wouldn't this be the norm? Are investors, particularly institutional investors, engaged? Well, some are, and a few are really at the vanguard. Hesta. Hesta is a retirement fund for health and community services employees in Australia, with assets of 22 billion [dollars]. They believe that ESG has the potential to impact risks and returns, so incorporating it into the investment process is core to their duty to act in the best interest of fund members, core to their duty. You gotta love the Aussies, right?
CalPERS is another example. CalPERS is the pension fund for public employees in

Figure 1. A snapshot of the PDTB annotation tool showing a NoRel relation that contains *so*

For each relation type (except the types EntRel and NoRel, which receive no sense tag), the annotator chooses a low-level sense of the PDTB 3.0 sense hierarchy. For instance, the sense tag Result is part of the broader set of Cause relations, which is itself contained in the top-level Contingency sense (Contingency:Cause:Result) (see (1b) below).

As in the PDTB, TED-MDB takes discourse connectives as discourse-level predicates with binary arguments that always have an abstract object interpretation (Asher 2012). The arguments are referred to as Arg1 and Arg2, where Arg2 is the text piece that hosts the connective. Discourse relations may be expressed both explicitly and implicitly. In Explicit discourse relations (cf. (1a)), the relation holding between two text segments is expressed by an overt discourse connective. In implicit discourse relations, there is no overt connective (cf. (1b)). The reader then infers the relation from the adjacency of the clauses and their lexical content, and the inferred relation is made explicit by inserting a possible discourse connective in the context (thus, an "implicit" relation). Explicit relations are annotated both at the intra- and inter-sentential levels, but implicit relations have so far been annotated only at the inter-sentential level. Throughout the present study, the annotation scheme is illustrated through different fonts. The Explicit connective is underlined, Arg1 is presented in italics, Arg2 in bold letters. Unannotated parts are shown in standard fonts.

(1) a. *About 80 percent of global CEOs see sustainability as the root to growth in innovation* <u>and</u> **leading to competitive advantage in their industries**. But 93 percent see ESG as the future. [Explicit] [Expansion:Conjunction]

b. *Prosthetists still use conventional processes like molding and casting to create single-material prosthetic sockets.* **Such sockets often leave intolerable amounts of pressure on the limbs of the patient, leaving them with pressure sores and blister**
[Implicit=consequently] [Contingency:Cause:Result]

Finally, discourse relations may be expressed by lexical means other than discourse connectives. These are named Alternative Lexicalizations, as in (2).

(2) … *long-term value creation requires the effective management of three forms of capital: financial, human, and physical.* <u>This is why</u> **we are concerned with ESG.**
[AltLex] [Contingency:Cause+Belief:Result+Belief]

A detailed account of the talks included in the corpus and the annotation process can be found in Zeyrek et al. (2018, 2019).

Related work: The view from speech annotation

The PDTB style of discourse annotation was initially performed over the written texts of a single language, English, and then applied to written genres in the discourse banks for other languages. During the annotation of TED-MDB, various aspects of the spoken discourse were encountered, such as DMs, which occur as part of discourse relations as well as question-response (QR) pairs, where the speaker asks a question and immediately answers it. These QR structures are motivated by the need to make the presentation livelier and less expository, used as a device to capture the attention of the audience, so they are associated with the rhetorical function of language. They are not, however, strictly speaking, rhetorical questions, as they do not express an assertion utilizing a question.

The DMs spotted in discourse relations and QR pairs in TED-MDB had to be tackled by means of an annotation framework designed for written texts. Thus, the TED-MDB annotation guidelines were extended with new instructions for the annotators explaining how DMs had to be distinguished from discourse connectives. The annotators were told to mark discourse connectives together with their binary arguments, and in case they spotted DMs, they should be left out of the annotated token. The QR pairs had to be annotated as Alternative Lexicalizations, again leaving out any DMs that occur with them. Nevertheless, this was only a temporary solution, and the TED-MDB team aims to enrich its annotation scheme with aspects of spoken genre in the future.

A scheme for spoken language (Crible 2017)

Crible (2017) has designed an annotation scheme for speech and applied it to spoken data. Different from the PDTB, which focuses on (the explicitly or implicitly marked) discourse relations, Crible (2017) targets discourse markers, which also include discourse connectives and does not consider implicit relations. The scheme is structured in four domains: Ideational, Rhetorical, Sequential, and Interpersonal. Each function is associated with a list of senses. For instance, Cause is linked to the Ideational domain, Motivation to Rhetorical, Opening Boundary to Sequential, and Monitoring to Interpersonal. A revised version of the scheme keeps the four main domains but uses a smaller list of functions that are potentially linked to all domains (cf. Table 3) (Crible and Degand 2017). Contrast can operate at any of the four domains, depending on the context. The following examples are taken from Crible and Degand (2017) to illustrate Contrast in the Ideational domain (Example (3)), in the Sequential domain (Example (4)) and the Interpersonal domain (Example (5)). See the full discussion in Crible and Degand (2017: 19–20).

(3) nous sommes animés par le désir de participer à notre échelle au progrès de la connaissance mais nos liens avec l'université sont aussi fragiles
‘we are moved by the desire to participate at our own scale to the progress of knowledge mais (‘but') our links with the university are fragile too'

(4) <spk1> euh j'aime les néologismes j'aime les les régionalismes mais euh je mets le point d'exlamation dessus euh pour dire euh attention
<spk2> mais la norme qu'est-ce qu'est-elle pour vous
<spk1> ‘uh I like neologisms I like regionalisms but uh I write an exclamation mark on them uh to say uh careful'
<spk2> mais (‘but') ‘the norm what is it to you'

(5) Alors cet auditeur vigilant il va vous dire tiens euh encore Jean d'Ormesson mais on entend Jean d'Ormesson à chaque automne
‘well this careful listener he will tell you look uh Jean d'Ormesson again mais (‘but') we hear Jean d'Ormesson every fall.'

Table 3. Cross-domain taxonomy in Crible and Degand (2017: 18)

<table>
<tr><th>Ideational</th><th>Rhetorical</th><th>Sequential</th><th>Interpersonal</th></tr>
<tr><td colspan="4">[addition] [alternative] [cause] [closing] [concession] [condition]
[consequence] [contrast] [enumeration] [opening] [punctuation]
[resuming][temporal] [topic-shift] [specification]</td></tr>
</table>

This scheme has been applied to TED Talks in a multilingual experiment that investigates English discourse markers, focusing on their functions, omission, and translation equivalents in Czech, French, Hungarian and Lithuanian (Crible et al. 2019). DMs are taken as "expressions that met the criteria of being syntactically optional (not integrated in any syntactic relation), formally fixed (grammaticalized), and having a procedural meaning and a discourse-structuring function" (Crible et al. 2019: 143). The study points out that there are two relatively frequent markers (with more than ten occurrences) which are never translated into all four languages, namely, *now* and *then*, followed closely by *and* and *so*. Discourse markers that are omitted in the translations are mainly cases of speech-specific markers functioning as "punctuators" (*okay, now, then*), and underspecified markers expressing a wide range of discourse relations (*and*).

Treatment of question-response sequences in various frameworks and TED-MDB

In a news corpus such as the PDTB, it is possible to find QR sequences in speech produced by a single writer/speaker or in reported speech strategies (e.g., in texts based on interviews that involve at least two speakers). Independently of this difference, in the PDTB 2.0 (Prasad et al. 2007, 2008), question-answer sequences are labelled as an explicit or implicit relation and bear an appropriate sense tag of the PDTB 2.0 sense hierarchy. Example (6) illustrates the annotation scheme used in the PDTB 2.0.

(6) **Why constructive**? <u>Because</u> *despite all the media prattle about comedy and politics, not mixing, they are similar in one respect: Both can serve as mechanisms for easing tensions and facilitating the co-existence of groups in conflict.*
[Explicit] [Contingency:Cause:Reason] [wsj-2369]

In Carlson and Marcu (2001), where a section of the Penn Treebank has been annotated in the RST framework, QR sequences are annotated as Question-Answer (with subtypes according to which segment is nucleus or satellite). This scheme distinguishes question-answer sequences (truly interactional) from rhetorical questions (assertions presented as questions), which receive a specific label.

In yet another corpus, namely the STAC dataset, QR sequences are treated as fully interactional, as the corpus is composed of multiparty dialogues (chats) extracted from on-line game sessions (Asher et al. 2016, 2017). These QR sequences are annotated in the style of the SDRT theory (Asher and Lascarides 1988) with the tags Question-Answer Pair (QAP), Question_Elaboration, and Question_Clarification (see Figure 2).

234	GWFS	anyone got wheat for a sheep?
235	inca	sorry, not me
236	CCG	[nope.]$_a$ [you seem to have lots of sheep!]$_b$
237	GWFS	yup baaa
238	dmm	I think I'd rather hang on to my wheat I'm afraid
239	GWFS	kk I'll take my chances then…

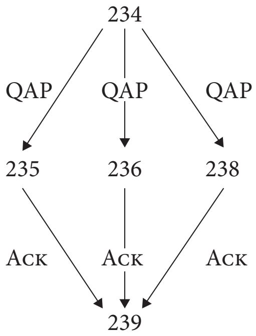


Figure 2. A sample of the STAC corpus (from Asher et al. 2019: 8)

In TED Talks, although the speeches are delivered by a single speaker with no participation of the audience, the QR pairs indicate where the speaker asks a question and immediately responds to it him/herself. This is a strategy that aims to capture and sustain the attention of the audience. In fact, this type of QR pairs is not unique to spoken discourse. They are also encountered in written texts, although less frequently, with the same purpose of capturing attention and creating dialogism. Grésillon and Lebrave's (1984: 126) comments about these interrogative constructions in written data are relevant here: "it is important to point out the potential of assertion that interrogation might carry. If this potential for assertion has been pointed out for rhetorical questions, we suggest it for all forms of written interrogation."

In TED talks, QR pairs are used with an appealing function oriented towards the audience (*subiectio*) (Lanham 1991; Mayoral 1994). In an attempt to capture their interactional and rhetorical function, in TED-MDB, they are labelled with a new top-level sense tag, Hypophora (Zeyrek et al. 2018). In the annotation, the wh-word or the auxiliary of polar questions is considered the anchor of Alternative Lexicalization, as illustrated in (7):

(7) Do **companies that take sustainability into account really do well financially**? *The answer that may surprise you is yes.* [AltLex] [Hypophora]

In Romance languages, polar questions may be expressed by intonation only in speech (and shown by the interrogation mark). These cases are annotated as implicit Hypophora relations, as demonstrated in the Portuguese case in (8), which is the translation of (7) above.

(8) **Estes casos são casos isolados?** … **As companhias que praticam a sustentabilidade estão mesmo bem financeiramente**? *A resposta pode surpreender-vos, mas é: "Estão, sim."* [Implicit=*será que* 'is it the case that'] [Hypophora]

The PDTB 3.0 is an extension of the PDTB 2.0 scheme, where the annotation of questions is also revised. QR sequences are now considered as a new coherence relation type called Hypophora involving the cases where "one argument (commonly Arg1) expresses a question, and the other argument (commonly Arg2) provides an answer" (Webber et al. 2019: 9).

As we hope to have shown so far, the nature of the QR sequences is related to the corpus data. For example, the STAC corpus is truly interactional. It involves no phatic or rhetorical questions, while the PDTB and the RST discourse banks, which are based on data from newspapers, can involve interactional material (as in the case of interviews). On the other hand, given that TED speeches are monologues and are not intended to elicit linguistic responses from the audience, any QR pairs uttered by the speaker would have a rhetorical effect.

Data analysis

The case of *well* and its translations to Portuguese and Turkish in TED-MDB

In TED-MDB, the discourse marker *well* is always used in a pragmatic function. An annotated discourse connective function does not exist in the corpus. For example, in (9), it functions as a structural marker or a "frame-marker," indicating topic change (Jucker 1997):

(9) … *the odds that it's not completely wrong are better than the odds that our house will burn down or we'll get in a car accident*. Well, **maybe not if you live in Boston**. [Implicit=except] [Expansion:Exception:Arg2-as-excpt]

The frequency of *well* in English and its equivalents in the Portuguese (PT) and Turkish (TR) subparts of TED-MDB was quantified for each talk. Table 4 provides the number of annotated (ann) and unannotated (unann) cases of *well*, and the number of cases where the equivalent discourse marker is kept or omitted in the Portuguese and Turkish translations.

Table 4. *Well* in the Portuguese and Turkish translations of TED-MDB

TED Talk ID	ENG		PT		TR	
	Ann	**Unann**	**Kept**	**Omitted**	**Kept**	**Omitted**
1927	0	5	2	3	0	5
1971	0	0	0	0	0	0
1976	0	5	3	2	1	4
1978	0	0	0	0	0	0
2009	0	0	0	0	0	0
2150	0	2	0	2	0	2
Total	**0**	**12**	**5**	**7**	**1**	**11**

The results point out that *well* is more frequently omitted (7 cases in 12 in PT, and 11 cases in 12 in Turkish) than kept in the translations. It is omitted more frequently in Turkish (11 cases, 91.6%) than in Portuguese (7 cases, 58.3%).

Example (10) illustrates a *well*-instance, which both Portuguese and Turkish translations eliminated. This could be attributed to the fact that the translators are focused on conveying the meaning of the sentences, and are less concerned with keeping the pragmatic function of the DM when transcribing spoken data in the written format. But we should note that in (10c), the Turkish translator is concerned with the interactional aspect of the speech and added *evet*, 'yes.' This can be viewed as a supplementation technique in the translation that helps the translator to capture the pragmatic/interactional aspect of the speech. In all three languages, the DM is left out of annotation (see Section 4.3 for a discussion of DMs in QR pairs).

(10) a. Why **did they do that**? Well, with apologies to the Home Improvement fans, *there's more growth in water than in power tools, and this company has their sights set on what they call "the new New World."*
[AltLex] [Hypophora]

b. Porque é que **fizeram isso**? [omitted=DM] Desculpem-me os fãs da Reforma de Habitações, mas *há um crescimento maior em água do que em ferramentas elétricas.* [AltLex] [Contingency:Cause:Reason; Hypophora]

c. Neden **bunu yaptılar**? [omitted=DM] Evet, *ev tamiri meraklılarından özür diliyorum ama su işinde elektrkli aletlerden daha fazla büyüme var ve bu şirket "yeni Yeni Dünya" adını verdikleri şeyi başarmak istiyorlardı.*
[Implicit=*çünkü* 'because'] [Contingency:Cause:Reason]

As already indicated, Turkish and Portuguese translations diverge in the strategy of keeping or omitting *well*. For instance, in (11), the DM is translated as *bem* in Portuguese (11b), but it is not translated literally to Turkish (11c). From the perspective of annotation choice, the DM is (correctly) left outside the annotated relation in both English and Portuguese:

(11) a. Are **investors, particularly institutional investors, engaged**? Well, *some are, and a few are really at the vanguard*. [AltLex] [Hypophora]

b. **Os investidores, em especial os investidores institucionais estão empenhados**? Bem, *alguns estão*.
[Implicit=*será que* ‘is it the case that’] [Hypophora]

c. **Yatırımcılar, özellikle kurumsal yatırımcılar bununla ilgili** mi? [omitted=DM] *Evet, bazıları öyle ve birkaçı da öncü konumda*.
[AltLex] [Hypophora]

In (11), the speaker asks a question and introduces the response via *well*, apparently to enhance the liveliness and immediacy of the question-response formula, using the pragmatic effect of the DM. This tactic is directly carried over to Portuguese through *bem*. The Turkish translator omits the equivalent DM, but again, she adds *evet* ‘yes’ to the QR pair, capturing the interactional feature of the text.

Example (12) illustrates the opposite choices taken by the Portuguese and Turkish translators: the DM *well* is omitted in Portuguese, while it is retained through an equivalent marker, *peki* ‘ok’ in the Turkish translation (12c):

(12) a. … *the odds that it’s not completely wrong are better than the odds that our house will burn down or we’ll get in a car accident*. Well, **maybe not if you live in Boston**.
[Implicit=except] [Expansion:Exception:Arg2-as-exception]

b. … [*as hipóteses de que não esteja completamente errado são melhores do que a hipótese de a nossa casa arder totalmente ou de termos um acidente de carro*$_{\text{SupArg1}}$]. [omitted=DM] *Isso*, se **não morarmos em Boston**.
[Explicit] [Contingency:Condition:Arg2-as-cond]

c. … *bu kabul edilmiş bilimsel mutabakata dayalı olduğu için tamamen yanlış olmaması ihtimali evimizin yanması veya araba kazası geçirmemiz ihtimalinden daha fazla*. Peki, **Boston’da yaşıyorsanız belki de değil.**
[Implicit=*ama* ‘but’] [Expansion:Exception:Arg2-as-exception]

To summarize, this section showed that the DM *well* always has a pragmatic function in TED-MDB, and tends to be omitted in the translations to Portuguese and Turkish. While both languages tend to omit it, Turkish omits it more frequently than Portuguese with a difference of 58.3–91.6%.

The case of *so* and its translations to Portuguese and Turkish in TED-MDB

It is a well-known fact that some lexical elements may be ambiguous between a propositional (semantic) and a pragmatic function and it is sometimes difficult to distinguish between these functions in context. For instance, the word *so* is multifunctional, with a propositional (discourse connective) function expressing a result relation, as well as a pragmatic function expressing the opening or the starting moment of a discourse unit. In TED-MDB, similar to other DMs, *so* should be annotated only when it conveys a propositional/semantic function. In (13a), for instance, *so* and its equivalents do not have a propositional meaning and the contexts are annotated as NoRel relations without considering *so*. The same procedure is used in all three languages.

(13) a. *But it does illustrate that environmental leadership is compatible with good returns.* **So if the returns are the same or better and the planet benefits, wouldn't this be the norm…?** [NoRel]

b. Mas *ilustra que a liderança em ambientalismo é compatível com bons retornos*. **Então, se o retorno é o mesmo ou melhor e o planeta beneficia, isto não devia ser a regra…?** [NoRel]

c. Ancak *bu, çevresel liderliğin iyi kazançla bağdaştığını gösteriyor*. **O zaman eğer kazanç aynı ya da daha iyi olursa ve gezegen bundan fayda görürse bunun norm olması gerekmez mi**? [NoRel]

The number of occurrences of *so* was compiled and differentiated as the instances that were kept or omitted in the Portuguese and Turkish translations (see Table 5). These numbers were paired with information on how many translations of *so* have been annotated and how many have been left out of the annotation. The annotated instances show that *so* and its equivalents were interpreted as discourse connectives that contribute to the meaning of the sentence. In contrast, the unannotated instances suggest that they were interpreted by the annotators as DMs with a pragmatic function.

Table 5. *So* in the Portuguese and Turkish translations of TED-MDB

Talk ID	**ENG**		**PT**			**TR**		
			Kept		**Omitted**	**Kept**		**Omitted**
	Ann	**Unann**	**Ann**	**Unann**		**Ann**	**Unann**	
1927	9	3	5	2	5	5	5	2
1971	0	2	1	0	1	0	1	1
1976	3	2	1	0	4	3	0	2
1978	1	1	1	0	1	0	1	1
2009	2	0	2	0	0	2	0	0
2150	5	2	2	1	4	3	2	2
Total	**20**	**10**	**15**		**15**	**22**		**8**
Total/Lang.	ENG = 30		PT = 30			TR = 30		

Table 5 shows that the number of occurrences of *so* that is omitted in the translations is equal to the number of translations in Portuguese (50%) and that only 8 cases are omitted in Turkish (26.6%). This is in contrast to the number of omitted occurrences of *well*, especially in the Turkish data. While 91.6% of occurrences of *well* are omitted, only 26.6% of the occurrences of *so* are omitted. This can be attributed to the pragmatic nature of *well*, which contrasts with the ambiguous nature of *so*, which can have either a semantic or a pragmatic function. Also, while there was a higher tendency of omission of *well* in Turkish than in Portuguese, the opposite pattern is found in the case of *so*: there are 15 cases of omission in the Portuguese translations and only 9 cases in the Turkish translations, suggesting that Turkish translators tended to interpret *so* as a connective rather than a DM with a pragmatic function.

Cases of so *omitted and kept in the translation*

Table 5 also shows that the number of contexts of *so* that is unannotated in Turkish is higher than in Portuguese. This situation seems to follow from the fact that more instances of *so* were omitted in the Portuguese translations (as we see in the numbers provided in Table 5 for Talk 1927) or the annotators tended to consider the translated *so* instances as not having a discourse connective function.

Example (14) illustrates a case of omission in both Portuguese and Turkish. As opposed to English and Portuguese, the Turkish annotator infers an implicit relation in (14c). In the absence of an explicit device, the annotator may indeed infer an implicit relation, as this example shows.[1]

(14) a. *Last summer, we did a really cool test out in California at Northrop Grumman.* **So those are four petals**. [NoRel]
b. *No último Verão, fizemos um teste mesmo fixe na Califórnia, na Northrop Grumman*. [omitted=DM] **São quatro pétalas**. [NoRel]
c. *Geçen yaz, Kaliforniya'da Northrop Grumman'da çok etkileyici bir sınama yaptık*. [omitted=DM] **Burada dört tane yaprak var**. [Implicit=*özellikle* 'in particular'] [Expansion:Level-of-Detail:Arg2-as-detail]

On the other hand, there is a total of 3 occurrences in Portuguese and 9 in Turkish, where an equivalent of *so* is kept in the translation but not annotated. In those cases, the annotators consider that *so* (and its equivalents in Portuguese and Turkish) do not contribute to the relation that holds between the two sentences. In many of such contexts, *so* may be interpreted as marking a topic change (there may even be

[1] In TED-MDB, topic shifts are labeled as NoRel, as in Example (13). This is an annotation decision, which we aim to revise in the future.

a paragraph mark in the transcripts between the two sentences), and there is no discourse relation marked (labeled as NoRel) (see (13a)–(c)).

In some contexts, languages diverge in the annotation of DMs. In one language, a discourse relation is inferred and annotated, while in the other language, either the DM is missing, or annotation is not provided. One such example where the annotators made different choices is provided in (15). In English and Portuguese, the annotator inferred and annotated an explicit discourse relation anchored to *so* and *então* conveying a Contingency relation (15a), (b). In Turkish, however, there is not an annotated token that corresponds to the *so*-relation though there is an equivalent DM *o zaman* 'so/then' (15c). Since a propositional meaning was not inferred from the text, the annotator decides not to annotate this relation.

(15) a. *There's tremendous opportunity in sustainability*. So **how are companies actually leveraging ESG to drive hard business results**?
[Explicit] [Contingency:Cause+SpeechAct:Result+SpeechAct]

b. *Existe uma tremenda oportunidade na sustentabilidade*. Então, **como é que as empresas estão a promover o ASG para alcançarem resultados sólidos**?
[Explicit] [Contingency:Cause+SpeechAct:Result+SpeechAct]

c. Sürdürülebilirlikte inanılmaz bir fırsat bulunur. O zaman şirketler görünür iş sonuçları almak üzere ÇSY'yi gerçekte nasıl kullanıyorlar?
[not annotated]

Cases of so *translated and annotated*

In this section, the equivalents of *so* translated to Portuguese and Turkish and annotated as propositional DMs (discourse connectives) are examined. In Portuguese, there are 12 annotated *so* instances, and in Turkish there are 13, which are typically labelled with the sense Contingency:Cause:Result, as in (16b), (c).

(16) a. *They believe that ESG has the potential to impact risks and returns*, so **incorporating it into the investment process is core**…
[Explicit] [Contingency:Cause:Result]

b. *Eles acreditam que o ASG tem o potencial de criar impacto em riscos e receitas*, assim, **incorporar o ASG no processo de investimento é fundamental**… [Explicit] [Contingency: Cause:Result]

c. *ÇSY'nin risk ve kazançları etkileme potansiyeli olduğuna inanıyorlar*. Bu yüzden **bunu yatırım sürecine dahil etmek onların vazifelerinin temeli, fon üyelerinin en çok yararı olacak şekilde davranmak onların vazifelerinin temeli**. [AltLex] [Contingency: Cause:Result]

However, the ambiguous nature of *so* leads to the frequent selection of the sense Result+speech act by the annotators (cf. (17a), (c)), as they feel this label captures both the connective value and the additional pragmatic value of *so* in certain contexts. The hesitation between Result and Result+SpeechAct extends to the label Result+Belief; for instance, the Portuguese context in (17b) is annotated as Result+Belief.

(17) a. *But 93 percent see ESG as the future, or as important to the future of their business.* So **the views of CEOs are clear**.
[Explicit] [Contingency:Cause+SpeechAct:Result+SpeechAct]

b. *Mas 93% consideram o ASG como o futuro, ou igualmente importante para o futuro dos seus negócios.* Assim, **a visão dos administradores é clara**.
[Explicit] [Contingency:Cause+Belief: Result+belief]

c. *... yüzde 93'ü ÇSY'yi gelecek olarak veya şirketlerinin geleceği için önemli olarak görüyor.* Gördüğünüz gibi ***genel müdürlerin bakışı belli***.
[AltLex] [Contingency:Cause+SpeechAct: Result+SpeechAct]

Cases of ommitted so *where the pragmatic function survives in the translation*
Certain cases of omission of *so* in the translation reveal more interesting aspects of the annotation process. In the English version of Example (18), *so* is annotated as a discourse connective but omitted in Portuguese and Turkish translations. Although *so* is not kept in the translations, the annotators in both languages identified a Result relation implicitly holding between the two sentences. The annotation restores the equivalents of *so* as implicit discourse connectives.

(18) a. *... only looking at race doesn't really contribute to our development of diversity.* So **if we're trying to use diversity as a way to tackle some of our more intractable problems....** [Explicit] [Contingency:Cause:Result]

b. *... olhar apenas para a etnia não contribui muito para o desenvolvimento da diversidade.* **Se tentarmos usar a diversidade como forma de resolver alguns dos problemas mais difíceis...**
[Implicit=*por conseguinte* 'as a result'] [Contingency:Cause:Result]

c. *Sadece ırka bakmak farklılıkları geliştirmemize gerçekte bir katkı sağlamaz.* **Eğer, farklılıkları bazı zor problemlerimizi çözmek için bir yöntem olarak kullanmayı deniyorsak farklıklara yeni bir tarzda bakmaya başlamamız gerekiyor**. [Implicit=*o halde* 'in that case'] [Contingency:Cause:Result]

In the English and Turkish versions of Example (19), *so (*and its equivalent, *yani*) is annotated as a discourse connective with a pragmatic value. In the Portuguese translation, however, *so* was eliminated. It is possible that the translator inferred the pragmatic function of the DM and found it redundant for the meaning of the text. Despite the removal of this marker, it is interesting to find that the discourse relation is still understood by the Portuguese annotator as having a pragmatic element, as the token retains the speech act label in the translated Portuguese version.

(19) a. *It means limiting future risk by minimizing harm to people and planet, and it means providing capital to users who deploy it towards productive and sustainable outcomes*. So **if sustainability matters financially today, and all signs indicate more tomorrow, is the private sector paying attention**?
[Explicit] [Contingency:Cause+SpeechAct:Result+SpeechAct]

b. *Isso significa providenciar capital aos utilizadores que o aplicam para a obtenção de resultados produtivos e sustentáveis*. [Omitted=DM] **Se a sustentabilidade hoje tem importância financeira.** [Implicit=*portanto* 'so']
[Contingency:Cause+SpeechAct:Result+SpeechAct]

c. *Bu, insanlara ve gezegene olan zararın minimize edilerek gelecekteki riskin sınırlanması anlamına gelir ve de üretken ve sürdürülebilir sonuçlar elde eden kullanıcılara sermaye sağlanması anlamına gelir*. Yani **eğer sürdürülebilirlik bugün finansal olarak önemli ise ve tüm belirtiler gelecekte daha önemli olacağını gösteriyorsa, özel sektör bu konuya dikkat ediyor mu**?
[Explicit] [Contingency:Condition+SpeechAct]

Discourse markers in question-response pairs

Many occurrences of *well* and *so* are found in Hypophora contexts, which is no surprise as Hypophora is a rhetoric figure related to dialogism, and *well* and *so* have a pragmatic value in many of the contexts that we have discussed above. In fact, 5 out of 7 cases of omission of *well* in Portuguese are found in contexts of Hypophora. *Well*, and its equivalents in Portuguese and Turkish, always occurs at the beginning of the second argument and introduces the response. Example (10), repeated below as (20), illustrates the utterance-initial position of *well* in English (20a) and *peki* 'ok' in Turkish (20c):

(20) a. Why **did they do that**? Well, with apologies to the Home Improvement fans, *there's more growth in water than in power tools, and this company has their sights set on what they call "the new New World."*
[AltLex] [Hypophora]

b. Porque é que **fizeram isso**? [omitted=DM] Desculpem-me os fãs da Reforma de Habitações, mas *há um crescimento maior em água do que em ferramentas elétricas.*
[AltLex] [Hypophora] [Contingency:Cause:Reason]

c. Neden **bunu yaptılar**? [omitted=DM] Evet, *ev tamiri meraklılarından özür diliyorum ama su işinde elektrkli aletlerden daha fazla büyüme var ve bu şirket "yeni Yeni Dünya" adını verdikleri şeyi başarmak istiyorlardı.*
[Implicit=*çünkü* 'because'] [Contingency:Cause:Reason]

In (20), *well* is used as a face-threat mitigator (Jucker 1997). Here, the face of the home improvement fans is threatened and *well* softens the potential threat introduced in the response. In fact, the speaker is obliged to apologise from the Home improvement fans at the beginning of the response, strengthening the mitigating function. In Turkish, the face-threat mitigating function of the utterance is retained by an equivalent marker *peki*, as well as the apology (*özür diliyorum)* at the start of the response. The Portuguese translation omits *well* but keeps the apology *desculpem-m*e as the mitigatory device.

Example (11) above presented another Hypophora context, where *well* is kept in Portuguese but rendered in Turkish, utilizing a different interactional marker. This example is repeated below as (21):

(21) a. Are **investors, particularly institutional investors, engaged**? Well, *some are, and a few are really at the vanguard.* [AltLex] [Hypophora]

b. **Os investidores, em especial os investidores institucionais estão empenhados**? Bem, *alguns estão.*
[Implicit=*será que* 'is it the case that'] [Hypophora]

c. **Yatırımcılar, özellikle kurumsal yatırımcılar bununla ilgili** mi? [omitted=DM] *Evet, bazıları öyle ve birkaçı da öncü konumda.*
[AltLex] [Hypophora]

In (22a), (b), *well* and its Portuguese translation, *bem*, introduce the response but the response does not confirm the question fully; it only approves it partially. Thus, in this instance, *well/bem* are used as a qualifier (Jucker 1997: 94), "indicating some problems on the content level of the current or the preceding utterance." In the Turkish translation, on the other hand, the word *evet* 'yes' chosen by the translator as a substitute for *well* conveys the interactional feature of the QR pair, and strengthens the rhetorical effect of the QR.

In English, *so* is found in 8 contexts of Hypophora and it is usually paired with *well* at the beginning of the response. In these instances, *so* sets the question and signals a summing up and/or a topic change. Example (22) presents this function of *so*:

(22) a. *So* if sustainability matters financially today, and all signs indicate more tomorrow, is the private sector paying attention? *Well*, the really cool thing is that most CEOs are.

b. [omitted=DM] Se a sustentabilidade hoje tem importância financeira, – e tudo indica que amanhã ainda será mais importante – será que o setor privado está a prestar atenção? Pelo menos, a maioria dos administradores está.

c. *Yani* eğer sürdürülebilirlik bugün finansal olarak önemli ise ve tüm belirtiler gelecekte daha önemli olacağını gösteriyorsa [omitted=DM] özel sektör bu konuya dikkat ediyor mu?

In (22), the pattern *so … well* is translated only partially to Turkish: *yani*, the equivalent of *so*, is used at the start of the question for summing up the preceding discourse, though the response lacks a DM that has an equivalent function to *well*. In Portuguese, *so* is omitted in the translation and the response starts with *pelo menos* 'at least,' which is not equivalent to *well*.

In sum, the structural pattern of Hypophora is closely linked to the pragmatic function of *so* and can be expressed as: *So, question? Well, answer*. Identifying this pattern is helpful for making decisions during the annotation stage and to distinguish between a marker with a propositional sense versus a pragmatic function.

Conclusion and future work

The present study focused on the multifunctional nature of two DMs found in the TED-MDB corpus. Given that TED-MDB uses the PDTB annotation scheme that targets written texts, specific issues arise when applying the PDTB scheme of annotation to TED talks. The present study targeted such contexts, and specifically examined the occurrence of *well* and *so* in English, as well as their translations in the Portuguese and Turkish transcripts.

The goal of constructing TED-MDB was to annotate discourse relations (explicitly or implicitly conveyed) – it mainly aimed to annotate discourse connectives with a semantic function. Although certain pragmatic features such as Belief and Speech Act are included in the PDTB 3.0 annotation scheme, the range of pragmatic or modal functions of DMs is not addressed. The present research puts DMs with a pragmatic function into focus. Methodologically, the DMs that appeared with a discourse relation were examined, and the omitted and unannotated instances were also taken into consideration to enable a fine-tuned analysis of DMs in translation.

At the beginning of the present research, it was hypothesized that DMs with a pragmatic/modal meaning would be more prone to implicitation (omission) than DMs with a propositional meaning. DMs with a propositional meaning were expected to be translated more frequently, and those with a non-propositional meaning were expected to be judged as non-essential and possibly omitted. The hypothesis is confirmed by the results of the DM *well*, especially in Turkish, where it is often omitted in the translation (91.6%). However, the translations of *so* showed that in Turkish, this marker tends to be kept in the translation rather than omitted (73.4%). When we analyze the results for the two DMs, we see that our initial hypothesis holds partially: translators tend to omit DMs that do not have a discourse connective use. Translators may differ in their decision to keep them (as in Portuguese *well*, which is omitted less frequently than in Turkish), or tend to keep a DM that is multifunctional (as in Turkish *so*).

Another hypothesis was that the translations of some languages were more prone to implicitation. Again, the results regarding *well* showed that Turkish tends to omit this DM more frequently than Portuguese (91.6% against 58.3%), but it behaves differently with the DM *so*. We can conclude that the translations of some languages are more prone to implicitation but only regarding DMs that do not have a discourse connective function.

In those contexts where *so* is translated but not annotated, we argued that the DM has pragmatic functions that the translator wishes to keep. In cases where *so* is translated and annotated, the DM turned out to encode a semantic relation with the value Result. The polyfunctionality of *so* creates some hesitation, and translations sometimes differ between its pragmatic function and discourse connective function in the two target languages. Furthermore, annotators sometimes assign Result+SpeechAct or Result+Belief tags to *so*. Thus, the pragmatic value of this DM could be captured along with its semantic function with the PDTB scheme.

The aim of the present paper was not to describe and distinguish between various functions of *well* and *so*. Nevertheless, their pragmatic functions were possible to spot, at least in specific contexts. For example, various functions of *well* mentioned by Jucker (1997) were found, in particular, its frame-marking, face-mitigating, and qualifier functions were observed and discussed. Various functions of *so* were also spotted during the analyses, such as its summing up or topic change function, as well as its pragmatic roles, which was possible to be captured by the Speech Act or Belief features of the PDTB scheme.

Many occurrences of *well* and *so* were found to be related to contexts where the speaker asks a question and immediately answers it, tagged as Hypophora in the TED-MDB scheme. These QR pairs frequently follow a pattern where a topic change is marked by the pragmatic marker *so* at the beginning of a question and the answer begins with the marker *well*. The use of pragmatic markers with Hypophora (which is itself a rhetorical device guiding the speaker to organize the speech) is well-aligned with the intended rhetorical effect of Hypophora.

It was also hypothesized that DMs with a pragmatic/modal meaning that are retained in the translation would be left out of the discourse relation annotation. We expected the annotators to identify these cases and distinguish them from discourse connectives, and even if they inferred a discourse relation, the DM would have to be left out of annotated spans. In most cases, our expectation is fulfilled. However, we observed that in annotating NoRel or EntRel cases, the annotators could not keep the pragmatic markers out, because according to TED-MDB guidelines, NoRels and EntRels are supposed to be annotated between two full sentences, and no part of the sentence should be left out (see Examples (13a)–(c)). This is an annotation decision and necessitates the revision of the guidelines in the future.

In future work, *well* and *so* as well as other DMs can be investigated with a more fine-grained approach to understand how precisely their various functions can be teased apart in relation to the discourse relations and QR pairs in TED-MDB. TED-MDB also offers opportunities to examine how the translation choices are related to the different functions of DMs, and how translation influences the decision to annotate the marker as part of the relation.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions.

Funding

The research has been partially supported by the FCT – Fundação para a Ciência e a Tecnologia (project UID/LIN/00214/2019) and by the Scientific and Technological Research Council of Turkey – BIDEB-2219 Postdoctoral Research program.

References

Aijmer, Karin, and Anne-Marie Simon-Vandenbergen. 2011. "Pragmatic markers." *Discursive pragmatics* 8: 223–247. https://doi.org/10.1075/hoph.8.13aij

Al-Saif, Amal, and Katja Markert. 2011. "Modelling Discourse Relations for Arabic." In *Proceedings of the Conference on Empirical Methods in Natural Language Proceedings of EMNLP '11*, Stroudsburg, PA, USA, 736–747. ACL.

Asher, Nicholas. 2012. *Reference to Abstract Objects in Discourse*. Volume 50. Springer Science & Business Media.

Asher, Nicholas, Julie Hunter, and Kate Thompson. 2019. "Comparing Discourse Structures between Purely Linguistic and Situated Messages in an Annotated Corpus." *Dialogue & Discourse* 11(1): 89–121. https://doi.org/10.5087/dad.2020.104

Asher, Nicolas, and Alex Lascarides. 1988. "The Semantics and Pragmatics of Presupposition." *Journal of Semantics* 15(2): 239–299. https://doi.org/10.1093/jos/15.3.239

Asher, Nicolas, Philippe Muller, Myriam Bras, Lydia Mai Ho-Dac, Farah Benamara, Stergos Afantenos, and Laure Vieu. 2017. "ANNODIS and Related Projects: Case Studies on the Annotation of Discourse Structure." In *Handbook of Linguistic Annotation*, ed. by Nancy Ide, and James Pustejovsky, 1241–1264. Springer. https://doi.org/10.1007/978-94-024-0881-2_47

Asher, Nicolas, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. "Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus." In *The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2721–2727. ELDA.

Brinton, Laurel. 1996. *Pragmatic markers in English. Grammaticalization and discourse function*. Mouton de Gruyter. https://doi.org/10.1515/9783110907582

Buysse, Lieven. 2015. 'Well it's not very ideal …' The pragmatic marker *well* in learner English. *Intercultural Pragmatics* 12(1): 59–89. https://doi.org/10.1515/ip-2015-0003

Carlson, Lynn, and Daniel Marcu. 2001. *Discourse Tagging Reference Manual*. Technical Report ISI-TR-545.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. "WIT3: Web Inventory of Transcribed and Translated Talks". In *Proceedings of EAMT*, Trento, Italy, 261–268.

Crible, Ludivine. 2017. Discourse Markers and (Dis)fluency across Registers: A Contrastive Usage-based Study in English and French. Ph.D. Dissertation. https://dial.uclouvain.be/pr/boreal/en/object/boreal%3A182932

Crible, Ludivine, and Liesbeth Degand. 2017. "Reliability vs. Granularity in Discourse Annotation: What is the Trade-off?" *Corpus Linguistics and Linguistic Theory* 15(1): 71–99. https://doi.org/10.1515/cllt-2016-0046

Crible, Ludivine, Ágnes Abuczki, Nijolė Burkšaitienė, Péter Furkó, Anna Nedoluzhko, Giedre Valunaite Oleskeviciene, Sigita Rackevičienė, and Šárka Zikánová. 2019. "Functions and Translations of Underspecified Discourse Markers in TED Talks: A Parallel Corpus Study on five Languages." *Journal of Pragmatics* 142: 139–155. https://doi.org/10.1016/j.pragma.2019.01.012

Cuenca, Maria Josep. 2008. "Pragmatic markers in contrast: The case of *well*." *Journal of Pragmatics* 40(8): 1373–1391. https://doi.org/10.1016/j.pragma.2008.02.013

Cuenca, Maria Josep, and Maria Josep Marín. 2009. "Co-occurrence of Discourse Markers in Catalan and Spanish Oral Narrative." *Journal of Pragmatics* 41: 899–914. https://doi.org/10.1016/j.pragma.2008.08.010

Grésillon, A., J.-L. Lebrave. 1984. "Qui interroge qui et pourquoi?". In *La langue au ras du texte*. Lille: Presses Universitaires de Lille, 57–132.

Hoek, Jet, Zufferey, Sandrine, Evers-Vermeul, Jacqueline, Sanders, Ted. 2017. "Cognitive complexity and the linguistic marking of coherence relations: a parallel corpus study." *Journal of Pragmatics* 121: 113–131. https://doi.org/10.1016/j.pragma.2017.10.010

Jucker, Andreas H. 1997. "The discourse marker well in the history of English." *English Language & Linguistics* 1, 91–110. https://doi.org/10.1017/S136067430000037X

Lanham, Richard. 1991. *A Handlist of Rhetorical Terms*. University of California Press, Berkeley. https://doi.org/10.1525/9780520912045

Lee, Alan, Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2016. "Annotating Discourse Relations with the PDTB Annotator." In *Proceedings of COLING* (Demos), 121–125.

Mayoral, José António. 1994. *Figuras Retóricas*. Madrid: Editorial Sintesis.

Oleskeviciene, Giedre V., Deniz Zeyrek, Viktorija Mazeikiene, and Murathan Kurfalı. 2018. "Observations on the Annotation of Discourse Relational Devices in TED Talk Transcripts in Lithuanian." In *Proceedings of the workshop on annotation in digital humanities co-located with ESSLLI*, vol. 2155, 53–58.

Oza, Umangi, Rashmi Prasad, Sudheer Kolachina, Dipti M. Sharma, and Aravind Joshi. 2009. "The Hindi Discourse Relation Bank." In *Proceedings of the third linguistic annotation workshop*, 158–161. ACL. https://doi.org/10.3115/1698381.1698410

Prasad, Rashmi, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. 2007. *The Penn Discourse Treebank 2.0 Annotation Manual*. https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. "The Penn Discourse Treebank 2.0." In *Proceedings of LREC*, 2961–2968. ELRA.

Schiffrin, Deborah. 2001. "Discourse markers: Language, meaning, and context." In Deborah Schiffrin, Deborah Tannen, and Heidi E. Hamilton (Eds.) *The handbook of discourse analysis* 1, 54–75. Blackwell Publishers.

Spooren, W., and Degand, L. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266. https://doi.org/10.1515/cllt.2010.009

Tonelli, Sara, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. "Annotation of Discourse Relations for Conversational Spoken Dialogs." In *Proceedings of LREC*, 2084–2090. ELRA.

Webber, Bonnie, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. *The Penn Discourse Treebank 3.0 Annotation Manual*. Technical Report, Institute for Research in Cognitive Science. University of Pennsylvania.

Zeyrek, Deniz, Işın Demirsahin, Ayışığı Sevdik Callı, and Ruket Çakıcı. 2013. "Turkish Discourse Bank: Porting a Discourse Annotation Style to a Morphologically Rich Language." *Dialogue and Discourse* 4(2): 174–184. https://doi.org/10.5087/dad.2013.208

Zeyrek, Deniz, Amália Mendes, and Murathan Kurfalı. 2018. "Multilingual Extension of PDTB-Style Annotation: The Case of TED Multilingual Discourse Bank." In *Proceedings of the 11th Language Resources and Evaluation Conference – LREC'2018*, 7–12 May 2018, Miyazaki, Japan, 1913–1919.

Zeyrek, Deniz, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. "TED Multilingual Discourse Bank (TED-MDB): A Parallel Corpus Annotated in the PDTB Style." *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-019-09445-9

Zhou, Yuping, and Nianwen Xue. 2015. "The Chinese Discourse Treebank: a Chinese Corpus Annotated with Discourse Relations." *Language Resources and Evaluation*, 49(2):397–431. https://doi.org/10.1007/s10579-014-9290-3

Variation of evidential values in discourse domains

A contrastive corpus-based study (English and Spanish)

Juana I. Marín-Arrese
Universidad Complutense de Madrid

This paper examines the variation of evidential values in oral conversation and written journalistic discourse in English and Spanish. The study focuses on indirect inferential and reportative values of evidentiality, and on the multifunctionality of a number of evidential expressions derived from the perceptual, conceptual and communicative experiential domains. The paper presents results of a contrastive corpus-based study on the expression of evidentiality, and on the multifunctionality of core evidential expressions (verbs and sentence adverbs). It is argued that variation in the use of particular values of evidentiality is sensitive to discourse domains and genres, and that multifunctionality would appear to favour expressions derived from the perceptual domain.

Keywords: evidentiality, multifunctionality, corpus annotation, comparable corpora

1. Introduction

Evidentials have been characterized as primarily indicating the source of information and the evidence on the basis of which the speaker feels entitled to make a claim (Anderson 1986; Aikhenvald 2004). A broader conception of evidentiality includes both the source of information and an estimation of its reliability, as Chafe (1986) posited in his seminal publication. From a cognitive-functional perspective, evidentiality is here conceived as a subdomain of the conceptual domain of epistemicity, which provides 'epistemic justification' for a proposition (Boye 2012).

Within evidential systems there is a basic distinction between direct (sensory) or indirect (inferential, reportative) modes of access to knowledge, the latter

https://doi.org/10.1075/btl.158.09mar

involving mediation by higher-level cognition in the case of inference or mediation through other individuals in the case of report (Langacker 2017). It has been argued that the different values of evidentiality are typically associated with different degrees of reliability of the source and mode of access to the evidence, and thus also of hearers' perception of degrees of speaker commitment to the proposition, and hearers' potential acceptance of the validity of the communicated information (Fitneva 2001; Papafragou et al. 2007; Marín-Arrese 2011b).

The values 'direct', 'indirect inferential', and 'indirect reportative' evidentiality are typically expressed by specific evidential expressions (Diewald and Smirnova 2010a). There are also a number of evidential expressions in English and Spanish, as in other languages, that are multifunctional, realizing two of the evidential values or even the three basic values identified within the domain of evidentiality (Boye 2012; Marín-Arrese 2017). This would seem to point to certain bidirectional connecting links between specific subspaces or notional regions, direct and indirect evidentiality, within the semantic map of epistemicity (cf. Boye 2012). This paper explores the link between the direct and indirect subspaces within the domain of evidentiality, as well as inter-subspace extensions, that is, on multifunctionality involving 'indirect inferential' (IIE) and 'indirect reportative' (IRE) values within the subspace of indirect evidentiality (Marín-Arrese 2017).

The aim of the study is to observe the degree to which the basic evidential values, IIE or IRE, are expressed preferably by evidential expressions derived from either the perceptual, conceptual or communicative experiential domains. The study also aims to explore the degree to which variation in the use of evidential expressions derived from the different experiential domains is sensitive to the variables of discourse domain, genre and language. The paper presents results of a contrastive corpus-based study on the expression of evidentiality, and on the multifunctionality of core evidential expressions (verbs and sentence adverbs), using comparable corpora in English and Spanish.

With these aims in mind, the paper addresses the following research questions:

i. Is there a difference in the degree to which evidentials derived from expressions involving the experiential domains[1] (ED) of perception, cognition and communication are present in the two discourse domains and genres (oral conversation vs. written press) (DD) in both languages?

[1] Domains of experience are understood as fields in which concepts are specified (see Langacker, 1987: 147–154, for a theory of conceptual domains), as for example the perceptual domain, the domain of cognitive or mental experiences, and the social domain of communication. As Geeraerts (2006: 27) also notes, "conceptualizations that are expressed in natural language have an experiential basis, i.e., they link up with the way in which human beings experience reality, both culturally and physiologically."

ii. Are there preferred frequencies of use for particular evidential values (EV), IIE vs. IRE, correlating with the experiential domains (ED) in the two discourse domains (DD) and across languages, English vs. Spanish?

In order to strive for the *tertio comparationis* (Chesterman 1998) at the semantic and notional level, the basic set of core evidential expressions (verbs and sentence adverbs) are examined for cross-linguistic correspondence across the two languages, English to Spanish and Spanish to English, using parallel corpora (Tiedemann 2012). The two resulting paradigms will form the potential cross-linguistic paradigm of corresponding evidential expressions (Altenberg and Granger 2002).

The data consists of naturally occurring examples from spoken and written corpora in the two languages: (i) Oral unscripted: (a) BNC-Baby (Oral subcorpus), (b) CORLEC (UAM); (ii) Written journalistic discourse: Corpus of English and Spanish Journalistic Discourse (CESJD-JMA). A basic level of *tertium comparationis* at the level of the corpora is similarly ensured, since the CESJD-JMA corpus was compiled applying comparison criteria for the selection of the texts on the basis of relevant similarity constraints or factors: E.g., mode, genre and text type, and subject matter or topic. In the oral unscripted corpora, however, factors such as register and degree of expertise of the speakers may differ and might thus affect the expression of evidentiality.

It will be argued that variation in the use of particular values of evidentiality is sensitive to the context of certain discourse domains and genres, and that the existence of particular evidential value-construction pairings would appear to indicate a process of entrenchment. Reference will also be made to certain parameters which may play a crucial role in facilitating these extensions of values within the domain of evidentiality, from inferential to reportative, resulting in the multifunctionality of particular evidential expressions (Marín-Arrese 2017).

This chapter is organized as follows: Section 2 discusses the main features and functions of the domain of evidentiality. Section 3 provides a description of methodological decisions and procedures. In Section 4, I present the results and discussion of the case studies. The conclusions are found in Section 5.

2. Evidentiality and multifunctionality

2.1 Evidentiality

Since the publication of the seminal work on evidentiality (Chafe and Nichols 1986), studies have for the most part centred on those systems of languages where the grammatical marking of the information source is obligatory (cf. Willett 1988; Frawley 1992; Aikhenvald 2004, *inter alia*). Aikhenvald (2004: 3) defines the notion in the following terms: “Evidentiality is a linguistic category whose primary meaning is source of information”. According to Aikhenvald (2004: 4), “All evidentiality does is supply the information source. The ways in which information is acquired – by seeing, hearing, or in any other way – is its core meaning”.

Recent years have witnessed a growing interest in the study of the domain of evidentiality in European languages, which rely on various strategies along the lexico-grammatical continuum (Squartini 2008; Diewald and Smirnova 2010b; Wiemer and Stathi 2010). Aikhenvald (2007: 222) has argued for a distinction between the term ‘evidentiality’, and evidential marker, to refer to the grammatical category of marking information source and the corresponding conceptual category ‘information source’ for other linguistic strategies. In this respect, Aikhenvald (2014: 3) notes that the expression of information source in languages lacking a grammatical evidential system is more versatile, and typically includes both “closed classes of particles and modal verbs, and a potentially open-ended array of verbs of opinion and belief”. Evidentiality in these languages has come to be conceived more broadly as including both grammatical and lexical means of indicating the source of evidence; thus the use of the more neutral and general term in this paper, ‘evidential expressions’ or ‘evidentials’ for short, to refer to any linguistic form along the lexico-gramatical continuum in English or Spanish whose function it is to convey either an indirect, inferential evidential meaning or a reportative evidential meaning.

Boye and Harder (2009: 14) have argued that evidentiality should be conceived as a “cognitive or functional substance phenomenon”, expressed by linguistic means that fulfil the function of indicating the source of information for the communicated content of a certain proposition, regardless of the grammatical vs. lexical status that the given linguistic device is ascribed to. This semantic-functional perspective is necessary when studying evidential meanings in languages which lack a specific system of grammatical evidentiality and do not possess fully grammaticalized evidentials. As Lampert and Lampert (2010: 319) argue, “the category of evidentiality is of use only, we conjecture, if a radical conceptual stance is taken in order to not miss capturing alternative linguistic strategies of expressing this notion”. However, in order to function as evidentials proper, these linguistic expressions need to be distinct, sufficiently conventionalized, holistic units with a meaning component that evokes or refers to source or access to information (cf. Cornillie et al. 2015). Apart from the basic condition that the meanings may be described in terms of the notion of ‘evidence’, Boye (2010: 304) observes that “for a given linguistic expression to be considered as having evidential meaning, it must be attested with a proposition-designating clause as its semantic scope”, that is, not a states-of-affairs-designating clause.

A broader conception of evidentiality, which includes both the function of indicating the source of evidence and providing 'epistemic justification' for the proposition (Boye 2012), also takes into account the discourse stance of the speaker and the potentially persuasive effects on the hearer. Evidentials would thus reflect the speaker's purported epistemic attitude towards the validity of the communicated information, which would relate to the degree of reliability typically assigned to the source and mode of access to the evidence (Marín-Arrese 2011a, 2013; Boye 2012). A related issue is the estimation of the hearer's epistemic vigilance filters and the degree to which they are likely to accept the information as valid (Sperber et al. 2010; Hart 2011). As Sperber et al. (2010: 359) have argued, humans "have a suite of cognitive mechanisms for epistemic vigilance, targeted at the risk of being misinformed by others". It is in the interest of speakers, and persuaders, to overcome these defences and strive for epistemic control in the discourse (Langacker 2013; Marín-Arrese 2013), and for hearer's 'epistemic trust' (Sperber et al. 2010).

2.2 Classification of evidential expressions

In the literature we find various subdivisions of the domain of evidentiality, which draw on classifications proposed by Willett (1988), Plungian (2001), De Haan (2001), Diewald and Smirnova (2010a), and Cornillie et al. (2015), among others. The most common subcategories found in most classifications are the following (cf. Marín-Arrese 2015, 2017):

i. Direct Perceptual Evidentiality (DPE): These expressions indicate direct, non-mediated, access to visual or other sensory evidence, which is external to the speaker/conceptualizer.

(1) When, on a hot day in London, **I see** <DPE> a woman wrapped in a black sack tagging along beside a guy in light T-shirt, jeans and sneakers, my first reaction is: "How bloody unfair!" (CESJD-EOG)[2]

ii. Indirect-Inferential Evidentiality (IIE): These expressions indicate personal indirect access to the information, through inferences triggered on the basis of perceptual or conceptual evidence, or inferences based on knowledge acquired through social communication sources (reports, documents, etc.).

[2] The reference system for the English texts of the *Comparable Corpus of English Spanish Journalistic Discourse* (CESJD-JMA) is the following:

ELG: English-Leading article-The Guardian; ELT: English-Leading article-The Times
EOG: English-Opinion column-The Guardian; EOT: English-Opinion column-The Times
ENG: English-News reports-The Guardian; ENT: English-News reports-The Times

(2) Instead of tinkering around the edges with the New Deal, community partnerships and affirmative action, they have finally embraced a bold initiative: water cannon and teargas. Twenty years to the month after Brixton, Handsworth and Toxteth went up in flames, it **appears** <IIE> as though nothing has been learnt. (CESJD-EOG)

iii. Indirect-Reportative Evidentiality (IRE): Expressions of indirect, mediated access to the information through social communication with some external source(s), that is, other conceptualizers.

(3) The US refusal to join the British-led stabilisation force and its apparently escalating proxy war with Iran around Herat in the west bodes ill for future security. So, too, does its policy of bolstering local chiefs and warlords in its quest for terrorists at the expense of central authority. Factionalism, banditry and crime are **reportedly** <IRE> on the rise in many parts of the country away from Kabul, especially in the Pashtun south. (CESJD-ELG)

In this paper, we restrict our focus to two of the values or subcategories of evidential expressions: (a) Indirect-Inferential evidence (IIE), and (b) Indirect-Reportative (IRE).

In most of these classifications, evidentiality is often viewed as a compact coherent 'functional-conceptual substance domain', with a number of subcategories; as we have seen, typically Direct, Indirect-Inferential and Reportative evidentiality (cf. Willett 1988; Diewald and Smirnova 2010a). However, Nuyts (2017: 58) has pointed out that "the notion of evidentiality is not a coherent semantic category", since "it covers dimensions of a quite different nature, which need to be kept apart and deserve a distinct status in a cognitively and functionally plausible semantic analysis". Under this view, the category 'inferential' would form part of a wider semantic class together with epistemic modality. Like epistemic modality, it is also a scalar category; though as Givon (1982: 42) has observed, "the match between the experiential scale of *evidentiality* and the scale of subjective (speaker's) *certainty* is not perfect". But whereas epistemic modality involves speaker-oriented assessments concerning the reality or veracity of the event and its likelihood, evidentials would seem to be hearer-oriented expressions reflecting the reliability of the mode of knowing and type of evidence, and indicating the source of the evidence (Marín-Arrese 2011b). Evidentiality involves degrees of "reliability of the process of inferencing in view of the strength or quality of the evidence available" (Nuyts 2017: 69). That is, certain inferential expressions (*must, obviously, it is evident*) seem to indicate a high degree of confidence of the speaker regarding the evidence and thus regarding the results of the inferential process from the evidence. In contrast, 'Direct' and 'Reportative' evidentiality are non-scalar. Direct evidential resources simply signal the mode of knowing and type of evidence (direct, sensory) and "differentiate between the different sense organs responsible for the experience (visual, acoustic, etc.)" (Nuyts 2017: 69). Reportative evidential resources are "'monolithic', one-valued" (Nuyts 2017: 69), indicating that the speaker has acquired or learned the information indirectly through communication with others.

The degree of confidence or commitment of the speaker with respect to the validity of the proposition, of the information communicated, is crucially linked to the 'reliability' attributed by default to the type of evidence and mode of knowing, "the process leading to the acquisition of the information (directly visual, indirectly through inferences, reports)" (Squartini 2008: 917): high in the case of direct visual perception, a gradation from strong to weak in inference, or in principle neutral in the case of report (Nuyts 2017). In this vein Fitneva (2001: 405) argues that "unlike speaker attitude, source of information provides hearers with an independent basis for assessing the reliability of the information and allows them to actively participate in finding out what is to be trusted". As Marín-Arrese (2015: 5) has observed, "On the basis of the source and the mode of access to the information, the different values of evidentiality are also typically assigned different degrees of reliability, which may be associated with distinctions in speakers' commitment to the communicated information, and in hearers' potential acceptance of the validity of the information".

An additional feature which differentiates Inferential from Direct and Reportative evidentiality, according to Nuyts (2017: 70), is the "'effort' they involve for the speaker" and the degree to which "the speaker can be said to be present in the meaning". This deictic character and the notion of strength is present in Frawley's (1992: 413) distinction between an internal Source of Knowledge (*Self*) and "the *scale of strength of knowledge*", to distinguish the scaled categories of "Inference" ("necessary> possible") and of "Sensation" ("visual> auditory> other senses …"), from an external Source of Knowledge (*Other*) and the "Scaled Categories of External Information" ("quote > report > hearsay> other"). This distinction between 'self' and 'other' hinges on the notion of egocentricity. Langacker (2017: 21) observes that "epistemic assessment involves degrees of centrality, with respect to several dimensions; each can be characterized in terms of immediacy – direct, unmediated access to an epistemic target – and increments of distance from that center". In the dimension 'source of information', evidentiality signals a basic distinction between direct, unmediated access to knowledge vs. indirect, mediated access. In sensory perception there is a direct connection between the speaker/conceptualizer and some external entity or stimulus. It involves "the speaker's own direct assessment, unmediated by the view of another conceptualizer" (Langacker 2017: 43). In the case of inference, the assessment is no longer direct, but mediated by processes of "higher-level cognition: thought, reasoning, generalization, inference, conceptual integration, and so on" (Langacker 2017: 21). Finally, in the case of report, the proposition is ascribed to another conceptualizer; there is mediation by "other conceptualizers, whether individually or in generalized fashion (e.g. as cultural knowledge)" (Langacker 2017: 21).

2.3 Multifunctionality

Boye (2012: 137ff) has drawn attention to a series of phenomena involving synchronic polyfunctionality and diachronic change in the meanings of evidential expressions, which point to a number of bidirectional connecting links between the semantic subspaces in the domain of evidentiality. One such case is the link between (i) direct perceptual evidentiality and indirect inferential evidentiality. Evidence for this link is the widespread phenomenon of extensions of meaning of expressions with a basic perceptual meaning, such as the verb '*see*' or '*ver*' in Spanish. In the following example, the evidential expression '*I see*' is signalling personal, direct, perceptual access to some external evidence.

(4) <hit n="14858" text="KBW">the one that we see We've seen the book yeah, ah, ah, she's seen the catalogue, ah she's got, on the, yeah I <kw>**see** <DPE> </kw> he's in the catalogue oh yeah</hit> (BNC-B)

The extension of perception verbs to indirect inferential evidentiality (IIE) is attested in a considerable number of languages, as in the example below, where '*see*' bears an indirect inferential value.

(5) <hit n="1831" text="KE4">I'll fill it up with water. Thank you. You want more, or have you got enough there? Enough there. I <kw>**see** <IIE> </kw>you like the tangerines. I'd like one. I have ta tangerines in our school. Mm.</hit> (BNC-B)

Within indirect evidentiality, we also find multifunctionality through the links between the semantic spaces of (ii) inferential evidentiality and reportative evidentiality. The most common cases are those of expressions with a basic perceptual meaning, such as '*appear, seem, apparently*', or '*parece, aparentemente, al parecer*' in Spanish. Most of these expressions are more frequently attested with an indirect-inferential meaning, though some instances are found with an indirect reportative meaning (Marín-Arrese 2017).

(6) <hit n="**157**" text="**KP5**">Picture's quite good until you get a play. Yeah. Can you get the telly down for me? Might have to. Oh is something in the way? <kw>**Appears** <IIE> </kw> **to** be. I don't know why Yeah. Ooh. Anyway Let's find the other side here. </hit> (BNC-B)

(7) <hit n="**4795**" text="**KCV**"> There are many in England there are quite a few you know No I know in Camden Town there are some very nice Yeah erm it's it's a very well-known supplier mm any it <kw>**appears** <IRE> </kw>that it a way a way to get in. I think To get a good place Hmm You have to give the doorman a ten pound note or something to get a nice place you know </hit> (BNC-B)

Less frequent are those evidentials derived from expressions with basic cognitive attitude meanings, such as '*supposedly*' or '*presumably*', which are found with both inferential (IIE) and reportative meaning (IRE), as in the following.

(8) Then in Iraq in 2003, confronted with a tyrant who had repeatedly thumbed his nose at the international system that Europe **supposedly** <IIE> revered, it instinctively recoiled, and a softened-up intellectual elite turned on the Americans instead. (CESJD-EOT)

(9) Last Friday the deal was agreed by UKFI, the new body **supposedly** <IRE> overseeing the taxpayers' stake in failed banks. Well, it would agree, wouldn't it? UKFI's board is packed with former directors of failed banks. (CESJD-EOG)

Much rarer are those cases of multifunctional expressions which bear witness to the cross-domain link between the semantic spaces of (iii) full epistemic support and inferential evidentiality, only marginally found in Spanish with the cognitive factive verbs '*conocer*' (know) or '*suponer*' (suppose), as in the following example.

(10) *<H2> <risas> Oye, **se conoce*** [se.IMP conoce.KNOW.3SG.PRS] *<IIE> que le encantan los <extranjero>"Holliwood"</extranjero> porque ya me lo encontré yo en otro con Inma, que fue cuando se me cayó todo el <extranjero>"Ketchup" </extranjero>.* (CORLEC-UAM)
(<H2> Hey, **apparently** [**it is known**] <IIE> s/he loves Hollywood (restaurants) because I already met him in another one with Inma, which was when I spilt all the Ketchup.)

The impersonal passive '*se conoce*' may also extend to a reportative value, as in the following example.

(11) *<H2> Tardó… tardó… tardó en bañarse cuatro días el Pablo, con esa <ininteligible> <risas>… Estuvo cuatro días sin ducharse, el guarro de él … <H1> Y **se conoce*** [se.IMP conoce.KNOW.3SG.PRS] *<IRE> que metió la mano en el bolsillo <silencio> y ahí estaban las gafas de la nena.* (CORLEC-UAM)
(<H2> He took … took … took four days to bathe that Pablo, with that … He did not take a shower for four days, the filthy guy … <H1> And **apparently** [**it is known**] <IRE> he put his hand into the pocket … and there he found the girl's glasses.)

Expressions with a basic communicative meaning (*say*, *tell*), typically used in speech representation, as in the example of indirect speech (ISR), derive evidential expressions with an indirect inferential value, in passive matrices of infinitival complement clauses such '*is said*', the impersonal '*they say*', or the impersonal passive '*se dice*' in Spanish, as in the following examples.

(12) US officials **say** <ISR> that food and other products that might carry the disease will be confiscated and destroyed. (CESJD-ENT)

(13) The shooting **is said** <IRE> to have especially angered Signor Berlusconi, who sent 3,000 Italian troops to Iraq despite widespread opposition in Italy. (CESJD-ENT)

The literature on evidentiality makes a distinction between reportative evidentiality and speech representation (indirect or reported speech), which may be viewed as a cline in terms of the dimensions of 'speaker perspective' and 'source realization'. As Chojnicka (2012: 179) has observed, "The original speaker's perspective is present to the largest extent in direct speech; in indirect speech, the current speaker attributes knowledge to another speaker from his/her own perspective. As the cline moves towards reportive evidentiality, the original speaker's perspective becomes gradually weaker and is finally lost. When it comes to source, in reported speech it is stated and linked to the reported information, whereas in evidentiality the source is not given."

Extensions involving the link between (iv) reportative evidentiality and inferential evidentiality are only found in the conditional form of the verb '*decir*' in Spanish, as in this example.

(14) *El Gobierno presenta en el Palacio Real de Nápoles su prometida vuelta de tuerca, un conjunto durísimo de medidas. En un 80%, se dedican a restringir la entrada, la libre circulación y los derechos de los ciudadanos extranjeros que residen en el país.* ***Se diría*** [se.IMP diría.SAY.3SG.COND] <*IIE*> *que la mayoría de las medidas, más que para garantizar la seguridad, han sido diseñadas para expulsar de forma inmediata a rumanos y gitanos.* (CESJD-SNP)
(… **It would seem** [**one would say**] that the majority of the measures, rather than serving to guarantee security, have been designed to expel the Rumanians and the gypsies immediately.)

As Marín-Arrese (2018: 101) has observed, "From a discourse-pragmatic perspective, the feature irrealis, presenting the situation as unrealized, contributes to distancing the speaker from the situation and diminishes speaker's responsibility for the communicated information". The modalising feature of conditionals (cf. Chilton 2014) may also have contributed to the extension of the meaning of this expression to the semantic space of inferential evidentiality.

3. Methodology

3.1 Hypotheses and research objectives

The paper posits the following hypotheses in relation to the research questions mentioned above.

i. With regard to the issue of whether evidentials derived from expressions of the experiential domains (ED) of perception, cognition and communication (the dependent variable) show a preference for a particular discourse domain and genre (unscripted oral conversation vs. written press) (DD) (the independent variable), the following null hypothesis is posited:

   $H_0$1 (Dependent variable: ED/Independent variable: DD): There is no difference between the proportion of evidential expressions derived from the three experiential domains (ED) of perception, cognition and communication present in each discourse domain (DD), oral unscripted conversation vs. written journalistic discourse.

   The null hypothesis is expressed as H_{01}: $\pi1 = \pi2 = \pi3$

ii. With respect to the preferred frequencies of use for particular evidential values (EV), IIE vs. IRE (the dependent variable), correlating with the experiential domains (ED) (the independent variable), in each of the discourse domains and genres (DD) and across languages, English vs. Spanish, the following null hypothesis is posited:

   $H_0$2 (Dependent variable: EV/Independent variable: ED): There is no difference between the proportion of expressions conveying the evidential values (EV), IIE vs. IRE, for each of the experiential domains (ED), in each discourse domain (DD).

   The null hypothesis is expressed as H_{02}: $\pi1 = \pi2$.

The following research objectives are set in relation to the above hypotheses:

a. Identification of the cross-linguistic paradigm of corresponding potential evidential expressions, through a two-stage process in the choice of data.
b. Identification of core evidential expressions derived from expressions of the experiential domains of perception, cognition and communication in our corpora.
c. Analysis of the data, identifying the values of evidentiality (inferential and reportative functions).
d. Quantification (using Monoconc), and comparison of the quantitative results in relation to discourse domains and genres, and languages.
e. Statistical analysis of the quantitative results across discourse domains (oral vs written discourse) and languages (English vs. Spanish).

3.2 Corpora and object of analysis: Evidential expressions

A basic set of core expressions (verbs and sentence adverbs) was selected from the literature on evidentiality in English and Spanish, and grouped according to the 'experiential domain', perceptual, cognitive, and communicative, of the basic meaning of the expressions typically undergoing semantic extension to evidential meanings:

a. Domain of perception:
 (v.) *appear, hear, look, give the impression, see, seem*; (adv.) *apparently, clearly, evidently, obviously, seemingly, visibly*.
 (v.) *dar la impresión, oir, parecer, ver*; (adv.) *aparentemente, al parecer, claramente, evidentemente, obviamente, por lo visto, visiblemente*.

b. Domain of cognition:
 (v.) *believe, reckon, suppose, think, understand*; (adv.) *presumably, supposedly*.
 (v.) *calcular, conocer, creer, entender, pensar, suponer*; (adv.) *supuestamente, presumiblemente, presuntamente*.

c. Domain of communication:
 (v.) *allege, report, say, tell*; (adv.) *allegedly, reportedly, purportedly*.
 (v.) *alegar, asegurar, decir, hablar, informar*.

For the basic meanings of the verbs and adverbs, the dictionary definitions provided by the *Oxford English Dictionary* (OED) and the *Diccionario de la Lengua Española, Real Academia Española* (DLE, RAE) have been used.

In order to strive for the *tertium comparationis* at the semantic or notional level, the basic set of core expressions (verbs and sentence adverbs) for the three experiential domains were examined for cross-linguistic correspondence across the two languages, English to Spanish and Spanish to English. The resulting paradigm formed the core cross-linguistic paradigm of corresponding potential evidential expressions in both languages (Chesterman 1998; Altenberg and Granger 2002).

Stage 1: The search for the core English expressions in the Global Voices Parallel Corpus 2017Q3[3] (English > Spanish) yielded the following translation equivalents (ordered according to frequency of occurrence of the English expression):

a. Domain of perception:
 (verbs) seem/s: *parece/n, pareciera, al parecer, sería*; look (like): *parece/n, se ve, parece ser, pinta*; appear/s: *aparece, parece/n, parecía, pareciera, al parecer, aparentemente, podría, resulta que*; (can) see: *ver, observar*; hear: *se oyen, puedo oír, he oído que, lo que oigo, puedo escuchar, escuchamos*.

[3] Corpus of news stories from the web site compiled and provided by CASMACAT (Multilingual version adjusted for OPUS) <http://casmacat.eu/corpus/global-voices.html>.

(adverbs) clearly: *claramente, obviamente, a todas luces, evidente, netamente*; apparently: *aparentemente, al parecer, según parece, parece, pareciera*; obviously: *obviamente, desde luego*; seemingly: *aparentemente, parece*; evidently: *veremos*; visibly: *visibles*.

b. Domain of cognition:

(verbs) I think: *creo, pienso, considero, me parece*; thought: *pensar, creer, parecen*; I believe: *creo, pienso*; believed: *se cree, se creía*; I suppose: *supongo*; supposed: *se supone, debe (debería, tendría)*; I would think; I reckon.

(adverbs) supposedly: *supuestamente, aparentemente*; presumably: *presumiblemente, presuntamente*.

c. Domain of communication:

(verbs) said: *se dice, según dicen, se habla*; reported: *se reportaron, se reportó, ha sido reportado, según se ha divulgado, se informó*; alleged: *se dice, supuestos*;

(adverbs) allegedly: *presuntamente, supuestamente, supuestos, al parecer, se presume, presuntos, según X*; reportedly: *se informa, según se informa, se dice, se cree, según informes*; purportedly: *supuestamente*.

The same search was carried out with the list of core expressions in Spanish from the three experiential domains in the Global Voices Parallel Corpus 2017Q3 (Spanish > English) to obtain a list of the most frequent cases with their correspondences in English.

Stage 2: The search for the above list of the most frequent expressions from the core potential evidential expressions in both languages in our oral and written corpora, yielded the following list of attested expressions used with an evidential meaning in the corpora:

a. Domain of perception:

English (10): (v.) *appear* (it appears that, appear/s to), *look* (looks like, looks as if), *give the impression, hear* (I hear, you hear), *see* (I/you/one can see, can be seen), *seem* (it seems that, seem/s to); (adv.) *apparently, clearly, evidently, obviously*.

Spanish (10): (v.) *da la impresión, parecer* (parece/n, según parece, parece ser), *oír* (he oído, por lo que he oído), *ver* (se ve, se puede ver, veo, por lo que se ve); (adv.) *aparentemente, al parecer, claramente, evidentemente, obviamente, por lo visto*.

b. Domain of conception:

English (7): (v.) *believe* (is/are believed), *reckon* (I/we reckon), *suppose* (is/are supposed to), *think* (is/are thought, you would think, you would have thought), *understand* (is/are understood); (adv.) *presumably, supposedly*.

Spanish (9): (v.) *calcular* (calculo, calculamos, se calcula), *conocer* (se conoce, por lo que se conoce), *creer* (se cree), *entender* (se entiende, entiendo, entendemos), *pensar* (se piensa), *suponer* (se supone); (adv.) *presumiblemente, presuntamente, supuestamente*.

c. Domain of communication:

English (6): (v.) *allege* (is/are alleged), *report* (is/are reported), *say* (is/are said, they say, you could/would/'d say), *tell* (am/are told, can tell); (adv.) *allegedly, reportedly*.

Spanish (3): (v.) *alegar* (se alega, alegan), *asegurar* (se asegura), *decir* (se dice, dicen, se diría).

The data consisted of all the tokens for the evidential expressions listed in Stage 2 collected from the following spoken and written corpora in the two languages:

a. Oral:
- *BNC-Baby* (unscripted conversation) (1,000,000 words)
- *Corpus Oral de Referencia de la Lengua Española Contemporánea* (CORLEC, UAM) (unscripted conversation) (1,100,000 words)

b. Written:
- *Corpus of English and Spanish Journalistic Discourse* (CESJD-JMA) (comparable corpus of journalistic texts: opinion columns, leading articles, and news-reports)
- English (The Guardian, The Times) (426,574 words)
- Spanish (El País, ABC) (368,883 words)

Total number of tokens with evidential functions, corresponding to the list of evidential expressions, found in the corpora:

- *BNC-Baby* (oral subcorpus): 584 tokens
- CORLEC: 789 tokens
- CESJD (English): 508 tokens
- CESJD (Spanish): 264 tokens

The search in the spoken BNC baby subcorpus was carried out with Xaira, and the searches in CORLEC and in CESJD were carried out using Monoconc.

3.3 Criteria for the analysis of evidential expressions

The analytical procedure draws on well-established criteria in the literature on evidentiality (cf. Anderson 1986; Diewald and Smirnova 2010a; Boye 2012). The following criteria have been applied in order to accept the use of a linguistic expression as a bonafide expression of evidentiality in English or Spanish:

i. Information source: In order to function as evidentials proper, linguistic expressions need to be distinct, holistic units with a meaning component that indicates some type of information source (Cornillie et al. 2015).

(15) … they allow Americans to compare the candidates over a prolonged session and in conditions of considerable pressure. They **obviously** <IIE> do make a difference. John Kerry's campaign seemed to be moving backwards until his strong performance in the debates revived it. (CESJD-ELT)

ii. High degree of conventionalization and entrenchment: Evidential expressions derived from lexical verbs are characteristically found as highly conventionalized forms (*it seems*), or as a fossilized 3SG form of the present indicative (*parece*), or in set expressions (*por lo que se ve*) or constructions (*parece ser que*).

iii. Propositional scope: As Boye (2010, p. 304) notes, "for a given linguistic expression to be considered as having evidential meaning, it must be attested with a proposition-designating clause as its semantic scope". We have accepted only those evidential expressions which scope over proposition-designating clauses. Cases like the following have been omitted from the data.

(16) *Higher education is not, in economists' jargon, a public good. It is a hybrid good, in which the benefit accrues most **obviously** to the student. (CESJD-EOT)

iv. Subjectification and pragmaticalization: There are cases where the evidential expression has undergone subjectification and has extended to discourse-pragmatic uses, thereby bleaching its evidential function. This is quite frequent in the case of adverbs. These cases are therefore omitted from the data.

(17) ?He said: "I am **obviously** very concerned about this and I think the responsibility of the BBC should not go unmentioned. (CESJD-ENT)

3.4 Procedure of annotation

The evidential expressions found in the oral and written corpora in our study were classified according to the two evidential functions focused on in this study.

a. IIE: expressions indicating speaker's inferences based on external sensory evidence or reasoning processes derived from general world knowledge, or personal assumptions.

b. IRE: expressions indicating that the information originates in some external voice(s), whose original perspective is defocused or lost.

The annotation procedure was carried out as follows:

i. The evidential expressions were classified and manually annotated according to: values of evidentiality, IIE or IRE (inferential and reportative functions).
ii. Classification of IIE or IRE evidentials derived from expressions in the experiential domains (ED) of perception, cognition and communication.
iii. Frequencies were counted and compared across discourse domains (oral vs written discourse) and languages (English vs. Spanish), using Monoconc;
iv. The distributions were analysed statistically to measure:
    a. Frequency comparison of evidential expressions found in the oral vs. the written texts: SIGIL online utilities for statistical inference (Corpus Frequency Test Wizard).
    b. Relative frequencies of evidential expressions, indicating either overuse or underuse of evidentials for each experiential domain in the oral vs. the written texts: UCREL's Log likelihood wizard.
    c. Frequencies indicating a possible association or correlation between the presence of expressions from three experiential domains (perceptual, cognitive, communicative) and the presence of evidential values (IIE, IRE) in each of the two discourse domains (oral vs. written): Chi-square[4] test.

4. Results and discussion

4.1 Results for evidential values, experiential domains, and discourse domains and genres in English

The results in Table 1 show the figures for evidential values, indirect-inferential (IIE) and indirect-reportative (IRE), of the expressions derived from the three experiental domains, perceptual, cognitive and communicative. The totals show journalistic discourse (1,191 per million words) as the preferred site for the use of evidential expressions.

i. $H_0$1 (Dependent variable: ED/Independent variable: DD): The results disprove the null hypothesis. The differences are significant for English, with respect to expressions from the perceptual and communicative experiential

[4] The chi-square provides a method for testing the degree to which there is association between the variables in a contingency table. The chi-square test measures the divergence of the observed values from the expected values under the null hypothesis of no association. If the P-value is significant, it indicates that there is some association between the variables, so that the observed values are not due to random variation. The value of significance was established at $p < 0.05$.

domains (ED). There is a marked preference in journalistic discourse for the use of perceptual-based evidential expressions (825.2 pmw), in contrast to the use in oral unscripted conversation (342.0 pmw). Though less marked there is also a significant preference in journalistic discourse for expressions derived from the domain of communication (173.5 pmw), in comparison with oral discourse (45.0 pmw).

Table 1. Experiential domains (ED) and Evidential values (IIE vs. IRE) in Oral and Journalistic discourse (English) (raw numbers, percentages, and ratios per million words)

English	**Oral (BNC-B) (1,000,000 words) evidential values**			**Journalistic (CESJD) (426,574 words) evidential values**			**Log-likelihood experiential domain**
Exp. Domains	**N-IIE %**	**N-IRE %**	**Total pmw**	**N-IIE %**	**N-IRE %**	**Total pmw**	**X^2 df = 1**
Percept	241	101	342	315	37	352	−130.96LL
	70.5	29.5	342.0	89.5	10.5	825.2	$X^2 = 142.57770$
Concept	160	37	197	15	67	82	+0.03LL
	81.2	18.8	197.0	18.3	81.7	192.2	*$X^2 = 0.01468$
Comm	19	26	45	2	72	74	−52.82LL
	42.2	57.8	45.0	2.7	97.3	173.5	$X^2 = 57.64253$
Total	**420**	**164**	**584**	**332**	**176**	**508**	**−132.98LL**
	71.9	**28.1**	**584.0**	**65.4**	**34.6**	**1,191**	**$X^2 = 143.19197$**

The Log-likelihood test shows relative frequencies of evidential expressions in the oral vs. the written texts, indicating either overuse or underuse of evidentials for each experiential domain in O1 (Oral) relative to O2 (Written)(– indicates underuse in O1 relative to O2). In English, there is a marked overuse of evidentials in journalistic discourse for evidential expressions derived from both the perceptual domain and the domain of communication. The differences for comparative results from both experiential domains in the oral and written discourse are significant in English.

ii. $H_0$2 (Dependent variable: EV/Independent variable: ED): The results allow us to reject the null hypothesis, since we find significant differences between the proportion of evidential values (EV), IIE vs. IRE, for each of the experiential domains (ED), and in both oral and written discourse (DD) in English. Evidential expressions derived from the experiential domain of perception occur significantly more frequently with inferential values (IIE): 70.5% for oral discourse and 89.5% for written discourse. Expressions derived from the domain of communication show a marked preference for the reportative value in written discourse, 97.3%, and less so in oral discourse, 57.8%.

An interesting case is that of the expressions derived from the domain of cognition, which show an opposing tendency in either discourse domain. As might be expected, these evidential expressions favour the indirect inferential value, 81.2%, in oral discourse. However, in written journalistic discourse these evidential expressions are markedly more frequent with a reportative value, 81.7%. A closer look at the data reveals the frequent use of expressions with passive matrices of infinitival complement clauses, '*is thought to*', '*is believed to*', whereby the journalist is signalling that the information has been accessed via some external source.

A chi-square test was run to establish whether there is association between the three experiential domains (ED: perceptual, cognitive, communicative) and the presence of evidential values (IIE, IRE) for each of the two discourse domains (oral vs. written). The null hypothesis, $H_0$2, assumes that there is no association between the variables 'IIE' and 'IRE' and each experiential domain. The alternative hypothesis, Ha2, claims that variation is due to association between the variables. The results were significant for the two evidential values in relation to the three experiential domains in both the oral corpus (Oral: EV/ED ($X^2 = 28.4419$; p-value: < 0.00001, significant at $p < .05$, and the written corpus (Written: EV/ED ($X^2 = 299.0436$; p-value: < 0.00001, significant at $p < .05$).

4.2 Results for evidential values, experiential domains, and discourse domains and genres in Spanish

The results in Table 2 show the figures for evidential values (EV), indirect-inferential (IIE) and indirect-reportative (IRE), and experiential domains (ED: perceptual, cognitive and communicative) for Spanish. The total figures show a balance in the use of evidential expressions in both oral (717.3 per million words) and written (715.7 per million words) discourse.

i. $H_0$1 (ED/DD): The results only partially disprove the null hypothesis. The figures for perceptual-based evidential expressions are very similar in both oral (591.8 pmw) and written discourse (596.4 pmw). The variation is quite limited in the case of expressions from the domains of cognition and communication. Cognition-derived expressions are more frequent in journalistic discourse (67.77 pmw), whereas communication-derived expressions are the preferred choice in oral communication (70.91 pmw). The global distribution of evidential expressions in Spanish thus differs quite clearly from English. The differences, however, are not significant (*X^2), according to the Log-likelihood test.

Table 2. Experiential domains and evidential values (IIE vs. IRE) in oral and journalistic discourse (Spanish) (raw numbers, percentages, and ratios per million words)

SPANISH	**Oral (CORLEC) (1,100,000 words) evidential values**			**Journalistic (CESJD) (368,883 words) evidential values**			**Log-likelihood Experiential domain**
EXP. DOMAINS	**N-IIE %**	**N-IRE %**	**Total pmw**	**N-IIE %**	**N-IRE %**	**Total pmw**	**X^2 df = 1**
PERCEPT	451	200	651	198	22	220	−0.01
	69.3	30.7	591.8	90.0	10.0	596.4	*X^2 = 0.00357
CONCEPT	41	19	60	7	18	25	−0.81
	68.33	31.66	54.55	28.0	72.00	67.77	*X^2 = 0.62226
COMM	20	58	78	6	13	19	+1.66
	25.6	74.4	70.91	31.6	68.4	51.51	*X^2 = 1.29474
TOTAL	**512**	**277**	**789**	**211**	**53**	**264**	**+0.00**
	64.9	**35.1**	**717.3**	**79.9**	**20.1**	**715.7**	***X^2 = 0.00000**

ii. $H_0$2 (Dependent variable: EV/Independent variable: ED): The results allow us to reject the null hypothesis, since we find significant differences between the proportion of evidential values (EV), IIE vs. IRE, for each of the experiential domains (ED), in both oral and written discourse. In oral discourse, inferential values (IIE) occur significantly more frequently in the case of evidential expressions derived from both the experiential domain of perception (69.3%) and the domain of cognition (68.33%), whereas reportative values are more frequent for communication-derived expressions (74.4%). In written journalistic discourse, inferential values (IIE) are markedly more frequent in the case of expressions from the perceptual (90.0%) domain. This contrasts with the percentages for the reportative value (IRE) in both the domain of cognition (72.00%) and communication (68.4%).

The results for the chi-square test show the figures are significant for the two evidential values in relation to the three experiential domains in both the oral corpus (Oral: ED/EV (X^2 = 58.5562; p-value: < .00001. significant at $p < .05$), and the written corpus (Written: ED/EV (X^2 = 83.6039; p-value: < 0.00001, significant at $p < .05$).

4.3 Discussion: Comparison of Results in English and Spanish

There are certain similarities across languages in the pattern of frequency distribution of evidential expressions in the three experiential domains, with a clear preference for expressions derived from the perceptual domain. The results from the English corpora are congruent with what might be expected. The feature 'intersubjectivity' is at the core of perceptual evidence; these expressions presuppose "intersubjective, manifest evidence, observable by more persons than the speaker" (Sanders and Spooren 1996: 258). This feature, together with the KNOWING IS SEEING metaphor, which motivates and structures these meaning-shifts in verbs of visual perception (cf. Matlock 1989), make these perceptual domain expressions the ideal candidates for the extension to inferential evidentials. The higher ratio of expressions from the perceptual domain thus stems from the overall higher numbers in the language of evidential expressions involving extensions from this domain. In our core list of expressions we had 10 types for perceptual, 7 for cognition, and 6 for communication. However, even with modified ratios for the tokens of perceptual origin ([342.0+825.2]÷10 = 116.72), cognition origin ([197.0+192.2]÷7 = 55.6), or from communication ([45.0+173.5]÷6 = 36.41), the perceptual domain is more prolific. The same arguments apply to Spanish.

The results for $H_0 1$ allow us to partially claim that variation is due to association between the variables, experiential domains and discourse domains (Dependent variable: ED/Independent variable: DD) in English, in the case of expressions derived from the perceptual domain and that of communication. However, results for variation in Spanish are not significant.

The higher total presence of evidential expressions in journalistic discourse in English can be said to be motivated by the felt need of the journalist to provide evidence for their claims, and to distance themselves from the content of the communicated proposition by displacing the responsibility to some external source. Paradoxical in this respect are the figures for Spanish, with almost equal ratios for oral discourse (717.3 pmw) and written discourse (715.7 pmw). The explanation must be sought either in cultural differences, or in the relative degree of reliability and validity of CORLEC as a bona fide corpus of oral unscripted conversation, and whether it is basically a general-purpose oral corpus. A possible solution would involve 'cleaning' the CORLEC corpus to leave only the oral unscripted conversations.

As regards $H_0 2$, involving the variables evidential values and experiential domains (Dependent variable: EV/Independent variable: ED) in both languages, there appears to be a significant association between the three experiential domains and the meaning extensions to inferential and reportative evidential values, so we can safely assume that the alternative hypothesis is valid.

5. Conclusions

This paper has presented a contrastive corpus-based study on the patterns of frequency distribution in the use of evidential expressions in the discourse domains of oral and written journalistic discourse in English and Spanish. The study has explored the degree to which evidentials derived from expressions from the experiential domains (ED) of perception, cognition and communication are present in similar proportions in the two discourse domains and genres (oral conversation vs. written press) (DD) in both languages (Hypothesis 1). The study has also looked at whether is a difference between the proportion of expressions conveying the evidential values (EV), IIE vs. IRE, for each of the experiential domains (ED), in each discourse domain (DD) (Hypothesis 2).

Expressions derived from the perceptual domain are more frequent by all accounts, and predominantly extend to the indirect inferential value in both discourse domains and languages. It has been found that variation in the use of values of evidentiality, IIE vs. IRE, is sensitive to choice of discourse domains, oral vs. written. This study has contributed empirical corpus-based results which have shown that expressions from the perceptual domain are the natural and ideal candidates for the extension to inferential processes, whereby speakers/writers feel they are providing adequate 'epistemic justification' for the communicated proposition. This well-attested extension may be motivated by the feature 'intersubjectivity', which is at the core of perceptual-based evidence, and the metaphor KNOWING IS SEEING, which motivates and structures these meaning-shifts in visual perception verbs.

Expressions derived from the domain of communication show a marked preference for the reportative value, and to a lesser extent for cognition-based expressions. It may be surmised that intersubjectivity is also at play in the case of expressions derived from the conceptual and communication domains, since the construal of some 'other source', the 'original voice', which is typically defocused, and unspecified, generalized or virtual, might pave the way for the extension of, for example, expressions with passive matrices of infinitival complement clauses to the reportative value in journalistic discourse in both languages.

The main limitations of the study stem from the problems observed regarding the reliability and validity of the Spanish corpus as truly representative of unscripted spoken discourse in European Spanish.

Acknowledgements

This research has received support from the project *Stance and Subjectivity in Discourse: Towards an Integrated Framework for the Analysis of Epistemicity, Effectivity, Evaluation and Inter/Subjectivity from a Critical Discourse Perspective*. (STANCEDISC), Ref. PGC2018-095798-B-I00, funded by the Ministerio de Ciencia, Innovación y Universidades (Spanish Ministry of Science, Innovation and Universities).

Dictionaries

DLE (*Diccionario de la Lengua Española, Real Academia Española*, RAE): http://dle.rae.es/
OED (*Oxford English Dictionary, OUP*): http://www.oed.com/

References

Aikhenvald, Alexandra. 2004. *Evidentiality*. Oxford: Oxford University Press.

Aikhenvald, Alexandra. 2007. "Information source and evidentiality: What can we conclude?" *Rivista di Linguistica* 19 (1): 209–227.

Aikhenvald, Alexandra. 2014. "The grammar of knowledge: a cross-linguistic view of evidentials and the expression of information source". In *The Grammar of Knowledge: A Cross-Linguistic Typology*, ed. by Alexandra Aikhenvald, and Robert M. W. Dixon, 1–51. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198701316.003.0001

Altenberg, Bengt, and Sylviane Granger. 2002. "Recent trends in cross-linguistic lexical studies". In *Lexis in Contrast*, ed. by Bengt Altenberg, and Sylviane Granger, 3–48. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.7.04alt

Anderson, Lloyd B. 1986. "Evidentials, paths of change, and mental maps: Typologically regular asymmetries". In *Evidentiality: The Linguistic Coding of Epistemology*, ed. by Wallace Chafe, and Johanna Nichols, 273–312. Norwood, NJ: Ablex.

Boye, Kasper. 2010. "Evidence for what? Evidentiality and scope". *STUF* 63 (4): 290–307. https://doi.org/10.1524/stuf.2010.0023

Boye, Kasper. 2012. *Epistemic meaning: A crosslinguistic and functional-cognitive study*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110219036

Boye, Kasper, and Peter Harder. 2009. "Evidentiality. Linguistic categories and grammaticalization". *Functions of Language* 16: 9–43.

Chafe, Wallace. 1986. "Evidentiality in English conversation and academic writing". In *Evidentiality: The Linguistic Coding of Epistemology*, ed. by Wallace Chafe, and Johanna Nichols, 261–272. Norwood, NJ: Ablex.

Chafe, Wallace, Johanna Nichols (eds.). 1986. *Evidentiality: The Linguistic Coding of Epistemology*. Norwood, NJ: Ablex.

Chesterman, Andrew. 1998. *Contrastive Functional Analysis*. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.47

Chilton, Paul. 2014. *Language, Space and Mind: The Conceptual Geometry of Linguistic Meaning*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511845703

Chojnicka, Joanna. 2012. “Reportive evidentiality and reported speech: Is there a boundary? Evidence of the Latvian oblique”. In *Multiple Perspectives in Linguistic Research on Baltic Languages*, ed. by Aurelia Usonienė, Nicole Nau, and Ineta Dabašinskienė, 170–192. Newcastle upon Tyne: Cambridge Scholars Publishing.

Cornillie, Bert, Juana I. Marín-Arrese, and Björn Wiemer. 2015. “Evidentiality and the semantics-pragmatics interface”. *Belgian Journal of Linguistics* 29: 1–17. https://doi.org/10.1075/bjl.29.001int

Diewald, Gabriele, and Elena Smirnova. 2010a. *Evidentiality in German. Linguistic Realization and Regularities in Grammaticalization*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110241037

Diewald, Gabriele, and Elena Smirnova (eds.). 2010b. *Linguistic Realization of Evidentiality in European Languages*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110223972

Fitneva, Stanka. 2001. “Epistemic marking and reliability judgements: Evidence from Bulgarian”. *Journal of Pragmatics* 33: 401–420. https://doi.org/10.1016/S0378-2166(00)00010-2

Frawley, William. 1992. *Linguistic Semantics*. Hillsdale (New Jersey): Lawrence Erlbaum Associates.

Geeraerts, Dirk. 2006. “Introduction. A rough guide to Cognitive Linguistics”. In *Cognitive Linguistics: Basic Readings*, ed. by Dirk Geeraerts, 1–28. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110199901.1

Givon, Tom. 1982. “Evidentiality and epistemic space”. *Studies in Language* 6 (1): 23–49. https://doi.org/10.1075/sl.6.1.03giv

de Haan, Ferdinand. 2001. “The place of inference within the evidential system”. *International Journal of American Linguistics (IJAL)* 67(2): 193–219. https://doi.org/10.1086/466455

Hart, Chris. 2011. “Legitimising Assertions and the Logico-Rhetorical Module: Evidence and Epistemic Vigilance in Media Discourse on Immigration”. *Discourse Studies* 13 (6): 751–769. https://doi.org/10.1177/1461445611421360

Lampert, Guenther, and Martina Lampert. 2010. “Where does evidentiality reside? Notes on (alleged) limiting cases: *seem* and *be like*”. *STUF* 63 (4): 308–321. https://doi.org/10.1524/stuf.2010.0024

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Vol. I: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Langacker, Ronald W. 2013. “Modals: Striving for Control”. In *English Modality: Core, Periphery and Evidentiality*, ed. by Juana I. Marín-Arrese, Marta Carretero, Jorge Arús and Johan van der Auwera, 3–55. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110286328.3

Langacker, Ronald W. 2017. “Evidentiality in Cognitive Grammar”. In *Evidentiality Revisited: Cognitive Grammar, Functional and Discourse-Pragmatic Perspectives*, ed. by Juana I. Marín-Arrese, Gerda Hassler, and Marta Carretero, 13–55. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/pbns.271.02lan

Marín-Arrese, Juana I. 2011a. “Effective vs. Epistemic stance and Subjectivity in political discourse: Legitimising strategies and mystification of responsibility”. In *Critical Discourse Studies in Context and Cognition*, ed. by Chris Hart, 193–224. Amsterdam: Benjamins. https://doi.org/10.1075/dapsac.43.10mar

Marín-Arrese, Juana I. 2011b. “Epistemic Legitimising Strategies, Commitment and Accountability in Discourse”. *Discourse Studies* 13 (6): 789–797. https://doi.org/10.1177/1461445611421360c

Marín Arrese, Juana I. 2013. “Stancetaking and inter/subjectivity in the Iraq inquiry: Blair vs. Brown”. In *English Modality: Core, Periphery and Evidentiality*, ed. by Juana I. Marín-Arrese, Marta Carretero, Jorge Arús, and Johan van der Auwera, 411–445. (eds.). Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110286328.411

Marín Arrese, Juana I. 2015. "Epistemicity and stance: A cross-linguistic study of epistemic stance strategies in journalistic discourse in English and Spanish. A cross-linguistic perspective". *Discourse Studies* 17 (2): 210–225. https://doi.org/10.1177/1461445614564523

Marín-Arrese, Juana I. 2017. "Multifunctionality of evidential expressions in discourse domains and genres: Evidence from cross-linguistic case studies". In *Evidentiality Revisited: Cognitive Grammar, Functional and Discourse-Pragmatic Perspectives*, ed. by Juana I. Marín-Arrese, Gerda Hassler, and Marta Carretero, 195–223. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/pbns.271.09mar

Marín Arrese, Juana I. 2018. "Evidentiality and the TAM systems in English and Spanish: A cognitive and cross-linguistic perspective". In *Tense, Aspect, Modality, and Evidentiality. Crosslinguistic perspectives*, ed. by Dalila Ayoun, Agnès Celle, and Laure Lansari, 81–106. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.197.05mar

Matlock, Teenie. 1989. "Metaphor and the grammaticalization of evidentials". *Proceedings of the Fifteenth Annual Meeting of the Berkeley Linguistic Society* 15: 215–225. https://doi.org/10.3765/bls.v15i0.1751

Nuyts, Jan. 2017. "Evidentiality reconsidered". In *Evidentiality Revisited: Cognitive Grammar, Functional and Discourse-Pragmatic Perspectives*, ed. by Juana I. Marín-Arrese, Gerda Hassler, and Marta Carretero, 57–83. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/pbns.271.03nuy

Papafragou, Anna, Peggy Li, Youngon Choi, and Chung-hye Han. 2007. "Evidentiality in Language and Cognition". *Cognition* 103 (2): 253–299. https://doi.org/10.1016/j.cognition.2006.04.001

Plungian, Vladimir. 2001. "The place of evidentiality within the universal grammatical space". *Journal of Pragmatics* 33: 349–357. https://doi.org/10.1016/S0378-2166(00)00006-0

Sanders, José, and Wilbert Spooren. 1996. "Subjectivity and certainty in epistemic modality: A study of Dutch epistemic modifiers". *Cognitive Linguistics* 7 (3): 241–64. https://doi.org/10.1515/cogl.1996.7.3.241

Sperber, Dan, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deidre Wilson. 2010. "Epistemic vigilance". *Mind and Language* 25: 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Squartini, Mario. 2008. "Lexical vs. grammatical evidentiality in French and Italian". *Linguistics* 46 (5): 917–947. https://doi.org/10.1515/LING.2008.030

Tiedemann, Jörg. 2012. "Parallel Data, Tools and Interfaces in OPUS". In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12), ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2214–2218. Istanbul, Turkey: European Language Resources Association (ELRA).

Wiemer, Björn, and Katerina Stathi. 2010. "Introduction: The database of evidential markers in European languages. A bird's eye view of the conception of the database (the template and problems hidden beneath it)". In *Database on Evidentiality Markers in European Languages*, ed. by Björn Wiemer, and Katerina Stathi. *STUF-Language Typology and Universals* 63 (4): 275–285.

Willett, Thomas. 1988. "A cross-linguistic survey of the grammaticalization of evidentiality". *Studies in Language* 12: 51–97. https://doi.org/10.1075/sl.12.1.04wil

The translation for dubbing of Westerns in Spain

An exploratory corpus-based analysis

John D. Sanderson
Universidad de Alicante

The aim of this paper is to analyze the process by which the audiovisual translation of American films has contributed to develop sociolects composed of distinctive lexis in various languages (Spanish or Italian, for instance) linked to specific genres such as Westerns or science fiction. With the compilation of a parallel corpus of source texts in English and their translation into Spanish, a diachronic analysis would enable researchers to identify linguistic recurrences that have developed them in a target culture. The Western, the most distinctive American film genre and culturally alien in origin to other cultural contexts, is chosen to analyze how a sociolect was formed in Spain with lexical elements, many of which did not follow word usage rules, but that became a requirement for acceptability in its polysystem.

Keywords: audiovisual translation, dubbing, film genre, Westerns, anglicisms, censorship

1. Introduction

The popularity of American cinema worldwide has contributed, through the translation of its films for dubbing or subtitling, to the development of sociolects in languages such as Spanish (Agost 1999; Naranjo 2015), Italian (Chiaro 2007; Keating 2014), French (Caron 2003; Armstrong 2004) or German (Queen 2004; Adamou and Knox 2011), related to specific film genres or sub-genres which were in origin alien to the target cultural contexts that received these productions. By means of screen translation, lexical elements which, in some cases, eluded word usage rules in these contexts, were progressively assimilated by the target audience. They would eventually become a requirement of acceptability (Toury 1995) in their polysystem (Even-Zohar 1978/2000), a cultural context made up of

https://doi.org/10.1075/btl.158.10san

stratified interconnected elements which integrates translated texts, in this case audiovisual, as a particular textual typology with a relevant effect on the receiving cultural system. Linguistic expectations are triggered by means of the visualization of an easily identifiable iconographic code (Chaume 2004), and this effect is more remarkable in countries where local audiovisual fiction consumption is far lower than American imported productions. These linguistic constructs become, then, more familiar for their audience even if they differ from common everyday speech. In Spain, for instance, where dubbing [“The replacement of the original speech by a voice track which attempts to follow as closely as possible the timing, phrasing, and lip movement of the original dialogue” (Luyken et al. 1991: 311)] is the prevalent mode for audiovisual linguistic transfer, the latest figures published by its *Instituto de Cinematografía y Artes Audiovisuales* show that, in 2018, attendance in Spanish cinemas was 17.92 per cent for Spanish films in contrast with 64.05 per cent for American films.[1]

The aim of this chapter is, by means of a parallel corpus which consists of twenty American Westerns and their translation for dubbing in Spain, to analyze diachronically how a sociolect for this film genre was developed and, eventually, embedded in the target cultural context.

2. Theoretical framework

Even though the concept of polysystem was initially applied mainly to the development and reactions of a literary system concerning the inclusion of translated texts as a system in itself (Even-Zohar 1978/2000; Snell-Hornby 1988; Shuttleworth & Cowie 1997), it has expanded to audiovisual translation concerning corpus-based analysis which proves a determining linguistic influence in, for instance, Spanish and Italian cultures (Baños & Chaume 2009; Laviosa 2011; Pavesi 2016, 2019). The overwhelming predominance of the American film industry results in a predilection for translated audiovisual texts as opposed to original productions of these target cultures, and the network of lexical relations within this source-oriented subsystem affects their polysystem as a whole.

In the adoption of lexical elements which are different from the norms of usage, dubbed Westerns not only included new models of reality, but also new patterns of language. The diachronic dynamics of this subsystem could even produce turning points within a target context, since dubbed productions assume a central role, with an own exclusive linguistic repertoire, which can prevail over the minority of local

[1] http://www.culturaydeporte.gob.es/dam/jcr:a802c21d-63ce-42c0-9097-ef55e3059444/24-rec-espec-nacionalidad..pdf

audiovisual films. Baños & Chaume (2009) and Baños (2009) refer to a Spanish audiovisual polysystem in itself, with translated texts which even influence target context creative writing, demonstrated with a comparable corpora of American and Spanish sitcoms.

'Sociolect' is defined as "a variety or lect which is thought of as being related to its speakers' social background rather than geographical background" (Trudgill 2003: 122), and therefore is a debatable term when referring to the distinctive lexis of a film genre, which has its own coherent audiovisual syntax, of which language is part of its stable contents (Altman 1984). I am using this term, 'sociolect', when referring to the distinctive lexis specific to Westerns, since they have a factual historical background in the source text related to their situational contextualization in the Wild West. The 'speech community' (Spolsky 1998: 3) portrayed in these American films has its own verbal norms, which have eventually gone through a process of standardization by means of audiovisual translation in target cultural contexts.

In the Spanish context, an artificial equivalent was eventually created by the recurrent professional practice of dubbing translators of a considerable amount of imported audiovisual productions. They were under the strict supervision of Franco's dictatorship during the first three decades covered in the corpus used for this research, and worked with total freedom after the advent of democracy, but some of the lexical elements were so rooted by then that they have prevailed well into the following century. The social interaction portrayed on screen by archetypical characters of the genre was verbally routinized in the target context by means of repetitive translation. As a result, this distinctive lexis has been assimilated by an audience who do not use it in their everyday life, but whose verbal expectations are triggered by the powerful iconographic code of the genre, and make it a requirement of acceptability for the community of spectators.

Since the term 'sociolect' is generally related to social background, and the distinctive lexis diachronically developed by translation for dubbing was assimilated by the audience of the target context as a whole, regardless of their background, 'jargon' could be an alternative debatable term. According to Spolsky (1998: 34): "closed peer groups often develop their own forms of jargon to serve as markers of group membership and also to make their speech less intelligible to outsiders." This might be applied to specialized language related to the science-fiction genre, or to strongly contextualized African-American films, but terms such as, for instance, 'to scalp', 'redskin' or 'posse', so specific to Westerns and alien to other cultural contexts, are historically documented as adapted to the norms of usage of the period in the source language, so their consideration as 'jargon' would be more difficult to generalize, even for their transfer to other languages. Perhaps a sociolinguistic neologism should be looming on the horizon.

Regarding general translation, it has usually been considered a positive issue that the more comparable the source and target situational contexts, the greater the degree of equivalence (Jakobson 1959/2000) that can be established between them (Rabadán 1991). Within the audiovisual field, this could be considered the case for standard film genres such as comedy and melodrama, not iconographically distinctive in general terms; or, in more specific cultural contexts, the presentation in American films, for instance, of clearly identifiable Italian or Spanish contexts for their translation in those countries (Parini 2009; Sanderson 2010). And concerning audiovisual translation, Pavesi (2005) points out that, in any case, translation for dubbing tends to neutralize markers of sociolinguistic and cultural variation which belong to the source cultural context in order to ease the reception of the target texts, generalizing a tendency already present in other modes of linguistic transfer.

Highly relevant for the corpus analyzed in this chapter is the seminal distinction established by Venuti (1995) between domestication, which would fit in the parameters presented in the previous paragraph, and foreignization: "choosing a foreign text and developing a translation method along lines which are excluded by dominant cultural values in the target language" (1997: 242). And the fact that the sociolect for Westerns in Spain was developed, in its initial period, under the ruling of a military dictatorship would be yet another source of interference to add to the warmly welcomed foreign film predominance. The ideological manipulation imposed on these films by strict censorial constraint in the target cultural context, without any objections from the American production companies, resulted, not unexpectedly, in a linguistic hybrid of interferences, in contrast with the neutralization appraised above.

The term 'dubbese', coined by Myers in 1973, bears a negative connotation that refers to this type of hybridity spoken in dubbed productions which sounds contrived because of the influence of the source text language. But lexical elements that initially made the dialogue seem unnatural would eventually become the norm (Even-Zohar 1978/2000), since audiovisual audiences refer to their memory of previous viewings of other dubbed films, hence forming a restricted intertext. As a consequence, one of the distinctive features of dubbese-prone audiovisual sociolects is their self-referentiality, with no other background to refer to but the translational routines established in the target cultural context because of their recurrent use in previous dubbed texts. Therefore, the experience of viewing a translated audiovisual production linked to a specific genre would generate a semantic 'noise' (Jakobson 1959/2000) when the dialogues did not meet those preconceived expectations. So in countries such as Spain, where consumption of local cinema is far lower than that of dubbed foreign films, linguistic referents based on audiovisual translation will be far more relevant for audiences than those which derive from national productions.

Toury (1995: 208) warned us that: “in translations, linguistic forms and structures often occur which are rarely, or perhaps even never encountered in utterances originally composed in the target language”. Those linguistic forms would eventually become the norm.

Studies which focus on the analysis of the translation into Spanish of audiovisual source texts related to film genre range from Agost’s (1999) research on television series to Naranjo’s (2015) article titled ‘Translating Blackness in Spanish Dubbing’. Specifically on Westerns, in Cronin’s *Translation goes to the Movies* (2009) there is a chapter devoted to the use of Spanish in the genre, and there are as well articles by Camus-Camus (2015) concerning censorship in the dubbing of Westerns and Valdeón (2018) about native-American stereotypes and translation. The recurrent issue that pervades all this research is foregrounded by Baños (2014) in her work on dubbed fictional dialogue: the presence of source text-induced features that are not natural in the target language.

In a strongly visually contextualized genre such as the Western, it would seemingly be more difficult to bridge the cultural gaps without a linguistic adaptation that smoothened differences. However, as we shall see in the following pages, there were enough culture specific items shared by both cultures, U.S.A. and Spain, to theoretically bring them closer but, with the analysis of some examples of distinctive lexis extracted from the parallel corpus, we shall see that this was not the case.

3. Parallel corpus selection and data compilation process

The twenty films that make up this corpus of American Westerns (Table 1) have been selected in such a way that the sample can be representative in this initial stage. There are at least two Westerns per decade premiered in Spain from the 1940s to the 2010s, distributed between eleven films released during Franco’s dictatorship (1939–1975), when the foundations of the sociolect were established amid the constraint of censorial practices, and nine films released during the current democracy, mainly to observe as well if the change in the political regime had an effect on its composition. Bearing this issue in mind, they have been listed on the chronological basis of their commercial release in Spain, since the delay in the premiere of three of the first eleven films, temporarily banned by the Spanish Censorhip Board, is relevant to the diachronic evolution of the sociolect. The choice of the films is also based on the regular inclusion in the source texts of lexical elements whose translation resulted in the formation of a sociolect for Westerns in Spain. The source text scripts and transcripts, available online from different sources, have only needed a partial correction once compared with the actual films, whereas the Spanish dubbed texts required a more time-consuming dictated transcript and revision.

Table 1. Corpus of twenty American Westerns and their dubbed versions in Spain

Original title and year	Word count	Dubbed version and year of release	Word count
1 *Stagecoach* (1939)	6,551	*La diligencia* (1944)	6,913
2 *My Darling Clementine* (1946)	5,069	*Pasión de los fuertes* (1948)	4.987
3 *The return of Frank James* (1940)	9,263	*La venganza de Frank James* (1950)	9,210
4 *High Noon* (1952)	6,209	*Solo ante el peligro* (1953)	6,318
5 *3.10 Train to Yuma* (1957)	6,018	*El tren de las 3.10* (1958)	6,134
6 *Rio Bravo* (1959)	14,884	*Río Bravo*(1959)	16, 223
7 *The Left-Handed Gun* (1958)	6,984	*El zurdo* (1962)	6,221
8 *The Man who Shot Liberty Valance* (1962)	8,448	*El hombre que mató a Liberty Valance* (1962)	7,788
9 *The Wild Bunch* (1969)	5,579	*Grupo salvaje* (1969)	6,123
10 *Chisum* (1970)	8,884	*Chisum* (1970)	7,873
11 *The Ox-Bow Incident* (1943)	8,006	*Incidente en Ox-Bow* (1973)	6,831
12 *The Outlaw Josey Wales*(1976)	7,639	*El fuera de la ley* (1977)	8,150
13 *The Mountain Men* (1980)	5,845	*El valle de la furia* (1981)	6,121
14 *Pale Rider* (1985)	5,885	*El jinete pálido* (1985)	5,734
15 *Unforgiven* (1992)	8,843	*Sin perdón* (1992)	9,220
16 *Geronimo: An American Legend* (1993)	6,823	*Gerónimo: una leyenda* (1994)	6,572
17 *The Assassination of Jesse James by the Coward Robert Ford* (2007)	10,474	*El asesinato de Jesse James por el cobarde Robert Ford* (2007)	8,763
18 *Appaloosa* (2008)	7,458	*Appaloosa* (2008)	6,502
19 *True Grit* (2010)	11,323	*Valor de ley* (2011)	10.088
20 *Django Unchained* (2012)	12,979	*Django desencadenado* (2013)	14,533
Total number of words	**163,164**	**Total number of words**	**160,304**

The choice of films has been based, therefore, on the presence of various distinctive terms in either the source or target texts which would be routinely related to Westerns as a genre. Some of them are grouped in the analysis of the corpus performed below regarding topics such as specific firearms or figures of authority, but others are put together based on the translation strategies applied for the dubbing into Spanish of specific archetypes such as ‘Stranger’, ‘Outlaw’, or ‘Bounty hunter’, vehicles such as ‘Stagecoach’ or ‘Buckboard’ or verbs such as ‘To scalp’, which contributed to form this sociolect in Spain with lexis which would not be commonly found in other genres. And the interference of censorial constraints of the Spanish dictatorship is also relevant, since it determined the standardized translation of potentially subversive issues, from expletives such as ‘Son of a bitch’ to references to archetypical individuals or groups which could imply a political sub-text, at least in the distorted minds of censors, such as ‘Hangman’ or ‘Posse’.

All in all, the twenty films chosen to form the corpus contribute with a significant amount of distinctive lexical elements to the viability of a statistical analysis which supplies relevant results on the diachronic development of the sociolect canonically related to Westerns in the target cultural context.

I then created a parallel corpus made up of the original texts in English and the Spanish translations for dubbing using the software *SketchEngine* (Kilgarriff et al. 2004). Both versions have been aligned on a turn-by-turn basis with a specific example database entry for distinctive terms such as those mentioned above. Therefore, when a search word is input in English, the system outputs all the lines of dialogue in the corpus which include that word, and then displays the entire turn that relates to it in the Spanish versions. Thus, translation problem-solution pairs are identified because of this alignment structure, and resulting glossaries and statistical results can then be open to interpretation and discussion from different perspectives.

The corpus is regularly increased in stages of five films at a time, with subsequent choices based on the existence of the distinctive lexis above mentioned in source and target texts and a balance in the diachronic distribution through decades to also allow for the analysis of the influence or the political context in the development of the sociolect. The time-consuming process of transcription and digitalization has an influence on the volume of the compilation, which currently stands at over 320.000 words extracted from about 70 hours of film, adding both the American originals and the dubbed versions. With the increase in the number of films and statistical results, the aim of this research is that the processed data will become helpful for audiovisual translators so that they can reduce time-consuming searches and justify the reliability of their choices. It could also be a contribution for researchers in order to understand how a sociolect related to an alien cultural context is developed, and an encouragement to analyze other specific film genres in various languages. In any case, with the available corpus compiled so far and the conclusions reached, some methodological steps and analytical issues can already be established.

4. Origin

The Western has probably been the film genre that most distinctively engulfs American culture and values. As Cronin (2009: 30) points out: "it is hardly surprising that cinema in the United States would not want to tell this story just to American audiences but to anybody else through the marked popularity of the Western genre". But the claim made by William Hays, head of the Motion Picture Producers and Distributors of America, back in 1923: "We are going to sell America to the world with American motion pictures" (in Gomery 1980: 8), required an adjustment with the advent of sound a few years later: its film industry had to decide how to transfer linguistically their productions abroad in order to maintain the hegemonic position it had deservedly achieved during the silent film period.

In the case of the Spanish language (it also happened with French, German and Italian; see Ellis 1995), Hollywood initially chose to shoot alternate versions of their most commercially attractive productions with Spanish-speaking actors: "A film would be shot, say, in English during the day, while another crew and cast employed the same set during the night to shoot another version for some major foreign market" (Audissino 2014: 98). The first Western to experience this process of multilingual filming was *The Big Trail* (Raoul Walsh 1930), starring John Wayne, whose Spanish language version was shot simultaneously with Jorge Lewis, bilingual actor of Mexican origin, playing Wayne's part, the also Mexican actress Carmen Guerrero in the lead female role, and Spanish actors Martín Galaraga, Julio Villarreal and Carlos Villarías in other supporting roles. Directed by David Howard and Samuel Schneider, under the supervision of Walsh himself, and with the script re-written by another Spaniard, Francisco Moré de la Torre, it was released in 1931 as *La gran jornada* in Latin America and as *Horizontes nuevos* in Spain.

Due to poor results in the box-office, once the cost of large scale production was considered (Jarvinen 2012), the American film industry very soon realized that international audiences would rather watch their worldwide famous stars, as they had in the silent era, than target language speaking stand-ins. The famous comedy duo Stan Laurel and Oliver Hardy, for instance, were made to star in the alternate versions of some of their own short films themselves reading their lines in foreign languages from cue cards (Nornes 2007), but this process of multi-language versions also proved to be very costly and artistically inefficient (Gomery 1980).

In the case of Spain, as mentioned above, dubbing was promptly chosen as the linguistic transfer mode for foreign films. In hindsight, Pérez-González (2014: 53) points out: "Within each audiovisual market/nation, the configuration of audiences as culturally homogeneous constituencies has allowed the film and television industries to perpetuate the deployment of representational conventions that were forged over several decades within each geographical context". So, unknowingly at the time, target context translators for dubbing would start developing linguistic representational conventions which eventually composed sociolects that contributed to perpetuate the cultural hegemony of the American film industry in Spain.

"That dubbing has a deep relationship to nationalism and fascism at its formative moment is significant." (Nornes 2007: 191). But this generalized perception should take into account that the adoption of this linguistic transfer mode actually precedes European fascist regimes; in Spain, for instance, dubbing was already a common practice during the II Republic (1931–1936). Its technical process took place first in the dubbing facilities that Paramount opened in the studios Des Reservoirs in

Joinville-le-Pont, France (Jarvinen 2012), which catered for several countries and languages and, since 1933, facilities were available in Spain: Metro Goldwyn Mayer set up its dubbing studios in Barcelona, and Italian dubbing studios Fono-Roma opened a branch, Estudios Fono España, in Madrid (Willis 2002). Actually, Moré de la Torre, the Spanish adaptor of *The Big Trail*, returned to Spain in 1934 to become the artistic director of the latter (Heinink et al. 1992). Unfortunately, none of the Westerns dubbed into Spanish during that period are available, so the corpus put together for this chapter starts with the most popular film of the genre produced in the 1930s whose dubbing has survived, *Stagecoach/ La diligencia* (John Ford 1939, released in Spain in 1944), and ends with *Django Unchained/ Django desencadenado* (Quentin Tarantino 2012, released in Spain in 2013).

5. Corpus analysis

5.1 Introduction

The aim of this chapter is to analyze the diachronic development of the translation of distinctive lexical elements which have formed the sociolect for Westerns in Spain. The choice of items is based mainly on terminology specifically contextualized by the source cultural context in the Wild West, from archetypical figures ("Sheriff", "Hangman", "Bounty-Hunter", etc.) to vehicles ('Stagecoach' or 'Buckboard'), but also on lexis which, in spite of its more general application to other contexts (from "Stranger" to "Son of a bitch"), its routine translation for dubbing in Spain eventually located it within the sociolect of the genre.

This analysis foregrounds the presence of anglicisms as the most favored strategy of linguistic transference to the target context of these terms. Görlach's (2001: 1) definition: "a word or idiom that is recognizably English in its form (spelling, pronunciation, morphology, or at least one of the three), but is accepted as an item in the vocabulary of the receptor language", is applied here, and a diachronic research allows for the establishment of a pattern of evolution in their acceptance, since some of them did not exist before the advent of sound in film. Statistical results, both in the number of occurrences and of specific periods in which one term is outstandingly chosen, may also provide an evaluation on the interference of censorship in linguistic choices.

The first eleven films of the corpus displayed above were released in Spain during the military dictatorship, the period where a disruption of chronology can be clearly appreciated between the American and Spanish releases. Both *Stagecoach* and *The Return of Frank James* were premiered in the U.S.A during World War II, so the delay in their release in Spain is understandable because of political instability. However, the 30-year gap between *The Ox-Bow Incident*'s premiere in the U.S.A. and its release in Spain as *Incidente en Ox-Bow* had more to do with censorial practices.

In her research on their incidence and effects during the dictatorship in the translation into Spanish of Westerns in both literary and filmic texts, Camus-Camus (2015) foregrounds the ideological constraints concerning the target cultural context, and even though her article focuses on *Duel in the Sun*/*Duelo al sol* (1946, released in Spain in 1953), a parallel and even more blatant example is provided with *The Ox-Bow Incident*. Every imported and national production had to receive the official permission for distribution from the Spanish Censorship Board during Franco's dictatorship, and this film, based on a novel of the same title by Walter Van Tilburg Clark (1940), "written in the late Thirties as a response to the rise of Fascism in Europe" (Calhoun 2004: 55), was unlikely to have a warm welcome from the Spanish authorities. In the film, three cowboys wrongly accused of murdering a rancher are caught by a posse and hanged without a trial in spite of their protestations of innocence; shortly afterwards, news will arrive that exonerates them from the crime. In a regime where random executions were commonplace, a film that mildly denounced them did not go down well with the powers-that-be, thus the delay in its release until the dictatorship itself was in its death throes.

As Merkle (2010: 19) points out concerning censorship: "selection mechanisms intervene to block the entry of those cultural products deemed undesirable or, when entry is allowed, to influence the form of cultural transfer". We shall see how, besides the delays, the established sociolect for the genre also contributed at the time to diffuse any controversial issue in this and other films. And Spain is now a democratic country, but the dubbed version of *The Ox-Bow Incident*/*Incidente en Ox-Bow* still has the Spanish soundtrack produced in the previous era. Therefore, for the purpose of a diachronic analysis, the relevant dates are those when the dubbed versions were released in Spain rather than the year they were premiered in the U.S.A., since the information collected and analyzed deals with when and which specific lexical elements established themselves in the target context.

The first film in the corpus is the most famous Western of the 1930s, *Stagecoach*/*La diligencia*, whose dubbing was produced at CEA Studios in Madrid for its release in 1944. Four minutes into the film, we can find the following dialogue:

(1) – **Marshall**, I'm looking for my **shotgun guard**. Is he here?
– Out with the **posse**.
– ***Comisario****, estoy buscando a mi* ***escopetero****. ¿Está aquí?*
– *Salió con los* ***rurales****.* (00.04.04)

If we analyze the translation of the three distinctive terms emphasized in this excerpt, we can first observe that, if 'Marshall' was considered a federal law enforcement authority, '*Comisario*' (used in the dubbed version of this film the seven times the word was uttered in the source text) was the most suitable translation, since in Spain it is also a figure of national authority with a similar status. And as '*Escopeta*' is the full equivalent in Spanish for 'Shotgun', and '*Escopetero*' is the etymological derivation, according to the dictionary of the *Real Academia Española* (*Diccionario de la Real Academia Española*; access online), to refer to the man who carries it, it would also stand as the closest translation for 'Shotgun guard'. 'Posse', however, was a more difficult linguistic issue, since it does not have a full equivalent term in Spanish to refer to: "In the past, a group of men in the U.S. who were brought together to catch a criminal" (Cambridge Dictionary; access online), even though this practice did exist, though unofficially, in Spain. The chosen translation, '*Rurales*', refers to guards that were in charge of protecting countryside property from theft or arson, a traditional figure of authority which dates back to mid- 18th century. Even though it was not an equivalent to the term of the source text, its shared etymological origin from Latin, '*rur-*' (Countryside), contributed to establish in Spanish a coherence with the iconographic code of the film.

We can therefore verify in this fragment how the conceptual similarities between source and target cultures are reflected in the lexical choices for the translation of these terms, in accordance with the neutralizing procedure enhanced above by Rabadán and Pavesi. This does not mean that these traditional Spanish words did not coexist with anglicisms in *Stagecoach/La diligencia*, as we can see in the translation for dubbing of the following dialogue:

(2) – How are you, **marshall**? Get my man through all right?
– I don't need them [handcuffs].
– If you don't want to lose your prisoner, **sheriff**, you'd better take care of him yourself!
– *Hola, **comisario**. Ya veo que me lo ha traído.*
– *No hacen falta esposas.*
– *Si no quiere quedarse sin su preso, **sheriff**, más vale que se lo lleve.* (01.23.30)

The latter emphasized term could have been translated with another Spanish equivalent, '*Alguacil*', an officer of local authority, but '*Sheriff*' was already part of the Spanish linguistic landscape before the advent of sound in film, even though this anglicism includes a phonetic borrowing, the fricative postalveolar consonant /ʃ/, which is non-existent in the Spanish language. The acclaimed writer Benito Pérez Galdós had used it in his novel *Aita Tettauen* (1905), and even earlier, it had been included by Francisco J. García Rodrigo in his *Historia verdadera de la Inquisición. Tomo 3* (1877), referring to a British context. Therefore, it was already a canonical term, so established, in fact, that its multiple presence in twelve of the twenty American films of the corpus has always been transferred with the same word.

The same cannot be said of the three traditional Spanish terms mentioned above, which practically disappeared from dubbed Westerns, being replaced by more unusual, at least initially, lexical items. As we shall see in the following pages, contrary to the assumed belief that finding a common ground for culture specific items would benefit the assimilation of an alien genre in a new context, the generalized tendency seemed to be the erasure of these shared issues, developing, in consequence, a sociolect that would flaunt its differences in the target context of reception.

In order to organize a detailed analysis, I will divide the distinctive lexis that forms the sociolect specific to Westerns into four groups: two devoted to lexical fields that stand out within the genre (figures of authority and firearms); one that, from a translatological perspective, will analyze different linguistic strategies involved in the development of the sociolect, and a final one concerning the influence of censorial coercion during the dictatorship and beyond its official demise.

5.2 Figures of authority

'*Comisario*' as a translation for 'Marshall' vanishes in the following four films of the corpus, being mainly replaced by '*Sheriff*', a first hint at the upcoming prioritized foreignization of the sociolect for Westerns. Even when 'Marshall' and 'Sheriff' coexist in the source text, as was the case in *3.10 Train to Yuma/El tren de las 3.10* (1957/1958), they were translated in both cases as '*Sheriff*'.

(3) – Hello, you the **marshall**?
– *¡Eh, oiga! ¿Es usted el* ***sheriff****?* (00.23.42)

(4) – Get the **sheriff**. Tell him to get as many **deputies** as he can.
– *Avise al* ***sheriff****. Dígale que reúna a todos los* ***hombres*** *que pueda.* (01.06.48)

There was an abrupt shift, however, in the following film in the corpus, *Rio Bravo/Río Bravo* (1959), where 'Marshall' was oddly translated as '*Representante*' (Representative) in the ten times the word was uttered, followed by a return to the normative use of 'Comisario' in *The Left Handed Gun/El zurdo* (1962) and *The Man who Shot Liberty Valance/El hombre que mató a Liberty Valance* (1962). In the 90s, since *Geronimo: An American Legend/Gerónimo: una leyenda* (1993/1994), there has been a generalized comeback of '*Sheriff*' but, in a further twist towards foreignization, we can observe that, in the previous film of the corpus, *Unforgiven/Sin Perdón* (1992), the only time 'Marshall' is uttered, the same word is used in the dubbed version as a translation: (5) "You'd come and kill him like you killed the U.S. **Marshall** in the '70"; "*Le mataría igual que mató a un* ***marshal*** *de los Estados*

Unidos en el 70" (01.47.23). And what could have been taken as a random choice, or mistake, is empowered in the most recent film of the corpus, *Django Unchained/Django desencadenado* (2012/2013), where '*Marshal*' was used in the Spanish dubbed version the sixteen times it was uttered in the source text as, for instance, in the following: (6) "Remember, get the **sheriff**, not the **marshall**."; "*Recuerde, llame al* ***sheriff****, no al* ***marshal***" (00:14:55).

The presence of another anglicism, with the same phonetic loan as '*Sheriff*', but which had never been used in earlier translations. shows how foreignization, at least in this instance, has come full circle, and contemporary Spanish words such as '*Comisario*' or '*Alguacil*' would nowadays hardly stand a chance in dubbed Westerns. As Gottlieb (2012: 56) warned: "In the 21st century, we see more and more examples of anglicisms that creep in under the lexical carpet, thus resulting in altered semantics and syntax."

In the previous quotation from *3.10 Train to Yuma/El tren de las 3.10* (1957/1958), I have emphasized another relevant word concerning this subgroup, 'Deputy': "a person who is given the power to act instead of, or to help do the work of another person: (…) a sheriff's deputy" (Cambridge Dictionary). This term had a far more inconsistent translation for the Spanish dubbing of Westerns, as can be anticipated by the use of the hyperonym '*Hombres*' (Men) in this fragment. Throughout the first half of the corpus, up to eight different terms can be randomly found in the dubbed versions of six films. As well as "*Hombres*", the three terms previously mentioned ('*Comisario*', '*Alguacil*' and, expectedly, '*Sheriff*') were also used, together with '*Agentes*' (Agents), '*Voluntarios*' (Volunteers), '*Ayudante*' (Assistant) and '*Encargados*' (Men in charge). In *The Left-Handed Gun/El zurdo* (1958/1962), for instance, we can observe how the tendency towards foreignization comes into conflict with the unsteadiness of the translation of the term:

(5) – Act like a **deputy**.
– Act like a **deputy**?
– Where's your **gun**, **deputy**? Pick it up.
– *Eres un* ***comisario****.*
– *¿Un* ***comisario****?*
– *¿Dónde está su* ***arma, sheriff****? Cójala.* (00.56.02)

From *Chisum* (1970) onwards, however, "*Ayudante*" was established as the canonical term, and it has remained this way ever since, which would also prove that the stability of the sociolect was mainly grounded on choices made during the dictatorship. Now I shall take the other distinctive element of this final quote, "Gun", to present the lexical field it belongs to.

5.3 Firearms

'Shotgun': "a long gun that fires a large number of small metal bullets at one time" (Cambridge Dictionary) has, as mentioned above, its exact equivalent in '*Escopeta*', used in the first film of the corpus together with its etymological derivation '*Escopetero*', when it referred to 'shotgun guard', the five times it was uttered in the original *Stagecoach*/*La diligencia*. But in the other twenty cases in the corpus where 'Shotgun' refers to the firearm, it has still been translated five times into Spanish as '*Escopeta*', but eleven times as the anglicism '*Rifle*', which refers to a long firearm that only shoots one bullet at a time (for which there is a Spanish term, by the way: "*Fusil*") and another four times, for a short period of time, as '*Carabina*' (Carbine), which has a shorter barrel than a shotgun; they were two co-hyponyms of '*Escopeta*', but referring to two other different firearms. In *The Left-Handed Gun/ El zurdo* (1958/1962), for instance, both 'Shotgun' and 'Rifle' were translated into Spanish as '*Rifle*', a choice of a term that, as with '*Sheriff*', confirmed the impression of a propensity towards getting closer to the source cultural context:

(6) You will hold this **rifle**. It's not dangerous.
Coge este ***rifle****. No es peligroso.* (01:06:23)

(7) Well, they hit him with a **shotgun**. He's dead
Le han disparado con un ***rifle****. Está muerto.* (01.14.11)

On the whole, there has been an unjustified oscillation between '*Escopeta*' and '*Rifle*' without considering that the former term is the one which specifically refers to 'Shotgun', but the latter contributes to developing a source-oriented target context sociolect. The diachronic evolution of its translation into Spanish (Figure 1) confirms this tendency.

A more blatant shift towards foreignization can be observed with the translation of 'Pistol'. Even though it shares its etymology with its Spanish equivalent, '*Pistola*', this word has only been used as a translation for dubbing in two of the seven films of the corpus in which 'Pistol' is included; in the other five films, yet another obvious anglicism is used, '*Revólver*' [in Spanish with a stressed vowel to differentiate it from the already existent verb '*Revolver*' (To turn)], defined in English as "a type of small gun which is held in one hand and can be fired several times without putting more bullets into it" (Cambridge Dictionary). The distinctive feature of this weapon is that it contains its ammunition in a revolving cylinder; hence, its etymology; a 'Pistol' does not have one, so '*Revólver*' is not the most suitable translation. And, on top of that, in the whole corpus of twenty original American Westerns, the English word 'Revolver' has not been used a single time in the source texts.

In order to conclude this sub-section, I will comment on the most usual term in the corpus of American Westerns to refer to firearms, the hyperonym 'Gun', used

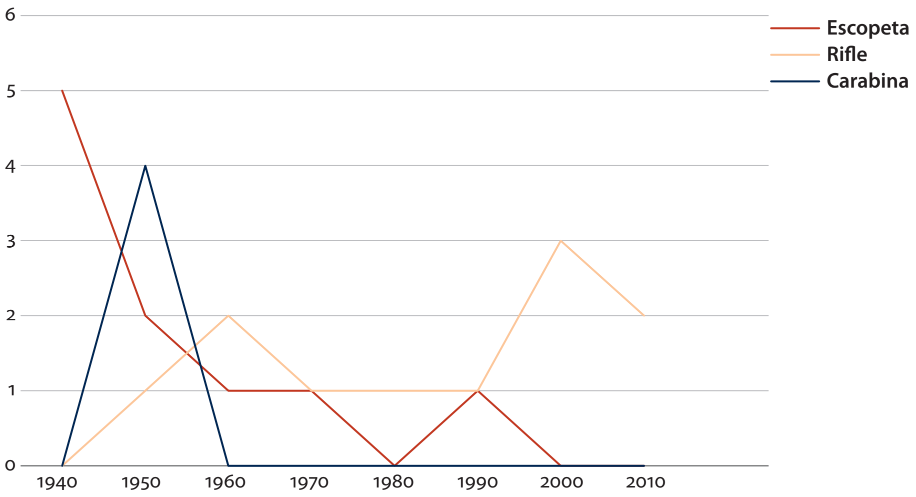


Figure 1. Diachronic evolution of the translation for dubbing into Spanish of 'Shotgun'

240 times throughout the twenty films. It is translated as '*Arma*' (short for '*Arma de fuego*', 'Firearm') in the final quote of the previous subsection, which would be the closest equivalent, and in 41 per cent of all the cases in the corpus. But in 43.7 per cent of the cases it has also been translated as the anglicised hyponym '*Revólver*'; the other terms used are '*Pistola*' (5.3 per cent), and, when referring to long guns, '*Rifle*' (9 per cent) and '*Fusil*' (0.53 per cent). Therefore, in the Spanish sociolect for Westerns, the hyperonym 'Gun' is mainly translated using another anglicism, further proof of a source-oriented film translation which erases any common links between cultural contexts in order to make the resulting sociolect far more distinctive.

5.4 Other translation strategies

Besides anglicisms, some recurrent terms established in the sociolect of dubbed Westerns are the result of other strategies that I would briefly like to point out: archaisms, structural calques and loan shifts.

5.4.1 *Archaisms*

Holmes' (1972: 102) conceptualization of the *cross-temporal factor*, related to "translating a text that not only was written in another language but derives from another time", has generally considered the use of archaic terms as a linguistic bridge between the past times portrayed in a text and the present of its translation. In the development of this sociolect linked to a genre that reenacts a previous era, some coincidences in established terms may well be worth noting. The most remarkable is the presence of the lexeme '*for*' ('Outside', whose etymological origin is from Latin, '*Foras/Foris*', as in 'Foreign') in the translation for dubbing into Spanish of 'Stranger' and 'Outlaw'. They have been translated, respectively, as '*Forastero*' (in six of the seven films in which it was uttered), and '*Forajido*' (in four out of five films), two archaisms that have become so embedded in the sociolect for Westerns that the use of a different word (in other film genres, '*Extranjero*' and '*Criminal*', for instance, are the most common terms) would produce a feared semantic noise, since it would contradict the expectations of the audience. These exceptional cases prove that anglicisms are not an essential condition to make a sociolect distinctive.

Interestingly, the latter English term is part of the title of a film of the corpus, *The Outlaw Josey Wales* (1976), which was translated with a non-normative multi-word unit, *El fuera de la ley*, a covert lexical borrowing which cannot be found at all within the dubbing of the film (the canonical '*Forastero*' is used), or in any other film. Concerning titles, however, two other Westerns that include that word, but are not part of the corpus, did make use of the canonical Spanish word in their translation: *The Outlaw/El forajido* (1943/1976, not released in Spain until Franco died, a 33-year gap that had a lot to do with the daring, for Spanish censors, performance of actress Jane Russell), and *The Last Outlaw/El último forajido* (1993/1994).

As for archaisms related to means of transport, '*Diligencia*' automatically became canonical because of the popularity of *Stagecoach*, whereas 'Buckboard', for instance, has never found a stable term. In that first film of the corpus it was translated as the contemporary '*Carreta*' (Cart), whereas in *My Darling Clementine/Pasión de los fuertes* (1946/1948), the archaic term '*Calesín*' (Gig) was used, and it has followed another unsteady path all the way to *Django Unchained /Django desencadenado* (2012/13), where another archaism etymologically connected to the previous one, '*Calesa*' (Calash), can be found. The generalized attempt to find terms that were differentiated from standard equivalents which sounded more contemporary in the target culture derived into a lexical oscillation whenever a canonical element had not been established in the formative stages of the sociolect.

5.4.2 *Structural calque*

By 'structural calque', Wach (2013: 162) defines the process: "in which the foreign morphological structure and meaning of a foreign element is transferred (…).The result is the origin of a completely new formation due to the translation of foreign elements." In the list of distinctive lexical elements compiled for this chapter, we can find 'Bounty hunter' for the first time in *The Wild Bunch/Grupo salvaje* (1969), transferred into Spanish by means of the structural calque '*Cazador de* *recompensas*'. This practice had existed in Spain since 1947 (Perea-Delgado 2012), when monetary rewards started being given for the capture and/or execution of '*maquis*', Republican guerrillas who had taken refuge in the mountains but there was no official term for it. *Cazador de recompensas* was thus adopted as a canonical term in the sociolect for Westerns, but not in everyday life. It eventually blended into '*Cazarrecompensas*' with *Geronimo: An American Legend / Gerónimo: una leyenda* (1993/1994), though both combinations have alternated in the last two decades.

Another structural calque can be found for the translation of "Redskin" as "*Pielroja*", found in the corpus for the first time in *The Man who Shot Liberty Valance/El hombre que mató a Liberty Valance* (1962), whereas a translation by paraphrase of "To scalp" was already used in the first film of the corpus, *Stagecoach/ La diligencia*: '*Arrancar la cabellera*' (tear off someone else's body of hair), whose verb has alternated with '*Cortar*' (cut off) ever since.

5.4.3 *Loan shift*

According to Hock's (1991: 398) definition, loan shifts: "arise from a shift in meaning of an established native word so as to accommodate the meaning of a foreign word". In the distinctive lexis of the sociolect for Westerns we can find an example with 'Saloon': "A place where alcoholic drinks are sold and drunk" (Cambridge Dictionary), which astoundingly became '*Salón*' in six of the twelve dubbed Westerns in which the term appeared, even though it already existed in Spanish to refer to 'Living room/Lounge'. The fact that it acquired in Spanish the meaning of the English cognate may have been due to the physical presence of the word 'Saloon' outside the premises in many Westerns, encouraging an adoption which included the semantic extension. Other options taken to translate it were yet another anglicism, '*Bar*', which had entered Spain through journalism (Stone 1957) and can be found in *High Noon/Solo ante el peligro* (1952/1953), or the traditional Spanish word '*Taberna*' (Tavern), in *3.10 Train to Yuma/El tren de las 3.10* (1957/1958)), though the latest choice, in *Django Unchained/Django desencadenado* (2012/2013) has been '*Cantina*' (Canteen).

On the whole, lexical elements which have become canonical in the Spanish sociolect for Westerns have also derived from other translation strategies besides anglicisms, but still favoring the generalized foreignizing tendency observed in some of these case studies.

5.5 Censorial issues

The fact that the diachronic analysis of this corpus includes Westerns released in Spain under a military dictatorship makes censorship a relevant issue, since it was during that period when many of the distinctive terms that make up this sociolect were established. Ninety per cent of audiovisual translation was from English into Spanish because of the predominance of the American film industry (Merino 2001), and, as Camus-Camus (2015: 9) points out: "With no exception, translations of US Westerns, one of the most popular genres during Franco's dictatorship, had to pass through the 'purifying' censorship filter". Interestingly, many of those manipulated lexical terms have remained after the advent of Spanish democracy because they were so deeply embedded in the target cultural context.

The third English term emphasized in the quotation of *Stagecoach/La diligencia* that opened the analysis of this corpus, 'Posse', was translated as '*Rurales*', a word which did not survive, as none of the other four unsteadily used during the dictatorship: '*Patrulla*' (Patrol), '*Fuerza*' (Force), '*Voluntarios*' (Volunteers) and '*Grupo*' (Group). The term had no official Spanish equivalent even though the issue existed and was put into practice at the time with a distinctive purpose: volunteers who sympathized with the military regime were recruited in order to capture political dissidents and, in most cases, execute them without a proper trial (Díaz Díaz 2016; Cabrera Martín, 2014; Barragán Mallofret & Castro Fernández, 2005). As mentioned earlier, *The Ox-Bow Incident/Incidente en Ox Bow* (1943/1973) tells the story of a posse that lynches three innocent cowboys; it was banned in Spain for three decades, and when the film was finally released, the term was translated as '*Patrulla*' the six times it was uttered in the source text. Earlier, when the dubbed version of *High Noon/Solo ante el peligro* (1952/1953) was premiered in Spain, the uncertainty was such that two different terms were used in the same film:

(8) …with a **posse** behind me, maybe there won't be trouble.
…respaldado por esa ***fuerza*** *podré reprimir cualquier tumulto.* (00.15.11)

(9) You must be crazy comin' in here to raise a **posse**.
Debe de estar loco cuando viene aquí a reclutar ***voluntarios****.*

With the advent of democracy, when no censorial practice interfered, there was still no established term, not only in the sociolect for Westerns, but neither in common everyday speech to refer to the recent historical past. The unsteadiness continues with still another three words found in the corpus to translate 'Posse': '*Soldados*' (Soldiers) in *The Outlaw Josey Wales/El fuera de la ley*; '*Pelotón*' (Squad) in *Geronimo: An American Legend/Gerónimo: una leyenda* and '*Partida*' (Party) in *True Grit/Valor de ley*.

I would also like to point out the fact that two American Westerns produced during the dictatorship which included the term in their original title, *Posse from Hell* (1961) and *Posse* (1975), were translated in both cases with yet another word: *Justicieros del infierno* (1961) and *Justicieros del oeste* (1976), using an adjective (‘Righteous’ would be its closest equivalent, with its positive connotation) as a noun. These inconsistencies are also worth emphasizing because they can portray a pattern of irregular translational behavior due to political reasons in origin. The glossary expanded even further: another American Western, *Posse* (1994), was released that same year in Spain under the title *Renegados* (Renegades).

A hyperonym worth foregrounding is ‘*Verdugo*’ (Executioner), term used to name the officer who enforced the capital punishment in Spain, and canonically used in *Westerns* to translate the hyponym ‘Hangman’, which can still be found most recently in *Django Unchained/Django desencadenado*: (12) “Ain’t nobody gonna cheat the **hangman** in my town.”; “*Nadie le va a quitar el gusto al* ***verdugo*** *de mi pueblo*.” (00:21:32). In a sociolect which erased so many normative Spanish words and supplied other alternatives, it is remarkable that no derivation was elaborated from the verb ‘*Ahorcar*’ (Execute by hanging) to refer to the officer who performed it in Westerns; the official capital punishment in Spain was the very local ‘Garrote’, “to kill someone by putting metal wire or collar around their neck and pulling it” (Cambridge Dictionary), with a different iconography. This could also have had a political motivation in origin, but this time to normalize the term in a period with a strong international pressure upon the Spanish fascist regime because of the illegality of its political executions, starting with the U.N. resolution in 1946 (Houston 1952).

Another censorial issue was the banning of swearwords under Franco’s ruling, which even affected Spanish classical literature (Bastianes 2018). Chronologically, the first films compiled for the corpus did not include them in the source texts, since the American film industry had its own filter with the Production Code Administration and the Catholic Legion of Decency (Camus-Camus 2015). But once the Motion Picture Association of America lifted its ban on swearing in 1968 (Jowett 1990), director Sam Peckinpah promptly included the expletive ‘Son of a bitch’ in *The Wild Bunch/Grupo salvaje* (1969), which seems to have caught Spanish dubbing translators by surprise.

(10) – **Son of a bitch**!
– Bounty hunters?
– ***¡Hijo de perra!***
– *¿Cazadores de recompensas?* (00:07:47)

Even though the commonly used equivalent '*Hijo de puta*' can be traced back to such respectable literary works as *Don Quixote* (Cervantes 1605), it was banned (as any other swearing) during the dictatorship. What we can find in this dubbed version is a structural calque of the original, '*Hijo de perra*' (Son of a female dog), a euphemism that no Spaniard would use in everyday life. The uncertainty produced in the Spanish cultural context under Franco by a term that, in American Westerns, would become widespread (absolutely all the films in the corpus which come after *The Wild Bunch* include 'Son of a bitch' in their scripts) can be verified by the watered down translations used in the following two films: '*Canalla*' (Scoundrel) in *Chisum* and '*Hijo de mala madre*' (Son of a bad mother) in *The Outlaw Josey Wales/El fuera de la ley*. But more remarkable is the fact that, during the Spanish democracy, in spite of the disappearance of censorial constraints, '*Hijo de perra*' had become canonical for the genre; it can still be found in *The Assassination of Jesse James by the Coward Robert Ford / El asesinato de Jesse James por el cobarde Robert Ford* (2007). An interesting case can be observed in *Unforgiven/Sin Perdón* (1992), where 'Son of a bitch' is repeatedly translated as '*Hijo de perra*', while 'Whore' is translated with the expletive taboo word '*Puta*', another example of how conventionalized terms become embedded in localized intertexts.

Other more recent dubbed westerns of the corpus do include the normative '*Hijo de puta*', starting with *Geronimo: An American Legend/Gerónimo: una leyenda* (1993/1994), and in the most recent film of the list, *Django Unchained/Django desencadenado* (2012/2013), we can even find a more contemporary blending: '*Hijoputa*',

6. Conclusions

From a diachronic perspective we can observe how, on the whole, the initial use of normative Spanish words in the translation for dubbing of American Westerns promptly gave way to a foreignizing tendency, blatant in the use of anglicisms such as '*Sheriff*', '*Rifle*' or '*Revólver*', but also as a result of other translation strategies that have formed a distinctive sociolect specifically related to this genre in the target context. The current state of the corpus, over 320.000 words, with its compilation still at an early stage, does not allow for a fixed pattern to be deduced, but the longevity of many lexical elements which do not follow the word usage rules of the Spanish language, but have survived well into the 21st century, enables me to suggest that the sociolect developed proves to be as stable as the standard mode of linguistic transfer, dubbing, in Spain. Nornes (2007: 191) remarks that "preference for either subtitling or dubbing is none other than a naturalized convention.

Audience research has shown that people tend to prefer whatever form of translation they grew up with." This seems to be the case as well concerning the sociolect for Westerns.

Therefore, the tendency to provide seemingly stilted dubbed versions which result from source-oriented translations, coined as dubbese, has not been reversed in this specific genre in spite of the awareness of linguistic researchers and translators. The distinctive terminology of the sociolect for Westerns in Spain has become so embedded that it is unusual to find innovative alternatives in the most recent dubbed versions. The prevalence of '*Sheriff*' to refer to various figures of authority, or the generalized erasure of '*Escopeta*' to translate 'Shotgun' and its replacement for '*Rifle*', would reveal a willingness to linguistically differentiate the genre within audiovisual translation, in accordance with its also exceptional iconographic code. We can also find '*Revólver*', yet another anglicism which is not even used in the corpus of source texts, to translate both 'Pistol' and 'Gun', when there are more normative Spanish terms available. But to flaunt foreignisation seems to be the favored pattern.

Other not so obvious features of dubbese have also shown their resilience. The use of '*Salón*', a Spanish term that means 'Living-room', which experienced a semantic extension in the sociolect for Westerns as a translation of the English cognate 'Saloon', is yet another relevant example of the survival of non-normative lexical elements: it is still found in *True Grit/Valor de ley* (2010/2011). Its coexistence with other terms established in the early stages of the development of the sociolect which do not result from a source-oriented translation, such as the archaisms '*Forastero*' or '*Forajido*', only proves the pervading force of conventionalization.

The censorial pressure during the period of the dictatorship also played an important role in the development of this sociolect, and, so far, the four decades of democracy reinforce the perception of predominance of the glossary established in the previous era, since some of the terms imposed for ideological reasons are still present in contemporary dubbed versions in spite of the absence of legal coercion. The most striking example is the survival of '*Hijo de perra*' well into the 21st century, a euphemism that would even seem non-commercial in a period when American Westerns are striving towards modernization, apparently implying an increase in the use of taboo language; in Spain, however, the contextualized tradition of the sociolect prevails. The established cohesion between the strongly codified iconography of the Western and a glossary of terms assimilated by means of constant exposure to previous dubbed versions resisted the change of political regimes. The artificial homogeneity achieved through decades of closely watched translations for dubbing still rules.

References

Adamou, Christina and Simone Knox. 2011. "Transforming Television Drama through Dubbing and Subtitling: *Sex and the Cities*." *Critical Studies in Television* 6/1: 1–21. https://doi.org/10.7227/CST.6.1.3

Agost, Rosa. 1999. "Serialitat i traducció." In *Cuerpos en serie*, ed. by Vicente Benet and Eloísa Nos Aldás, 91–106. Castellón: Universitat Jaume I.

Altman, Rick. 1984. "A Semantic/Syntactic Approach to Film Genre." *Cinema Journal* 23 (3): 6–18. https://doi.org/10.2307/1225093

Armstrong, Nigel. 2004. "Voicing 'The Simpsons' from English into French: a story of variable success." *JosTrans: The Journal of Specialised Translation* 2: 97–109, available online https://www.jostrans.org/issue02/art_armstrong.php (last access 13 April 2020).

Audissino, Emilio. 2014. "Dubbing as a Formal Interference: Reflections and Examples." In *Media and Translation: An Interdisciplinary Approach*, ed. by Dror Abend-David, 97–117. New York & London: Bloomsbury.

Baños, Rocío. 2014. "Insights into the false orality of dubbed fictional dialogue and the language of dubbing." In *Media and Translation: An Interdisciplinary Approach*, ed. by Dror Abend-David, 75–95. New York & London: Bloomsbury.

Baños, Rocío. 2009. *La oralidad prefabricada en la traducción para el doblaje. Estudio descriptivo-contrastivo del español de dos comedias de situación: 'Siete vidas' y 'Friends'*. Unpublished Doctoral Thesis. University of Granada, Spain.

Baños, Rocío and Fredric Chaume. 2009. "Prefabricated Orality: A Challenge in Audiovisual Translation." *Intralinea* 6, available online http://www.intralinea.org/specials/article/Prefabricated_Orality (Last access 20 April 2020).

Barragán Mallofret, Daniel and Juan Luis Castro Fernández. 2005. "Arqueología de la justicia. Arqueología de las víctimas de la Guerra Civil Española y de la represión franquista." *Rampas* 7: 149–174. https://doi.org/10.25267/Rev_atl-mediterr_prehist_arqueol_soc.2004.v7.07

Bastianes, María. 2018. "Un clásico difícil. Censura y adaptación escénica de *La Celestina* bajo el franquismo." *Hispanic Research Journal. Iberian and Latin American Studies* 19 (2): 117–134. https://doi.org/10.1080/14682737.2018.1444417

Cabrera Martín, Marta. 2014. *La impunidad de los crímenes cometidos durante el fascismo. Obligaciones del Estado español bajo el derecho internacional*. Luarca, Asturias: Asociación Española para el Derecho Internacional de los Derechos Humanos.

Calhoun, John. 2004. "The Ox-Bow Incident." *Cineaste* 29 (3): 55–6.

Camus-Camus, Carmen. 2015. "Negotiation, Censorship or Translation Constraints? A Case Study of *Duel in the Sun*." In *Audiovisual Translation. Taking Stock*, ed. by Jorge Díaz Cintas, and Josela Neves, 8–27. Newcastle upon Tyne: Cambridge Scholars Publishing.

Caron, Caroline-Isabelle. 2003. "Translating Trek: Rewriting an American Icon in a Francophone Context." *The Journal of American Culture* 26 (3): 329–355. https://doi.org/10.1111/1542-734X.00095

Cervantes Saavedra, Miguel de. 1605. *El ingenioso hidalgo Don Quijote de la Mancha*. Madrid: Ediciones de la Lectura, 1911.

Chaume, Fredric. 2004. "Film Studies and Translation Studies: Two Disciplines at Stake in Audiovisual Translation." *Meta: journal des traducteurs*, 49 (1): 12–24. https://doi.org/10.7202/009016ar

Chiaro, Delia. 2007. “The Effect of Translation on Humour Response: The Case of Dubbed Comedy in Italy.” In *Doubts and Directions in Translation Studies: Selected Contributions from the EST Congress, Lisbon 2004*, ed by Yves Gambier, Miriam Schlesinger and Radegundis Stolxe, 137–152. Amsterdam and Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/btl.72.16chi

Clark, Walter Van Tilburg. 1940. *The Ox-Bow Incident*. New York: Random House, The Modern Library Classics, 2004.

Cronin, Michael. 2009. *Translation goes to the Movies*. New York & London: Routledge.

Díaz Díaz, Benito. 2016. ”Tiempos de violencia desigual: guerrilleros contra Franco (1939–1952).” *Vínculos de Historia* 5: 105–120. https://doi.org/10.18239/vdh.v015.008

Ellis, Jack C. 1995. *A History of Film*. Boston: Allyn and Bacon.

Even-Zohar, Itamar. 1978. “The Position of Translated Literature within the Literary Polysystem.” In *The Translation Studies Reader*, ed. by Lawrence Venuti (2000), 192–197. London & New York: Routledge.

García Rodrigo, Francisco Javier. 1887. *Historia verdadera de la Inquisición. Tomo 3*. Madrid: Alejandro Gómez Fuentenebro.

Gomery, Douglas. 1980. “Economic Struggle and Hollywood Imperialism: Europe Converts to Sound.” *Yale French Studies 60*, Cinema/Sound: 80–93. https://doi.org/10.2307/2930006

Görlach, Manfred. 2001. *A Dictionary of European Anglicisms*. Oxford: Oxford University Press.

Gottlieb, Henrik. 2012. “Old Films, New Subtitles, More Anglicisms.” In *Audiovisual Translation and Media Accessibility at the Crossroads. Media for All 3*, ed. by Aline Remael, Pilar Orero and Mary Carroll, 249–272. Amsterdam: Rodopi.

Heinink, Juan B., Florentino Hernández Girbal and Robert G. Dickson. 1992. *Los que pasaron por Hollywood*. Madrid: Verdoux.

Hock, Hans Heinrich. 1991. *Principles of Historical Linguistics*. Berlin & New York: Mouton de Gruyter. https://doi.org/10.1515/9783110219135

Holmes, James S. 1972. “The Cross-temporal Factor in Verse Translation.” *Meta: journal des traducteurs* 17 (2): 102–110. https://doi.org/10.7202/003078ar

Houston, John A. 1952. “The United Nations and Spain.” *The Journal of Politics* 14 (4): 683–709. https://doi.org/10.2307/2126447

Jakobson, Roman. 1959. “On linguistic aspects of translation.” In *On Translation*, ed. by Reuben A. Brower, 232–239. Cambridge: Harvard University Press, 2000. https://doi.org/10.4159/harvard.9780674731615.c18

Jarvinen, Lisa. 2012. *The Rise of Spanish-Language Filmmaking. Out from Hollywood’s Shadow, 1929–1939*. New Brunswick, New Jersey & London: Rutgers University Press. https://doi.org/10.36019/9780813553283

Jowett, G. S. 1990. “Moral Responsibility and Commercial Entertainment: Social Control in the American Film Industry 1907–1968.” *Historical Journal of Film, Radio and Television* 10 (1): 3–31. https://doi.org/10.1080/01439689000260011

Keating, Carla Mereu. 2014. “The translation of ethnonyms and racial slurs in films. American blackness in Italian dubbing and subtitling.” *European Journal of English Studies* 18 (3): 295–315. https://doi.org/10.1080/13825577.2014.944020

Kilgarriff, Adam; Rychlý Pavel; Srmrz Pavel & David Tugwell. 2004. “The Sketch Engine.” In *Proceedings of the Eleventh EURALEX International Congress*, 105–115. Lorient: Université de Bretagne-Sud.

Laviosa, Sara. 2011. "Corpus-Based Translation Studies: Where does it come from? Where is it going?" In *Corpus-Based Translation Studies. Research and Application*, ed. by Alet Kruger, Kim Wallmach and Jeremy Munday, 13–32. London: Bloomsbury. https://doi.org/10.1080/10228190408566201

Luyken, Georg-Michael, Thomas Herbst, Jo Langham Brown, Helen Reid and Hermann Spinhof. 1991. *Overcoming Language Barriers in Television Dubbing and Subtitling for the European Audience*. Manchester: European Institute for the Media.

Merino, Raquel. 2001. "Presentación de la Base de Datos TRACE (Traducciones Censuradas Inglés-Español)." In *Trasvases culturales: literatura, cine y traducción 3*, ed. by Eterio Pajares, Raquel Merino and José Miguel Santamaría, 287–295. Zarauz: Universidad del País Vasco.

Merkle, Denise. 2010. "Censorship." In *Handbook of Translation Studies. Volume 2*, ed. by Yves Gambier and Luc Van Doorslaer, 18–21. Amsterdam & Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/hts.1.cen1

Myers, Lora. 1973. "The art of dubbing." *Filmakers Newsletter* 6: 56–58.

Naranjo, Beatriz. 2015. "Translating Blackness in Spanish Dubbing". *Revista Española de Lingüística Aplicada* 28 (2): 416–441. https://doi.org/10.1075/resla28.28.2.03nar

Nornes, Abé Mark. 2007. *Cinema Babel. Translating Global Cinema*. Minneapolis & London: Minnesota University Press.

Pavesi, María. 2019. "Corpus-based audiovisual translation studies: ample room for development." In *The Routledge Handbook of Audiovisual Translation*, ed. by Luis Perez-González, 315–333. London & New York: Routledge.

Pavesi, María. 2016. "The Space of Italian Dubbing: From Naturalness to Creativity in Fictive Orality." In *The Languages of Dubbing. Mainstream Audiovisual Translation in Italy*, ed. by Michela Canepari, Gillian Mansfield and Franca Poppi, 13–30. Roma: Carocci.

Pavesi, María. 2005. *La traduzione filmica. Aspetti del parlato doppiato dall'inglese all'italiano*. Roma: Carocci.

Parini, Ilaria. 2009. "The transposition of Italian American in Italian dubbing." In *Translating Regionalised Voices in Audiovisuals*, ed. by Federico Federici, 157–178. Roma: Aracne. https://doi.org/10.4399/97888548288588

Perea Delgado, Luisa M. 2012. "'Los de la Sierra', presencia de la guerrilla antifranquista en los montes de Tarifa." *Al Qantir* 12: 184–197.

Pérez Galdós, Benito. 1905. *Aita Tettauen*. Madrid: Viuda e hijos de Tello.

Pérez-González, Luis. 2014. *Audiovisual Translation. Theories, Methods and Issues*. London & New York: Routledge. https://doi.org/10.4324/9781315762975

Queen, Robin. 2004. "'Du hast jar keene Ahnung': African American English dubbed into German." *Journal of Sociolinguistics* 8 (4): 515–537. https://doi.org/10.1111/j.1467-9841.2004.00272.x

Rabadán, Rosa. 1991. *Equivalencia y traducción. Problemática de la equivalencia translémica inglés-español*. León: Universidad de León.

Sanderson, John D. 2010. "The Other You. Translating the Hispanic for the Spanish Screen." In *Polyglot Cinema. Migration and Transcultural Narration in France, Italy, Portugal and Spain*, ed. by Verena Berger and Miya Komori. 49–71. Vienna & Berlin: Lit Verlag.

Shuttleworth, Mark and Moira Cowie. 1997. *Dictionary of Translation Studies*. Manchester: Saint Jerome.

Snell-Hornby, Mary. 1988. *Translation Studies: An Integrated Approach*. Amsterdam & Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/z.38

Spolsky, Bernard. 1998. *Sociolinguistics*. Oxford: Oxford University Press.

Stone, Howard. 1957. "Los anglicismos en España y su papel en la lengua oral." *Revista de Filología Española* Vol. XLI (1/4): 141–160. https://doi.org/10.3989/rfe.1957.v41.i1/4.1046

Toury, Gideon. 1995. *Descriptive Translation Studies – And Beyond*. Amsterdam & Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/btl.4

Trudgill, Peter. 2003. *A Glossary of Sociolinguistics*. Oxford: Oxford University Press.

Valdeón, Roberto A. 2018. "Language and translation in classic westerns: revising stereotypes in *They Died with their Boots on* and *Fort Apache*." *Language and Intercultural Communication* 18 (6): 681–693. https://doi.org/10.1080/14708477.2018.1445746

Venuti, Lawrence. 1995. *The Translator's Invisibility: A History of Translation*. London & New York: Routledge. https://doi.org/10.4324/9781315098746

Wach, Szymon. 2013. "Calquing English Terminology into Polish." *Academic Journal of Modern Philology* 2: 161–169.

Willis García-Talavera, James F. 2002. "El programa de mano Metro Goldwyn Mayer en los inicios del cine sonoro: las versiones en habla castellana." *Revista Latina de Comunicación Social* 52: 602–618.

Databases and dictionaries

Cambridge Dictionary. Open access. https://dictionary.cambridge.org

Diccionario de la lengua española. Real Academia Española. Open access. https://dle.rae.es

Generic analysis of mobile application reviews in English and Spanish

A contrastive corpus-based study

Natalia Mora López
Universidad Complutense de Madrid

In this paper, the Systemic Functional Linguistics (SFL) approach to genre is applied to English and Spanish mobile application reviews from Google Play Store. These reviews are shown to be divisible into two optional stages, namely Evaluation and Description, the first part being more subjective and based on the author's opinion and the second one more objective and focused on describing issues. The frequency of use of the Attitude axis from Appraisal Theory provides support for the distinction of those stages. The combination of the polarity of the contents of each stage together with the number of stars given to the item reviewed creates six patterns for the reviews. However, no noticeable differences between English and Spanish reviews were observed.

Keywords: product reviews, application stores, appraisal theory, genre, SFL

Introduction

In the last few decades, users have witnessed the emergence and spread of new mobile technologies that have led to new ways of using language. One of these widespread media is mobile application stores, in which users can download for free or after due payment an application for their mobile devices. When deciding whether getting or dismissing it, users commonly take a look at reviews written by other users to know whether the app is what they are looking for.

These reviews express their author's opinion, although this may be done in several ways. The common way to do it would be by using subjective language. Subjectivity in language refers to those elements or aspects of language that convey an opinion, evaluation or speculation (Banfield, 1982; Wiebe 1994). However, reviews may also contain objective sentences, which are those that present "factual

https://doi.org/10.1075/btl.158.11mor

information about the world" (Liu, 2012: 19). However, this does not mean that when authors mention factual information they are not expressing their opinion on the item reviewed, since, as Liu (2010: 634) does also point out, there are opinionated objective sentences: "an opinionated sentence is a sentence that expresses explicit or implicit positive or negative opinions. It can be a subjective or objective sentence." This difference is highly relevant for reviews, since not only subjective expressions convey opinions, but also factual elements. For example, an objective statement like "this app does not work," when referring to a mobile application, entails a very negative meaning.

Previous studies on reviews from application stores have usually focused on general linguistic aspects and automatic analysis. For example, Vasa et al. (2012) analysed 8.7 million reviews of 17,330 apps and concluded that comments tend to be short but informative and that there is a relation between the length of the review and the rating given: the poorer the rating was, the longer the comment. Khalid (2013) analysed 6,390 reviews, paying special attention to the complaints users made, which typically included objective functional errors and requests for additional features. Finally, Guzman and Maalej (2014) analysed 32,310 reviews and proposed an automatic approach to help developers filter and analyse user reviews by extracting features automatically from the reviews.

What is missing in these previous analyses is a characterisation of mobile application reviews as constituting or belonging to a specific genre because of its structural characteristics, that is, its staging, and the lexicogrammatical features which can provide evidence for this staging. Also, to my knowledge, there are no contrastive analyses of mobile application reviews in English and Spanish which investigate their similarities and differences. My aim in this paper is, therefore, to explore the staging and the lexicogrammatical characteristics of mobile application reviews in a bilingual comparable sample randomly collected from a larger corpus of Google Play Store reviews (Mora López 2017).

The remainder of the paper is organised as follows: firstly, the theoretical framework and tools used for the analysis are presented; secondly, a description of the mobile application review genre and the corpus used in this study is provided; thirdly, the chapter focuses on the generic stages of mobile application reviews from the lexicogrammatical point of view, based on appraisal features, and the structural perspective, based on the patterns found in the English and the Spanish samples; finally, the last section summarises the results of the corpus analysis and provides some concluding remarks.

Theoretical framework

The analysis of the reviews presented here is based on two main theoretical concepts: the notion of genre as a sequence of stages and the classification of emotions, judgements and evaluations as the axis called Attitude inside Appraisal Theory. These two concepts are developed in the next two sections.

Genre analysis

As said above, these reviews share some features. It is proposed here that these features characterise reviews in a way that makes them different from other types of reviews. In the 20th century, long before the creation of application stores, Bakhtin (1986: 60) referred to genres in the following way: "Each separate utterance is individual, of course, but each sphere in which language is used develops its own relatively stable types of these utterances. These we may call speech genre." However, this was a very broad definition that focused on the fact that language in communication is an individual and concrete realisation produced by participants' utterances. These utterances are, in turn, grouped together, and create types of speech genres.

More recently, Bhatia (2004: 23) proposed a set of conditions that texts should follow for it to be considered a genre.

> Genre essentially refers to language use in a conventionalised communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms by imposing constraints on the use of lexico-grammatical as well as discoursal resources.

Therefore, a genre has a conventionalised communicative setting, a specific set of communicative goals, stable structural forms and uses lexicogrammatical and discoursal resources. The way these aspects are realised in reviews of games and applications was shown in Mora López (2017), when analysing the main general features of application and game reviews, as well as music, book and film reviews from Google Play Store, and it is also summarised in Section 3.1.

The definition of genre and the aspects that characterise it can also benefit from the notion of genre in Systemic Functional Linguistics (SFL). This perspective focuses on the goal-oriented aspect of the text as the basis for its definition (Martin 1984; Eggins & Martin 1997). Additionally, genres are also characterised according to their 'generic stages' or 'moves'. The idea of analysing genres as consisting of stages (potential or realised) goes back to work by SFL scholars such as Hasan and Halliday (Hasan 1984; Halliday & Hasan 1985) and has been elaborated on by Martin, Eggins and others (Martin 1985; Swales 1990, 2004; Eggins 1994; Skelton 1994; Eggins & Martin 1997; Eggins & Slade 1997; Nwogu 1997).

Eggins (1994: 37) presents staging, that is, the schematic structure of a genre, as a description of the parts that form the whole, and what the relation among those parts is. Swales (2004: 228) defines 'moves' or 'stages' more specifically as "discoursal or rhetorical units that perform a coherent communicative function in a written or spoken discourse." Those moves are identified by the linguistic features they have (lexical meaning, propositional meanings, illocutionary force, etc.), since they not only provide the segments with a uniform orientation but also signal the content of discourse that can be found in them (Nwogu 1997: 122).

This means that, apart from the conditions Bhatia proposes, genres do also show specific stages in their structure. This approach to the analysis of genre has been useful when dealing with task-oriented dialogues (Taboada 2003, 2004a; Taboada and Lavid, 2003), electronic bulletin boards (Taboada 2004b) and book reviews (Taboada, 2011). Therefore, this paper will show which stages are found in the reviews studied here.

Appraisal theory

Appraisal Theory, mainly developed by Martin and White (Martin 2000; White 2003; Martin and White 2005), presents a system to classify evaluative language or those expressions that indicate "the subjective presence of writers/speakers in texts as they adopt stances towards both the material they present and those with whom they communicate" (Martin and White 2005: 1). Appraisal resources are divided in this system into three axes: Attitude, which is mainly concerned with feelings, such as emotions, judgements or evaluations; Engagement, which focuses on the ways speakers position themselves towards the text; and Graduation, which deals with the degrees of intensity of the meanings expressed by the two previous categories.

The distinction of the stages of application and game reviews draws on the use of Attitude expressions (see Lexicogrammatical features of stages). These expressions are in turn divided into three subcategories, namely Affect, Judgement and Appreciation. Affect expresses positive or negative feelings as well as emotional reactions, such as happiness, anger or disgust (Example 1). Judgement provides moral evaluations about people and the way they behave in relation to normative rules (Example 2). Finally, Appreciation expressions address evaluations on things, considering aspects such as reactions they produce, their composition or their value (Example 3). Attitude expressions have been underlined in the examples below.

(1) *I love this game*

(2) *All Booster got erased – Unfair*

(3) *Amazing*!!

The identification of Attitude expressions is particularly relevant in the analysis of the stages in application and game reviews since it will be shown how the amount of this type of items is significantly more profuse in one stage than in the other. This difference is the main basis to distinguish the information presented in them.

The mobile application review genre and the corpus

There are some characteristics that make the application review genre different from other product reviews. Before analysing the specific features found in this study, some basic considerations about this genre are explained below: firstly, some general aspects of the genre are presented; secondly, they are compared to the results from previous studies on other product reviews; finally, the corpus analysed is described.

General characteristics of mobile application and game reviews

Writing an online review on Play Store is restricted by some limitations imposed by the setting and users' communication goals, while other characteristics of these reviews are also reflected on users' writing style (Mora López 2017). Additionally, the platform provides a list of indications for posting reviews. Some of them forbid the use of sexual, abusive or fake content and are focused on respect to other users; however, some others refer specifically to the content of the review in a more relevant way for the style of this type of reviews, such as "Try to include both positives and drawbacks", "Make your comments useful and informative", "Post clear, valuable, and honest information", "Use proper grammar and check your spelling" and "Keep it readable; don't use excessive capitalization and punctuation" (Google, n.d.). In these cases, the platform is advising a specific use of the language and restriction of the content so that the comments posted by users become adequate for the setting, communicative goals and style of these reviews. However, frequently, users do not totally respect these rules.

The platform Play Store can be accessed via mobile application or website. Therefore, users can post their comments either with the help of a physical keyboard or tapping the letters on the screens of their devices. These reviews will be displayed inside the page of the app users are checking.

The original communicative goal to write a review is to tell others your opinion about the app, including reasons. However, users have deviated from assessing the item (whether they like it or not) and have sometimes created a second and third purpose: to describe the performance of the items, which is actually not a personal, subjective opinion but an objective description of an item; and to establish direct communication with developers, usually to ask for updates, changes and help (either to developers or other users), which again is not an opinion but a request.

It is interesting to note that the interface and indications to users are different depending on the platform used to write the review. When writing a review via website, the message users see is "Tell others what you think about this app. Would you recommend it, and why?" but, when using a mobile device, the message changes to "Describe your experience (optional)." This second instruction is much wider than the first one and promotes a more descriptive review, despite the fact that it refers to users' experiences, than the one seen in the website environment.

Only users who have downloaded the item can post a review about it. These reviews are limited to about 4000 characters if written via website and 500 characters if via mobile device and in either case must be given together with a 1-to-5 star rating. Actually, the star rating is necessary when reviewing the item, but including a message is optional. The written messages provided usually contain internet and mobile language (abbreviations, emoticons, etc.) and grammar rules are not always followed: it is easy to find absent subjects and connectors. Regarding their structure, reviews typically consist of two stages: first, users provide a personal evaluation; second, users explain the positive or negative features they have experienced with the app or game, or why they give that rating. A deeper analysis of these stages is presented in the section Generic stages in mobile application reviews.

Comparison with reviews about other products and from other platforms

Previous studies on the analysis of product reviews have shed some light on the characteristics of the genre. Some of them are partially shared with the reviews analysed here while others are not.

Pollach examined the generic features of 358 online reviews on digital cameras. She found that these reviews were comprised of "comments, and evaluations, and personal stories (e.g. weddings, vacations, christenings) involving the products reviews" (Pollach 2006: 4). Apart from the fact that the basic structure presented here has an Evaluation part, it can also be seen that, when Descriptions relate personal experiences and problems users have been encountering, they can be similar to narratives (Examples (17), (20)). She also points out the use of emoticons and appeals to other users (as in Example (20)), and the use of informal language with grammar and spelling mistakes. However, reviews on Play Store sometimes include messages addressing the developers of the application or game (Examples (16), (24)).

A more specific study on the content of product reviews is Vasquez's (2012). She focuses on "small narratives" in 100 TripAdvisor negative reviews on hotels, and considers them to be canonical in terms of their structure (optional abstract, orientation, complication and action, resolution, coda), with evaluation appearing at any point in the narrative. These more elaborated narratives have not been found in reviews from Play Store, since they are brief and descriptive of very specific issues related to the item. Additionally, although some spared evaluative expressions may be found in the Description stage, the overall meaning is not evaluative but descriptive.

An analysis of the generic stages of consumer reviews was performed by Taboada (2011), who studied a corpus of 50 movie reviews and concluded that, similarly to the structure suggested here, movie reviews consisted of a Description and an Evaluation stages. However, she indicates that the Evaluation stage is necessary, and actually most of the reviews included it at the end of the review, in contrast to the findings from Play Store application and game reviews, where Evaluation is found at the beginning and none of the stages is intrinsically necessary.

Corpus

The corpus analysed consists of 200 reviews (100 English and 100 Spanish) written since the respective creation of each app until 2017. They address well-known products: *Candy Crush* and *Instagram*. In turn, they are separated into two groups of 50 game and 50 application reviews each. Finally, those groups contain half of the reviews with positive ratings (4 and 5 starts out of 5) and half of them with negative ratings (1 or 2 starts out of 5). This distribution is summarised in Table 1.

Table 1. Distribution of the texts analysed

Complete corpus (200 texts / 12064 words)							
Spanish (100 texts / 5189 words)				**English (100 texts / 6875 words)**			
Games (50)		**Applications (50)**		**Games (50)**		**Applications (50)**	
+ (25)	– (25)	+ (25)	– (25)	+ (25)	– (25)	+ (25)	– (25)

These reviews belong to a larger corpus which was the result of a process of automatic extraction and filtering collected by Mora López (2017). This larger corpus includes not only applications and games, but also reviews about books, films and music from Google Play Store. Table 2 presents the distribution of number of reviews per category in the larger corpus.

For this analysis, as explained above, only 200 reviews about applications and games were selected randomly, although manually in order to assure an even distribution, as seen in Table 1.

Table 2. Summary of English and Spanish comments extracted in larger corpus (Mora López 2017)

	Spanish	English
Applications	15225	15721
Games	15328	15288
Books	2223	4909
Films	1595	7793
Music	2933	5976
Total	**37304**	**49687**

It must be said that the authors of the reviews may or may not be native speakers of the language they use, since access to the Spanish and English sites is not restricted. This is especially important in the case of English, whose use as second and foreign language is widespread. In the case of Spanish, many users show features of Latin American Spanish (lexis, spelling which reflects a specific pronunciation and accent, etc.).

Although app reviews are very common and the reader may be familiarised with them, a typical example of such a review is given in Example (4).

(4) *I love Instagram but it crashes on my galaxy note 3 I'm almost always on Instagram, but lately I've been noticing that only when I use Instagram that it freezes my phone. I would be watching a video and then I'll click out of it and it will either stop, or it will continue playing the sound of the video while I'm on something else… pls fix this issue*

In this example, the user starts by indicating an emotion or feeling ("I love Instagram") and develops a justification by describing several issues experienced ("but it crashes […] else…"). A direct request can be seen at the end of the review ("pls fix this issue").

Generic stages in mobile application reviews

A more detailed analysis of the structure and content of these reviews is presented below. The results show that there are two basic stages, Evaluation and Description, which can be distinguished by the frequency of use of Attitude expressions. Additionally, these two stages and the polarity of their contents create several combinations which are systematically repeated throughout the corpus.

In the analysis of the structure of the 200 reviews, it was observed that reviews are consistently divided into two stages: firstly, users provide a personal evaluation; secondly, users explain the positive or negative features they have experienced with the app or game, or why they give that rating.

The first stage, the Evaluation, is subjective and describes how the app makes users feel or the impression they have about the app. The contents in this stage are typically related to meanings of Attitude. As mentioned above, Attitude meanings are those that express feelings, like emotional reactions (Affect), judgments of behaviour (Judgement) and evaluation of qualities of things (Appreciation). More specifically, the Evaluation stage comprises expressions of the first and last aspects: Affect, that is, positive or negative emotions (e.g. 'like', 'love', 'hate', 'happy', 'disappointed', etc.), and Appreciation, that is, evaluations of things (e.g. 'addictive', 'boring', 'useless', 'original', 'great', 'bad').

This first stage is much shorter than the second one, the Description. It may consist of a single word, phrase, or sentence. However, the second stage is longer and may range from a single sentence to an extended narration (within the limits of 1200 characters). This stage is more objective than the Evaluation. Here, users explain the reasons why they have a given opinion on the reviewed item or describe the item's performance, characteristics and issues. This may include a description of or mention to how the item works, new features included or former features now removed, bugs they have found and problems they are experiencing.

However, although these two stages are likely to be related, none of them is necessary, that is, both of them may appear together or in isolation, as shown in Examples (5)–(8) (only Evaluation) and (9)–(12) (only Description).

(5) *Just awesome* (5 stars)

(6) *Bad* (1 star)

(7) *Me gusta mucho* ('I like it a lot') (5 stars)

(8) *X Muy muy mala* ('X Very very bad') (1 star)

(9) *Level 1807 doesn't work right S6 edge and iPhone 6s* (1 star)

(10) *Something's wrong When i try to follow an account it says I'm following it and goes green but then it goes back to white and says "follow" and when i check thr numbers for that I'm following it shows that i follow that account by the number but when i refresh it, it goes back to the number before that number and if i press follow over and over it will add more numbers to it. Please fix this* (4 stars)

(11) *No me dejan seguir a nadie* ('I can't follow anyone') (1 star)

(12) *Falla desde la actualización* ('It doesn't work since it was last updated') (2 stars)

Therefore, the formula that summarises the basic structure of the stages of mobile application reviews can be stated as (13). In this formula, the caret indicates sequence and the brackets, optionality.

(13) (Evaluation) ^ (Description)

Lexicogrammatical features of stages

The distinction between Evaluation and Description is based on the type of information conveyed by each of them, as explained above. The Attitude axis inside Appraisal Theory is especially relevant to distinguish them since it covers expressions related to emotions, feelings and opinions. These subjective, personal meanings are likely to be included in a stage such as Evaluation but not so commonly in the Description, which is more narrative and objective.

The 200 texts in this corpus were analysed in search of Attitude expressions. These were manually annotated in an Excel file by the author of this paper and classified into one of the three categories inside Attitude (Affect, Judgement and Appreciation). An example of an Attitude expression in a review from the corpus can be seen in Example (14), where the word "awesome" (underlined for clarifying purposes) was classified as Appreciation.

(14) Evaluation: *It's awesome,*
Description: *but for a while I didn't have a phone number but I had instagram after a while the thing popped up when you had to verify your phone number, my friend said that his just went away once he exited out but mine I didn't let me exit out, If I didn't have a phone number and verify I could follow people, post, like or comment. Please fix that for future people.*
Star rating: 4

Table 3 presents a list of the five most repeated positive and negative Attitude items in English and Spanish and the number of times they were found. It must be noted that the shorter the span is (that is, if is a single word), the more likely it is to be repeated. Therefore, the results show either adjectives or verbs as the most common resource.

Table 3. Five most common positive and negative Attitude items in English and Spanish

English				Spanish			
Positive	**#**	**Negative**	**#**	**Positive**	**#**	**Negative**	**#**
love	20	annoying	8	me gusta ([I] like)	17	mala (bad-femenine)	7
fun	12	unfair	7	me encanta ([I] love)	9	malo (bad-masculine)	4
like	12	stupid	5	bien (well)	9	peor (worse, worst)	4
awesome	11	bad	5	buena (good-femenine)	9	copia (copy)	3
good	7	disappointed	3	mejor (better, best)	9	mal (bad, badly)	3

Table 4 summarises the findings related to the use of Attitude in the English and Spanish corpora. The difference between the amount of words included in attitudinal spans in Evaluation and Description is significant in both languages, with a p-value lower than 0.05 (0.00014751 in English and 9.4871E-05 in Spanish).

Table 4. Distribution of Attitude words per stage in English and Spanish

	English				Spanish			
	Evaluation		Description		Evaluation		Description	
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Total words	445	100	6430	100	705	100	4484	100
Attitude words	91	20.44	150	2.33	159	22.55	129	2.87

It can be seen that the number of words covered by the Evaluation stage is much lower than by the Description, which represents the difference in length between the stages, as pointed out in the previous section. Also, although Descriptions may also contain Attitude realisations, they are only a very small portion of the number of words used in that stage, since it is mostly devoted to narratives and descriptions.

Regarding the distribution of Attitude spans per category (Affect, Judgement and Appreciation), Table 5 shows the raw number of spans and their percentage in the total amount of Attitude spans in each language. As with Table 4, it can be seen that English and Spanish users share similar habits not only in the use of Attitude in each stage but also the choice of subcategory.

Table 5. Distribution of Affect, judgement and Appreciation spans in English and Spanish

	English				Spanish			
	Evaluation		Description		Evaluation		Description	
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Affect spans	36	15.5	42	18.1	35	16.9	27	13.1
Judgement spans	4	1.7	17	7.3	3	1.4	11	5.3
Appreciation spans	57	24.5	76	32.7	77	37.3	53	25.7
Attitude spans	232 (100%)				206 (100%)			

Patterns observed in the stages

The combination of the two elements and the star rating given (users must provide, together with their comment, a 1–5 star rating) produces six different patterns in the reviews. These patterns are summarised in Table 6.

Table 6. Generic stages and ratings found in the corpus

	Evaluation	&/or	Description	Stars
A	Positive		positive aspects due to which 4–5 stars are given	4–5
B	Positive		negative aspects due to which 5 stars are not given	4
C	Positive		negative aspects, despite which 5 stars are given	5
D	Negative		negative aspects due to which 1–2 stars are given	1–2
E	Negative		*used to like it* negative aspects due to which 1–2 stars are given	1–2
F	Positive		negative aspects due to which 1–2 stars are given	1–2

It can be seen that positive Evaluations do not necessarily match positive Descriptions or even positive ratings. However, a negative Evaluation was not found to be followed by a positive Description (unless pattern E is considered as such), although it may have a positive rating. Each of these patterns is explained and exemplified below.

Pattern A: *Positive (Evaluation) + Positive aspects (Description) + 4–5 stars (Rating)*

It might be expected that positive Evaluations were followed by positive Descriptions and ratings. However, this is only one of the patterns found for positive Evaluations. In these cases, although users may not give an excellent 5 star rating, they do not mention any problems or reasons why they do not do so, but focus on the positive aspects of the app (Examples (15)–(16)).

(15) Evaluation: *Fun to play,')*
Description: *I'm slowly progressing to level 1805!? WOW! I SHOULD hope to get there on that level, b4 I'm 60!Finally made it to level 184! 😆 I see you added at my request a free days pass at a cost $$ to Unlock those New higher Levels 😊 This is a plus + Kudos 🙌 CCS is very fun 2play, ') TY CC FOR changing the sending/receiving lies format 😆🙌KUDOS! & Perfect! Thank you! 😊 free wheel spin cud use upd8! ie: add gold bars a +! Almost had Big BONUS/wheel, ') Big BONUS wheel is rigged! Nearly had jackpot several times! It stops to hesitate, Yet never Win it!? 😯 Has Anyone Ever? Probably Not! 😐 Ty! 4 sum upd8 changes 😆 You have listenNd & Ty for those changes 😄💋I'm now @ lvl 184! 😆*
Star rating: 5

(16) Evaluation: *Me ha encantado Me encanta esta red social y encima la actualización que an hecho es fantástica seguro que le gusta a todo el mundo y por eso yo le doy 5 estrellas……..*
Description: *E leído los comentarios y a muchas personas le falla la nueva actualización. A mi no me falla nada 👍 les doy la enhorabuena a los creadores de instagram y por favor no cambien la nueva actualización.*
Star rating: 5

The Evaluation stages are the first part of the comment. They may include a short title for the comment and be followed by the explanation. For example, (16) opens with the short title "*Me ha encantado*" 'I loved it', and then starts the comment "*Me encanta esta red…*" 'I love this net…'. The first part, which is considered to be Evaluation, includes Attitude meanings like "fun", "love", "*ha encantado*" '(I) loved', "*encanta*" '(I) love', "*fantásica*" 'fantastic', "*gusta*" '(everybody) likes.

The Description stage includes explanations about the user's experience. For example, (15) tells other users how their progress is going on and the level they have reached. They also thank the developers of the app and mention specific features the game has as well as complaining about a problem they are having with an element in the game; finally, they make some requests to the developers. Some Attitude items can be found in the Description: "plus", "very fun", "perfect", "thank you". However, this represents a very small part of the Description, which focuses on the experience.

Example (16) shows the same structure. It starts with a positive Evaluation based on subjective perceptions like "*encanta*" '(I) love' or "*fantástica*" 'fantastic', among others. Then, the Description merely refutes other users' comments who say that the app does not work. Finally, they thank developers and request not to make any changes. Thanking is considered an Attitude expression, since it is related to the speaker's feelings. However, it is the only element in all the Description that can be seen as such.

Pattern B: *Positive (Evaluation) + Negative aspects (Description) + 4 stars (Rating)*

This pattern consists of a positive Evaluation followed by some negative aspects in the Description that justify the choice of a 4 star rating. If a positive Evaluation appears with a 4 star rating but no Description is included, these reviews would be regarded as realisations of pattern A.

(17) Evaluation: *I love Instagram*
Description: *but it crashes on my galaxy note 3 I'm almost always on Instagram, but lately I've been noticing that only when I use Instagram that it freezes my phone. I would be watching a video and then I'll click out of it and it will either stop, or it will continue playing the sound of the video while I'm on something else… pls fix this issue*
Star rating: 4

(18) Evaluation: *Instagram Stories De verdad la app es demasiado buena, la mejor red social que existe,*
Description: *desde mi punto de vista, sólo tengo un inconveniente: No se me activa Instagram Stories, a partir de esa Novedad lo he actualizado 4 veces y Nada.. ¿Que pasa?*
Star rating: 4

Example (17) shows the distribution of the contents into Evaluation and Description, where the former includes the Attitude verb "love". The Description focuses on a problem the user has and the actions the user does to trigger the bug. There are no Attitude elements in the Description.

The Evaluation in (18) shows Attitude expressions like "*buena*" 'good', "*mejor*" 'best', and then describes the issue that justifies a 4 star rating: Instagram Stories does not work. Finally, the author asks other users or developers about the issue. If the word "inconveniente" 'disadvantage' is considered an Attitude expression, Evaluation would be extended up to that point. In this case, it has been understood as a synonym of "issue" or "problem" and included in the Description, so no Attitude items appear in that stage in this review.

Pattern C: *Positive (Evaluation) + Negative aspects (Description) + 5 stars (Rating)*

The third pattern consists of a positive Evaluation and a Description including negative aspects that the user found in the application or game, but which do not prevent them from giving a 5 star rating, unlike what was found in the previous pattern, where only 4 stars were given. If a positive Evaluation appears with a 5 star rating but no Description is included, these reviews would be regarded as realisation of the structure in pattern A.

(19) Evaluation: *5 STARS !!!!! Yes… I Rate This App 5 Stars! Who Doesn't Like Instagram!!!???*
Description: *But There Has Been At least 3 Updates And The Tagging Still Isn't Fixed! I Highlighted Their Handle Names And Pressed That Arrow Button Which Tags Them Automatically.… But After Doing So, No Tagged Names Appear In The Comment Box. I Have To Actually Type Their Handle Name In Order For It To Work.*
Star rating: *5*

(20) Evaluation: *Me encanta esta app desde que la conozco*
Description: *Aunque ahora tengo problemillas al grabar en instagram stories, al grabar directamente se ve muy oscuro ya sea la cámara delantera o trasera. No lo consideré un problema ya que al darme cuenta de que podía subir videos y fotos de hace menos*

de 24 horas pensé en grabarlo directamente con mi cámara y así subirlo. Pero ahí no terminó, ya que al subir algún activo de video que ya tenía grabado no se podía ver para nada nítido… Se me veía con muchos pixels los videos. Alguien sabe por qué? Alguna solución? ☺
Star rating: 5

In Example (19), the author of the review states and actually exclaims that they are giving this application 5 stars, which may be interpreted as Attitude, in the sense that it is 'very good'. The Evaluation also contains the Attitude expression "like". The Description talks about an issue which has not been fixed after several updates and what this problem entails. There are no Attitude expressions in this stage.

Example (20), similarly to other examples, includes the verb "encanta" 'love' in the Evaluation. The Description revolves around a problem the user experiences when recording videos which are uploaded to the application and ends asking other users whether they know why that problem occurs. Again, there are not Attitude meanings in this Description.

Pattern D: *Negative (Evaluation) + Negative aspects (Description) + 1–2 stars (Rating)*

This structure turns to negative comments, that is, those rated with only 1 or 2 stars. In this case, the Evaluation they include is also negative, and the Description focuses on those negative aspects that justify their opinion.

(21) Evaluation: *Very disappointed*
Description: *This game is having 2 many problems. I had a striped booster pressed it so i could use it & the game acted like it was going to play & came back twice for me to press play again. The game took my striped booster without me playing it. Fix the bugs in this game!!! Also as of 7/24/16 this game is doing the same thing it was before when u play 1 world & have no more lives left then u have none in either world. Whats the point of having 2 worlds & 5 lives in each if u cant use all from each world. Very MAD!!!!*
Star rating: 1

(22) Evaluation: *Mala*
Description: *Se cierra sola la aplicación no puedo jugarr*
Star rating: 2

Example (21) shows an Attitudinal Evaluation, "disappointed", and the Description explains the two main problems the user finds in the game. Apart from the explanation of the issues, the user asks developers to fix the bugs. After more than 100 descriptive words, the user includes an Attitude expression, “mad”, to close the review and reinforce the Evaluation given at the beginning.

Example (22) is a very short review. The Evaluation is reduced to a single word, “mala” ‘bad’, and the Description simply states that the game force closes and, consequently, the user cannot play, so no Attitude items are included.

Pattern E: *Negative (Evaluation) + Negative aspects (Description) + 1–2 stars (Rating) (‘Used to like it’ variant)*

There is a recurrent strategy in reviews following the structure seen in pattern D which is referred to as pattern E in Table 6. Users state that they “used to like” the application or game, but now they do not, due to updates and changes carried out. This variant may also be realised by stating that previous versions were better than the current one, in a way that it can be assumed that they liked the previous version better. Descriptions typically include mentions to the changes and comparisons to previous versions and performance. This variant is shown in Examples (23)–(24).

(23) Evaluation: *I’m Being Kind This use to be a nice app.*
Description: *Now, I have a ripoff of Snapchat, disappearing messages in my DMs, and my followers hardly see any of my content because of your new algorithm which has messed everything up! A few suggestions: Take out the “Stories” feature, you practically stole it from Snapchat. Bring back the old algorithm so everthing is seen not just what “we’d like” to see first. I want people to see the things I make as an artist. Fix the bugs with the DMs, because they’re beyond annoying.*
Star rating: 2

(24) Evaluation: *Muy mal, me frustra Despues de la ultima actualizacion no me gusta nada,*
Description: *añadiron una lista de puntuacion, para que pese mas la app, los peses antes se combinaban con la galleta y se volvian todos del mismo color ahora no, y lo peor delo peor en las misiones se traba me salgo de la aplicacion y cuando regreso ya no estan las misiones de tiffi, siempre les habia dado 5 estrellas pero esta vez no, meti una memoria especial para tener los juegos de king, todo para nada*
Star rating: 1

Pattern F: *Positive (Evaluation) + Negative aspects (Description) + 1–2 stars (Rating)*

Although negative aspects are described and a low rating is given (1–2 stars), users may still include a positive Evaluation at the beginning of the review. When no Evaluation is given, these examples are regarded as pattern D.

(25) Evaluation: *I like Instagram,*
Description: *but this app seriously crashes every time I open it. I have to clear the cache and data to get it open, and then leave it running, because if I close it, and then try to open it, it crashes. (Edited to add: this has been going on for months, and I've tried uninstalling, reinstalling, moving it to my SD card, buying a new SD card---nothing fixes it.)(Edited again, months later, to add: I send in a crash report every single time this happens, with my logs. Developers should have DOZENS of reports on this issue.)*
Star rating: 1

(26) Evaluation: *Entretiene*
Description: *Pero hay veces se sale solo*
Star rating: 2

The Evaluation stage in Example (25) simply states "I like Instagram", with the Attitude expression "like" in it. However, 1 star is given and the Description relates the main issue the user experiences: the app crashes constantly. The review is edited twice, as the author indicates, and adds further explanation on attempted solutions and complaints, but no Attitude expressions are found in this stage.

Example (26) is shorter. The Evaluation stage refers to a positive quality of the game, which is that it "Entretiene" 'entertains', but it has a bug, as mentioned in the Description: it force closes.

Distribution of staging patterns in the English and Spanish corpora

Although several staging patterns have been found in the corpus, the frequency of appearance is not the same for all of them. Table 7 presents the distribution of the staging patterns in the English and Spanish corpora. Each language consists of 100 reviews, comprising 50 positive and 50 negative reviews about the application *Instagram* and the game *Candy Crush*.

Table 7. Distribution of staging patterns A-F and junk messages in English and Spanish reviews

	English	**Spanish**
Pattern A: P.Eval. + P.Desc. + 4–5 stars	24	30
Pattern B: P.Eval + N.Desc. + 4 stars	19	12
Pattern C: P.Eval + N.Desc. + 5 stars	5	8
Pattern D: N.Eval + N.Desc. + 1–2 stars	37	38
Pattern E: N.Eval + N.Desc + 1–2 stars ('*used to like it*' variant)	5	8
Pattern F: P.Eval + N.Desc + 1–2 stars	8	4
Junk messages	2	0

The figures show very little difference in the frequency of staging patterns A-F in the English and Spanish reviews, although no cases of junk messages were found in the Spanish sample while two cases were found in the English one. The most common patterns are A (positive Evaluation + positive Description + 4–5 star rating) and D (negative Evaluation + negative Description + 1–2 star rating), as could be expected. Additionally, these two patterns also include positive and negative reviews, respectively, which only consist of an Evaluation stage that matches the polarity of the star rating given, thus increasing the frequency of appearance.

Most important problems in the (automatic) analysis of application and game reviews

The automatic detection of subjectivity, opinions and sentiment has been attracting increasing interest especially since the beginning of the 2000s (Das and Chen 2001; Dini and Mazzini 2002; Pang, Lee, and Vaithyanathan 2002; Turney 2002; Dave, Lawrence, and Pennock 2003; Nasukawa and Yi 2003; Wiebe et al. 2003; Yu and Hatzivassiloglou 2003 inter alia). More specifically, some attempts have been made to automatically classify linguistic units as members of Martin and White's Appraisal categories (Taboada and Grieve 2004; Whitelaw, Garg, and Argamon 2005; Argamon et al. 2007; Taboada, Brooke, and Stede 2009; Khoo, Nourbakhsh, and Na 2012; Dotti 2013). Most studies focus on the Attitude axis and on specific word classes, typically adjectives (Taboada and Grieve 2004; Whitelaw, Garg, and Argamon 2005; Khoo, Nourbakhsh, and Na 2012) since, when working at the lexis level, Attitude meanings are more frequently found in them than in other word classes. However, they have shown that a successful automatic procedure to identify Attitude and Appraisal is still to be found and it remains as an open research line.

If the two stages found in this paper as basic constituents of Play Store comments were to be analysed automatically, three steps should be taken into account: firstly, Attitude meanings should be identified successfully; secondly, the ratio of these spans should be used to distinguish whether only one or both of the stages are included; thirdly, if both stages are present, a line should be drawn between them according to the distribution of Attitude items. In these steps, there are two main issues to be solved: the lack of a large, manually annotated sample to be used as a training corpus or a gold-standard for posterior automatic analysis in this type of texts; and, consequently, the lack of an automatic procedure that can analyse Attitude meanings successfully and reliably enough to substitute manual analysis.

Finally, some difficulties are added to the automatic analysis of these texts by some linguistic characteristics they show, such as ungrammaticality, misspellings, acronyms, abbreviations, emoticons and words specifically related to the applications reviewed. These features would make it harder to correctly identify words themselves and dependencies among the elements of the clauses. For example, in English, the word "great" was found to be abbreviated as "gr8" and "favourite" as "fav", while "greatest" was misspelt as "greates". In Spanish, some letters were found to be repeated in words to convey emphasis, as in "Bueniiisiiimo" 'veeeryyy goood', "malooo" 'baaad' and "Roboooo" 'theeeeft.' Also, the word "enrritas" '[you] get enrritated' was found as a misspelling of "irritas" '[you] get irritated', and sometimes a "k" is used instead of "qu", a very common resource in texts messages, for example, since they are pronounced similarly (e.g. "porkeria" 'muk' instead of "porquería" 'muck'), apart from the fact that written accents are largely disregarded.

Additionally, there are two very specific problems found in the analysis of this type of reviews. On the one hand, reviews can be found with a wrong rating assigned by the author; on the other hand, reviews may be cases of spam or off-topic.

A rating may be assumed to be wrong when both the Evaluation and the Description (or the one that appears if one of them is omitted) is absolutely positive or negative and the rating has the maximum score in the opposite polarity. This is illustrated by Examples (27)–(28), which do not belong to the 200 reviews corpus analysed previously, but to the larger corpus collected by Mora López (2017). Since the examples for the 200 reviews corpus were selected randomly, no cases of mismatches between comment and rating were found, although they are known to exist.

(27) Evaluation: *Insta It nice I love it*
Star rating: 1

(28) Evaluation: *I like it :-)*
Star rating: 1

In these examples, a star rating of 1 (the lowest rating, so the user is supposed to totally dislike the item) is given but the review has a positive polarity. It can be then assumed that users misunderstood the scale and thought that 1 was the highest score.

Regarding spam, off-topic content or junk messages, users may post reviews which do not provide any useful information or do not even make sense. In these cases, users may do personal promotion by sharing information about themselves or write irrelevant messages, as shown in Examples (29)–(35). Examples (29)–(33) were collected manually by taking a quick look at the most recent comments posted in May 2019, while 34–35 belong to the 200 reviews corpus analysed here.

(29) Comment: *XD no se que escribir*
Star rating: 5

(30) Comment: *n*
Star rating: 5

(31) Comment: *dadas en la dddddddd*
Star rating: 2

(32) Comment: *pomo ñ.*
Star rating: 5

(33) Comment: *like my photo instagram id @avi_yadav_pati__official*
Star rating: 5

(34) Evaluation: *Awesome.e It's such a fun game to play*
Comment: *when I'm alone and no one will talk to me. I wish I had a friend. One time I cut my foot with a can. I also had a dog once. I like to hangout with the ants outside and follow there trails. I ate dog poop one time that had a ring in it. Fun time. But anyway this game is alright. I hope I can be cool like these people. I wanted to be a ninja but I could get into the school for ninjas in ninja turtle town. HOW you are not the only thing that has I don't know how you are a couple ki*
Star rating: 5

(35) Comment: *Good luck with everything you need anything further please let us know if you have any other information contained in this email and any attachments are handled by the way to get the same time as I have a great day and I will be in the morning and I will be in the morning and I will be in the morning and I will be able or willing to do it for a few days ago and I will be in the morning and I will be in the morning and I will send the money is not the best way is to get the chance to look at this point in th*
Star rating: 5

These two aspects, the wrong rating and the irrelevant messages may make these reviews more difficult to analyse automatically, in addition to the other issues mentioned above. In the first case, they may hinder the automatic matching of a review and a rating by drawing conclusions from their content, while in the second case, the fact that these comments do not fit into the genre style and information is not related to the items may be misleading.

Summary and concluding remarks

After analysing a bilingual comparable English-Spanish corpus of 200 reviews on games and applications from Google Play Store, it can be concluded that this type of reviews shows a specific structure and content that distinguishes them from traditional reviews. On the one hand, there are two specific stages that characterise mobile application reviews: Evaluation and Description.

The first one provides a more subjective content while the second stage deals with the narrative or description of the reasons why users have that specific opinion about the item reviewed. Other contents that may be included in the Description are direct requests to developers and help requests to either developers or other users. Both of them may appear either in isolation or combination although in that case Evaluation does always come first.

The difference between these two stages is also supported by the use of Attitude expressions, as defined by Appraisal Theory, since the Evaluation stage has a significantly higher presence of these items while it is not so profuse in Descriptions. Additionally, the combination of the polarity of the Evaluation, the Description and the star rating associated with the review were observed to create six systematic patterns.

Regarding English and Spanish reviews, no noticeable differences were observed between languages: they have a very similar amount of Attitude expressions in both stages as well as a similar distribution of the six patterns found, which means that English and Spanish speaking users adjust to the genre in a similar way.

Although some light has been shed on the content and structure of this type of reviews, the analysis of a wider corpus might give room to additional patterns, like a negative Evaluation with a positive rating (4–5 stars), which had no samples found in this corpus.

References

Argamon, Shlomo, Kenneth Bloom, Andrea Esuli, and Fabrizio Sebastiani. 2007. "Automatically Determining Attitude Type and Force for Sentiment Analysis." In *Proceedings of the 3rd Language & Technology Conference (LTC-07)*, 218–231.

Bakhtin, Mikhail Mikhaĭlovich. 1986. *Speech genres and other late essays*, Vern W. McGee (trans.). Austin: University of Texas Press.

Banfield, Ann. 1982. *Unspeakable Sentences*. Boston: Routledge and Kegan Paul.

Bhatia, Vijay Kumar. 2004. *Worlds of Written Discourse: A Genre-Based View*. London: Continuum International.

Das, Sanjiv and Mike Chen. 2001. “Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards.” In *Proceedings of the Asia Pacific Finance Association Annual Conference APFA*, 37–56.

Dave, Kushal, Steve Lawrence, and David M. Pennock. 2003. “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews.” In *Proceedings of WWW*, 519–528. https://doi.org/10.1145/775152.775226

Dini, Luca and Giampaolo Mazzini. 2002. “Opinion Classification through Information Extraction.” in *Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining)*, 299–310.

Dotti, Fiorella Carla. 2013. “Overcoming Problems in Automated Appraisal Recognition: The Attitude System in Inscribed Appraisal.” *Procedia-Social and Behavioral Sciences* 95: 442–446. https://doi.org/10.1016/j.sbspro.2013.10.667

Eggins, Suzanne. 1994. *An Introduction to Systemic Functional Linguistics*. London: Pinter Publishers.

Eggins, Suzanne and James R. Martin. 1997. “Genres and registers of discourse”. In *Discourse as Structure and Process. Discourse studies: A Multidisciplinary Introduction*, ed. by T. A. van Dijk, 230–256. London: Sage. https://doi.org/10.4135/9781446221884.n9

Eggins, Suzanne and Diana Slade. 1997. *Analysing casual conversation*. London: Cassell.

Google. (n.d.). Comment posting policy. Retrieved from <https://play.google.com/intl/en_ae/about/comment-posting-policy/> (accessed on November, 2019).

Guzmán, Emitza, and Walid Maalej. 2014. “How do users like this feature? A fine grained sentiment analysis of app reviews.” In *22nd International Requirements Engineering Conference (RE)*. IEEE Press. https://doi.org/10.1109/RE.2014.6912257

Halliday, Michael Alexander Kirkwood, and Ruqaiya Hasan. 1985. *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Hasan, Ruqaiya. 1984. “The nursery tale as genre.” *Nottingham Linguistics Circular* 13. 71–102.

Khalid, Hammad. 2013. “On identifying user complaints of iOS apps.” In *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press. https://doi.org/10.1109/ICSE.2013.6606749

Khoo, Christopher Soo-Guan, Armineh Nourbakhsh, and Jin-Cheon Na. 2012. “Sentiment Analysis of Online News Text: A Case Study of Appraisal Theory.” *Online Information Review* 36 (6): 858–878. https://doi.org/10.1108/14684521211287936

Liu, Bing. 2010. “Sentiment analysis and subjectivity.” *Handbook of natural language processing* 2: 627–666.

Liu, Bing. 2012. “Sentiment analysis and opinion mining.” *Synthesis Lectures on Human Language Technologies* 5 (1): 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Martin, James R. 1984. “Language, register and genre.” In *Children Writing: Reader*, ed. by F. Christie, 21–30. Geelong, Victoria: Deakin University Press.

Martin, James R. 1985. “Process and text: Two aspects of human semiosis.” In *Systemic perspectives on discourse*, ed. by James Benson and William Greaves, 248–274. Norwood, NJ: Ablex.

Martin, James R. 2000. Beyond exchange: Appraisal systems in English. In *Evaluation in Text: Authorial Distance and the Construction of Discourse*, ed. by Susan Hunston and Geoffrey Thompson, 142–175. Oxford: Oxford University Press.

Martin, James. R. and Peter R. White. 2005. *The Language of Evaluation. Appraisal in English*. New York: Palgrave. https://doi.org/10.1057/9780230511910

Mora López, Natalia. 2017. Annotating Appraisal in English and Spanish product reviews from mobile application stores a contrastive study for linguistic and computational purposes. Unpublished PhD dissertation. Available online at <https://eprints.ucm.es/46878/1/T39717.pdf>

Nasukawa, Tetsuya and Jeonghee Yi. 2003. "Sentiment Analysis: Capturing Favorability using Natural Language Processing." In *Proceedings of the Conference on Knowledge Capture (K-CAP)*. https://doi.org/10.1145/945645.945658

Nwogu, Kevin Ngozi. 1997. "The medical research paper: Structure and functions." *English for Specific Purposes* 16 (2): 119–138. https://doi.org/10.1016/S0889-4906(97)85388-4

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs Up? Sentiment Classification using Machine Learning Techniques." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79–86. https://doi.org/10.3115/1118693.1118704

Pollach, Irene. 2006. "Electronic word of mouth: a genre analysis of product reviews on consumer opinion web sites." In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*. IEEE Press. https://doi.org/10.1109/HICSS.2006.146

Skelton, John. 1994. "Analysis of the structure of original research papers: an aid to writing original papers for publication." *British Journal of General Practice* 44: 455–459.

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

Swales, John. 2004. *Research Genres*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139524827

Taboada, Maite. 2003. "Modeling task-oriented dialogue." *Computers and the Humanities* 37 (4): 431–454. https://doi.org/10.1023/A:1025729107628

Taboada, Maite. 2004a. *Building Coherence and Cohesion: Task-oriented Dialogue in English and Spanish*. Amsterdam and Philadelphia: John Benjamins. https://doi.org/10.1075/pbns.129

Taboada, Maite. 2004b. "The genre structure of bulletin board messages." *Text Technology* 13 (2): 55–82.

Taboada, Maite. 2011. "Stages in an online review genre." *Text & Talk-An Interdisciplinary Journal of Language, Discourse & Communication Studies* 31 (2): 247–269. https://doi.org/10.1515/text.2011.011

Taboada, Maite and Jack Grieve. 2004. "Analyzing Appraisal Automatically." *American Association for Artificial Intelligence Spring Symposium on Exploring Attitude and Affect in Text*. Stanford. March 2004. AAAI Technical Report SS-04-07, 158–161.

Taboada, Maite, and Julia Lavid. 2003. "Rhetorical and thematic patterns in scheduling dialogues: A generic characterization." *Functions of Language* 10 (2): 147–179. https://doi.org/10.1075/fol.10.2.02tab

Taboada, Maite, Julian Brooke, and Manfred Stede. 2009. "Genre-Based Paragraph Classification for Sentiment Analysis." In *Proceedings of 10th Annual SIGDIAL Conference on Discourse and Dialogue*. London, UK. September 2009, 62–70. https://doi.org/10.3115/1708376.1708385

Turney, Peter D. 2002. "Thumbs Up Or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews." In *Proceedings of the Association for Computational Linguistics (ACL)*, 417–424.

Vasa, Rajesh, Leonard Hoon, Kon Mouzakis, and Akihiro Noguchi. 2012. "A preliminary analysis of mobile app user reviews." *Proceedings of the 24th Australian Computer-Human Interaction Conference*. ACM. https://doi.org/10.1145/2414536.2414577

Vásquez, Camilla. 2012. "Narrativity and involvement in online consumer reviews: The case of TripAdvisor." *Narrative Inquiry* 22(1): 105–121. https://doi.org/10.1075/ni.22.1.07vas

White, Peter R. 2003. "Beyond modality and hedging: A dialogic view of the language of intersubjective stance." *Text*, 23(2): 259–284. https://doi.org/10.1515/text.2003.011

Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. 2005. "Using Appraisal Groups for Sentiment Analysis." In *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*, 625–631. https://doi.org/10.1145/1099554.1099714

Wiebe, Janyce M. 1994. "Tracking point of view in narrative." *Computational Linguistics* 20 (2): 233–287.

Wiebe, Janyce M., Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane J. Litman, David R. Pierce, Ellen Riloff, and Theresa Wilson. 2003. "Recognizing and Organizing Opinions Expressed in the World Press." In *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, 12–19.

Yu, Hong and Vasileios Hatzivassiloglou. 2003. "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 129–136. https://doi.org/10.3115/1119355.1119372

Exploring variation in translation with probabilistic language models

Alina Karakanta[1], Heike Przybyl[2] and Elke Teich[2]
[1]Fondazione Bruno Kessler / University of Trento[1] / [2]Department of Language Science and Technology, Saarland University

While some authors have suggested that translationese fingerprints are universal, others have shown that there is a fair amount of variation among translations due to source language shining through, translation type or translation mode. In our work, we attempt to gain empirical insights into variation in translation, focusing here on translation mode (translation vs. interpreting). Our goal is to discover features of translationese and interpretese that distinguish translated and interpreted output from comparable original text/speech as well as from each other at different linguistic levels. We use relative entropy (Kullback-Leibler Divergence) and visualization with word clouds. Our analysis shows differences in typical words between originals vs. non-originals as well as between translation modes both at lexical and grammatical levels.

Keywords: translationese, interpretese, relative entropy, language models, parallel corpora, comparable corpora

1. Introduction

It is widely acknowledged that the process of translation leaves specific linguistic traces in the translation product, commonly referred to as translationese (Gellerstam 1986). While some authors have suggested that the fingerprints of translation are universal (Baker 1993; Chesterman 2004), others have shown that there is a fair amount of variation among translations related to source language interference (shining through), translation mode (oral vs. written) or level of expertise (e.g. learner vs. professional). Translationese effects are observed at all linguistic levels, from lexis and grammar to semantics and discourse. For example, looking at the

[1] Work performed while the author was at Saarland University.

https://doi.org/10.1075/btl.158.12kar

language pair English-German, Teich (2003) shows that translations exhibit shining through with regard to some grammatical features (e.g. word order) and (over-) normalization with regard to others (e.g. use of passive) and suggests to look for explanations in the contrastive typology of English and German. Shlesinger and Ordan (2012) demonstrate that mode (oral vs. written) has a stronger effect on translation output than status as translation (translation vs. original), indexed e.g. by type-token ratio or part-of-speech distributions in Hebrew as a target language. Baroni and Bernardini (2006) are the first to show that translations can be automatically identified comparing translations to original texts in the target language in an automatic classification task; and Koppel and Ordan (2011) pursue the hypothesis that translations from different source languages into the same target language are sufficiently different from each other so that the source language of a given translated text can be identified by means of automatic classification, here using function words (e.g. pronouns, discourse connectors) as features. More recently, Lapshinova-Koltunski and Zampieri (2018) look at genre and translation method (e.g. human vs. computer-aided) as additional factors impacting on translation output. Except for Rabinovich and Wintner (2015), who describe a clustering approach to translationese identification, most of the existing approaches use supervised methods, specifically automatic text classification, with predefined features that are assumed to be affected and are fairly easy to obtain. Specifically, the features are frequency distributions on (different kinds of) words, often with a bias towards high frequency phenomena. Also, depending on the concrete variable of interest (genre, mode of translation, target language, source language etc), different kinds of features are considered. This makes it hard, if not impossible, to compare the results of different studies. Obtaining a systematic picture of variation in translation is further impeded by the fact that often the corpora needed are compiled fairly opportunistically, so that there are few corpus resources that control for all or many relevant variables (see also Hareide (2019) for similar observations).

We here propose a generic approach to the analysis of translationese and variation in translation that avoids some of these drawbacks by being able to incorporate any variable of interest in corpus comparison and by supporting detection of features from corpus data. The approach uses probabilistic language models (LM) and methods of model quality assessment (here: relative entropy) and applies them to corpus comparison (Fankhauser et al. 2014). A language model is an account of the probabilities of all the words (or other units) in a corpus, typically in the context of two or three preceding words (*n*-grams). Using language models as representations of corpora, we can compare e.g. the probability distributions of words in original vs. translated texts in the same language. A standard method of comparing probability distributions is relative entropy, which measures the overall difference between two language models (in bits) and allows us to identify those features (e.g. words) that contribute most to a given difference (see also Teich et al. 2020 on the application of information theory in translation studies).

To obtain a rich enough corpus, we gathered relevant existing corpora of the EuroParl family (Koehn 2005) compiling them into a uniform representation and encoding as many relevant factors of variation as possible as corpus metadata, published as the EuroParl-UdS corpus (Section 2). We explain what probabilistic language models do and how they can be employed for corpus comparison (Sections 3.1–3.2). We showcase the application of our approach on the EuroParl-UdS corpus and on interpreting corpora of the Europarl family (after encoding the same relevant factors), demonstrating how we detect features involved in translation variation/translationese on two relevant variables: language pair/target language (here: English→German and English→Spanish) and translation mode (translation vs. interpreting) (Section 4). Finally, we briefly assess our approach and discuss future extensions, refinements and applications (Section 5).

2. Corpus data

Our study requires high-quality corpora, specifically tailored for translation studies. The large, publicly available parallel and comparable corpora usually used in Machine Translation (MT) (Koehn 2005) are unfortunately not suitable for exploring translation variation and detecting features of translationese. Since the size of the training corpus is decisive for the success of MT applications, corpora for training MT systems were created by concatenating all available data in a language pair, regardless of the translation direction or the status of the speaker (native vs. non-native). In a corpus used for studying human translation we at least need to know about the status of a given text (original vs. translation), otherwise no analysis or interpretation regarding the translation relation is possible. Other variables, such as speaker status, can also be very useful, e.g. for filtering out productions by non-native speakers.

We thus compiled a new data set from the EuroParl domain that reflects a number of relevant parameters – the Europarl-UdS corpus (Karakanta et al. 2018).[2] This corpus contains verbatim reports of proceedings of the European Parliament (EP) and their official translations, with additional information about translation direction and speaker status. It has to be noted that the basis of the written original texts is a spoken event which was edited before being published in order to fulfill written conventions. Although typical spoken language features such as hesitations

[2] Europarl-UdS is publicly available from http://fedora.clarin-d.uni-saarland.de/europarl-uds/

or false starts are corrected (and the speaker may also make corrections to the text), this text genre can be described as written with spoken characteristics. Translations from these written originals will also reflect the spoken character of the original. The EP proceedings were enriched with speaker metadata, such as nationality, political party, etc., which allows for selecting data based on specific criteria. We filtered the proceedings based on that metadata in order to obtain parallel and comparable corpora. More specifically, we created comparable corpora of written original texts produced only by native speakers of German (WO_DE) and Spanish (WO_ES). Similarly, we obtained corpora of translations into German (T_EN→DE) and Spanish (T_EN→ES) of texts produced only by native speakers of English. The specific languages were chosen to represent Germanic and Romance languages.

For interpreting, we selected the speeches delivered in Spanish by Spanish native speakers (SO_ES) and the simultaneous interpretations of speeches delivered by English native speakers into Spanish (SI_EN→ES) from the European Parliament Interpretation Corpus (EPIC; Sandrelli and Bendazzoli 2005). For German, we created a similar corpus of interpreted speech. We selected speeches delivered by English native speakers in the Translation and Interpreting Corpus (TIC; Kajzer-Wietrzny 2012) and EPIC Ghent (EPICG; Defrancq 2015), revised them according to our transcription guidelines and transcribed their interpreted versions into German in order to create a parallel corpus of simultaneous interpreting (SI_EN→DE).[3] We also transcribed German spoken originals by German native speakers for the existing English interpretations in TIC (SO_DE) and collected speaker and speech-related metadata. The corpus is still under compilation, therefore more speeches are added continuously. The spoken original dataset includes speeches that were prepared and read out by the speakers (e.g. some committee reports), but also impromptu speeches that were delivered without any written aid (e.g. replies to questions), or a mixture of both delivery modes. Therefore, the spoken original material can be described as partially written to be spoken, whereas SI output is seen as spontaneous speech. In general, Bendazzoli and Sandrelli (2005) describe EP data as a specific genre of simultaneous interpreting where the setting is very specific and homogeneous: strict rules for allocation of speaking time to Members of the European Parliament (MEPs) and fixed structure of debates lead to short speaking interventions at generally high speaking rates (medium EP speaking rate of 150 w/m is considered as high in other settings (Monti et al. 2005)). Moreover, interpreters are all professionals with a certain qualification, working into their mother tongue for the major European languages (Bendazzoli and Sandrelli 2005).

[3] The transcription guidelines are based on the EPICG transcription approach (Bernardini et al. 2016).

In order to ensure compatibility across the different transcription guidelines and to prepare the data for language modeling, we removed all information relating to the phonetic transcription (intonation, non-verbalized noises, non-standard pronunciation, etc).

Statistics of the corpora are presented in Table 1. The small size of the corpora, especially for interpreting, definitely poses challenges to the application of probabilistic methods for analysing variation in translation. However, this choice is driven by the particularity of the translationese signal being weak in comparison to other factors such as genre, therefore many of the features of interest may be low-frequency features. This calls for high-quality data, where the variables relating to their production have been strictly controlled. For this reason, we have filtered the data based on translation direction and whether they were produced by a native speaker. In addition, creating corpora of interpreting data is a particularly tedious and time-consuming activity, due to the effort necessary for reliable transcription. This work has been based on collecting the right data, well-structured and representative of our research questions, rather than collecting a large amount of data. Moreover, our models (see Section 3) use smoothing techniques and allow for controlling the statistical significance values, and therefore are robust to low-frequency phenomena.

Table 1. Corpus sizes for comparable written originals (WO) and spoken originals (SO), as well as for parallel data; translations from English (T_EN) and simultaneous interpretations from English (SI_EN) for German and Spanish. For the parallel data words are reported on the target side

Translation	Europarl-UdS		Interpreting	EPIC-UdS (DE) / EPIC (ES)	
	Sentences	Words		Sentences	Words
T_EN→DE	137,813	3,100,647	SI_EN→DE	582	15,662
T_EN→ES	125,852	3,162,915	SI_EN→ES	1,780	39,101
WO_DE	427,779	7,869,289	SO_DE	766	15,234
WO_ES	183,361	5,661,374	SO_ES	587	14,579

3. Methods

3.1 Probabilistic language models and analysis of translation variation

Representation of the probabilities of linguistic entities (words, sentences etc.) is a key component in natural language processing and empirical computational linguistics, finding application in a wide range of tasks, from spell checking to machine translation. A language model (LM) assigns probabilities to the sentences in a corpus by computing the probability of the sequence of words W they contain:

(1) $P(W) = P(w_1, w_2, w_3, \ldots, w_n)$

For instance, a sentence “It’s raining cats and dogs” is on the whole more likely than e.g. “It’s raining cats or dogs” or “It’s raining dogs and cats” or “It’s raining pigs and chickens”. Thus, a language model is a summative representation of a given linguistic behavior. Similarly, LMs are used to compute the probability of an upcoming word by taking into account the word’s preceding context, standardly represented as a conditional probability:

(2) $P(w_n \mid w_1, w_2, \ldots, w_{n-1})$

where the probability of the next word w_n is estimated in the context of the preceding words w_1, w_2 until w_{n-1}. Such *n-gram* models may consider one (bigram), two (trigram) or more preceding words in context in order to predict the next word.[4] Commonly, such models operate with a logarithmic scale, which has been shown to directly index human online language processing behavior (known as ‘surprisal’; Crocker et al. 2016).

From a corpus-linguistic perspective, language models are simply summative models of relative frequencies of the linguistic units contained in a given corpus (typically words) considered in their *n*-gram context. Thus, the representation of a corpus in terms of a language model provides a suitable basis for all kinds of linguistic analysis where relative frequency plays a role, from register analysis to diachronic accounts. In the same way, when applied to translation corpora, language models allow us to inspect translation behavior(s), e.g. to assess the plausibility of alternative translation solutions. For example, an expression *pervasive impact* will be a less common combination in English than *big impact*, suggesting that *big impact* might be the safer translation solution of the two overall. The probability of items in the target language is therefore related to trying to meet target language expectations in the translation. However, from the perspective of finding the best match for a source language expression in the target language, we are concerned with finding the most probable target language expression *E_TL* given a source language (*E_SL*), i.e. *P(E_TL | E_SL)*. Because of this additional constraint, it may turn out that other expressions are favored in translations into a language A compared to original texts in language A. Extending the previous example, *big impact* may be less likely in translations of German *große Auswirkung* than *major effect* (Examples 1a and 1b).

(1) a. *daß jede Frage, wie die Steuern behandelt werden, natürlich eine* ***große Auswirkung*** *auf die gesamte globale Wirtschaft der Welt hat.*

[4] For an introduction to *n*-gram language models see Jurafsky and Martin (2008: Chapter 3).

b. *that every aspect of tax management naturally has a* ***major effect*** *on the entire global economy.*

In the present study, we build word-based unigram language models (LMs) with Jelinek-Mercer smoothing for different types of texts, using the corpora described in Section 2. Smoothing allows us to obtain reliable statistics even for corpora with few documents and words. More specifically, we create LMs for written originals (WO), spoken originals (SO), translations (T) and simultaneous interpretations (SI) for the two language pairs. As units/features we use words, after applying tokenization, lowercasing and number conflation on the corpora. To obtain insights into differences according to our variables of interest, we need to compare the probability distributions obtained across the corpora. To this aim we use the information-theoretic measure of relative entropy.

3.2 Comparing language models by relative entropy

For comparing the probability distributions computed with language models trained on the different corpora in Section 2 across modes, we use relative entropy, also known as Kullback–Leibler Divergence (KLD). KLD is a well-known information theoretic measure of similarity/dissimilarity between probability distributions and has recently been increasingly used in corpus-based analyses of linguistic varieties and styles (Hughes et al. 2012; Klingenstein et al. 2014; Degaetano-Ortlieb and Teich 2018). It allows us to measure divergences in the probability distributions over specific linguistic features. More specifically, relative entropy shows how many extra bits of information are needed to encode a distribution *P*, with a model created to perfectly encode a distribution *Q*.

(3) $$D_{KL}(Q||P) = \sum_{i} Q(i) \log\left(\frac{Q(i)}{P(i)}\right)$$

The more additional bits a unit (*i*) needs for encoding, the more distinctive it is for a specific corpus.

To illustrate how KLD is applied in our analysis scenario, consider the following example. We build a unigram language model on the Spanish written originals (WO_ES), which corresponds to distribution *Q*. Then, we build a second unigram language model on the translations into Spanish (T_EN→ES); this is distribution *P*. For a word *i=España*, which appears in both models, we compute its probability based on models *Q* and *P*. Suppose we have *Q(España) = 0.05* and *P(España) = 0.0001*. We compute KLD between the distribution *P* and *Q* for the word *España* and find it equal to 0.0849. This means that 0.0849 extra bits of information would be required for encoding *España* appearing in the translations with our distribution based on the original texts. We repeat the same procedure for all words *i* in the corpus and collect the sum for all words to obtain the KLD for the whole distribution.

For visualization of KLD scores on individual words we use word clouds (Fankhauser et al. 2014). For an example of a word cloud displaying all words that contribute to the distinction between written mode originals vs. translations in the target language German see Figure 1a–b. Size corresponds to degree of distinctivity of a word for a given corpus (KLD score) and color encodes relative frequency (red=high frequency, blue=low frequency). The value for the relative frequency (per million) appears at the left hand-side border of the word clouds. To ensure that the findings are statistically significant we filter words in the word clouds by setting a *p*-value = 0.05.

4. Analysis and results

In order to detect the units that contribute to the distinctiveness among corpora, we inspect the word clouds provided by the KLD visualization. Our analysis shows differences for the factors of translation direction and translation mode. Since the KLD visualisations are interactive, it should be noted that we include screenshots of the comparisons between pairs of distributions, i.e. Figure 1a is compared with Figure 1b, and Figure 1c with Figure 1d.

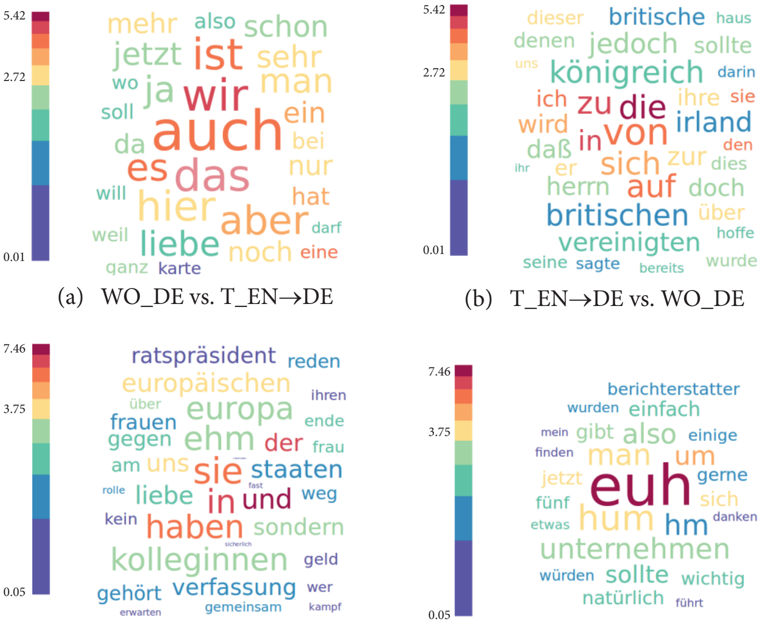
4.1 Translation direction: Originals vs. Translation/Interpreting

We observe the distinctive units for written originals vs. translations and spoken originals vs. interpretations (see Figure 1).[5] Figure 1(a) shows typical words of original German texts in written mode in the domain of European Parliament, while Figure 1(b) shows typical words in German translations from English in the same domain. The differences shown are all significant at $p = 0.05$. At the lexical level, topic-related differences prevail; the German originals deal mostly with general topics (very few lexical items), while the translations from English use more diverse lexis, especially dealing with national issues (*Irland, britischen, vereinigten*), etc. At the level of grammar, words such as *es, man, ist, wer* suggest the use of impersonal structures for the originals, which are typical for German. We also observe different

[5] Written original speeches are the officially published versions by the European Parliament, which have been edited in order to convey to written language conventions. The spoken originals are true transcripts containing spoken language characteristics like interjections, unfinished sentences etc.

pronouns (*wir* in originals vs. *ich* in translations), which suggests a different cultural convention in self-mention in this domain. The written originals show more spoken language features such as deictic markers (*jetzt, hier*) and speech particles (*also, ja, aber*), which are not present in translations. We could interpret this fact as an effect of normalisation of the edited text in the translation process in order to make it more representative of the written domain.

Turning to the spoken mode, Figures 1(c) and 1(d) show typical words for spoken originals and for simultaneous interpreting, respectively. Interpreting is characterised by speech and hesitation markers (*also, euh, hum, hm*) as well as very few high frequency words. Even though the originals are also spoken transcription, these elements are not so dominant and the texts are richer in vocabulary, which can be attributed to the fact that some of the original speeches are prepared, while the interpreters face a high cognitive effort and simplify their utterances because of time constraints. Here we observe an interesting effect; in the written mode, the originals show more elements of the spoken domain compared to translations, while in the spoken mode, the simultaneous interpretations are more representative of the spoken domain. This shows a process of adapting the translation (or interpreting) output to the expectations of the target audience.



(a) WO_DE vs. T_EN→DE

(b) T_EN→DE vs. WO_DE

(c) SO_DE vs. SL_EN→DE

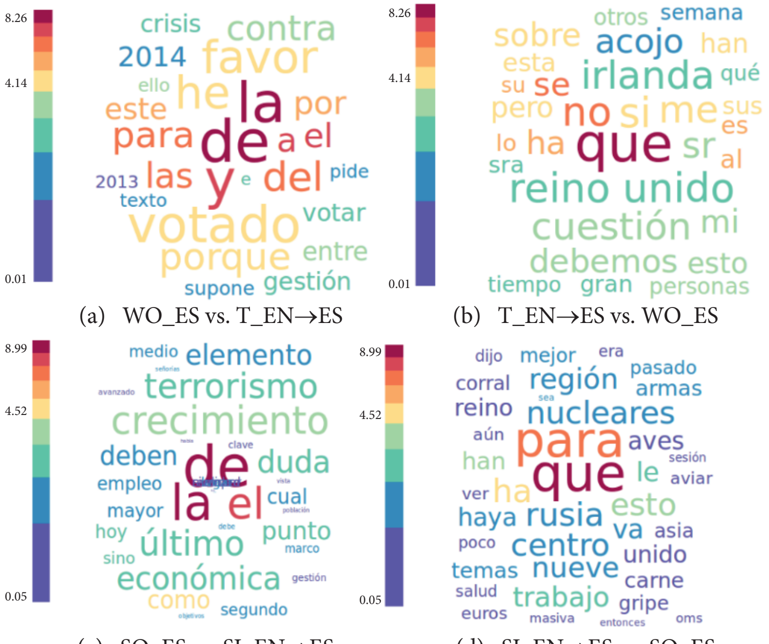
(d) SI_EN→DE vs. SO_DE

Figure 1. Written mode originals vs. translations (a–b) and spoken mode originals vs. interpreting German (c–d)

For Spanish, we observe similar effects when comparing originals and non-originals in both modes (Figure 2). At the lexical level, for both modes, topic-specific differences are again predominant. The topics of interest differ according to the topics that were of high national importance in the different countries. For example, Spanish originals deal with elections and the economic crisis (*votado, crisis, economica, crecimiento*), while the translations deal with national and international relations (*Ireland, reino unido, Rusia, nucleares*).

At the level of grammar, we observe some common distinctive elements for both models. The words that most frequently distinguish between originals and non-originals are *de* for originals and *que* for translations and interpretations. *Que*, as a relative pronoun, points towards more subordinate structures for translations and interpretations. On the other hand, for originals, *de*, in combination with articles such as *el* and *la*, is an indication of more complex main clauses, with longer noun phrases (noun + *de* + article + noun). This observation is corroborated by the presence of the coordinating conjunctions *y* and *e* in the written originals (Figure 2a).

Specifically for the spoken mode (Figures 2c and 2d), the spoken originals seem to contain more elements of formal language. An example is *deben* for originals, compared to *ha/han que* for simultaneous interpretations. An indication of the



(a) WO_ES vs. T_EN→ES

(b) T_EN→ES vs. WO_ES

(c) SO_ES vs. SI_EN→ES

(d) SI_EN→ES vs. SO_ES

Figure 2. Written mode originals vs. translations (a–b) and spoken mode originals vs. interpreting Spanish (c–d)

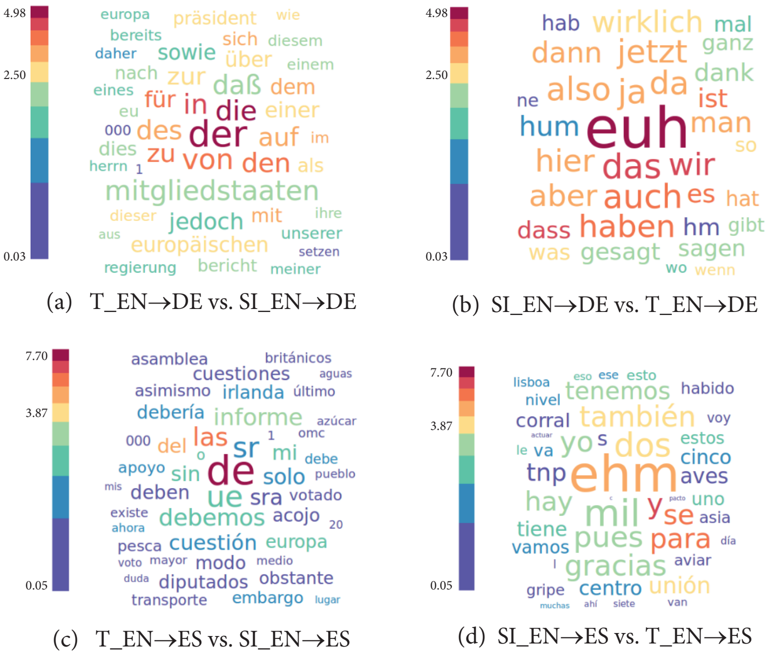
effect of time constraints on the interpreting process is given when comparing the length of the words in spoken originals and interpretations. Interpretations are characterised by much shorter words than originals. This suggests that interpreters, because of the limited time available, avoid using long words, possibly replacing nouns (which are often long) with pronouns (*esto*).

Finally, it is worth mentioning that the size of the interpreting corpora is reflected in the colors present in Figures 2c and 2d. The blue color of most words shows that, while highly distinctive, they are not frequent. One of the main issues with applying computational linguistic methods is the data sparsity issue. Several methods, such as distributional semantics (word vectors) or methods based on neural networks, are unable to provide accurate results unless they are fed with large amounts of data (often millions of sentences), therefore they are inefficient for studying the phenomena of interest. Here, we show that relative entropy is a robust measure of distinctiveness of corpora even in cases of very limited data.

4.2 Translation mode: Translation vs. Interpreting

At a second level, we observe differences between translation modes (translation vs. interpreting) for our two language pairs (see Figure 3). For German (Figures 3a and 3b), written translation is characterised by subordinate clauses starting with relative pronouns *der, die* and conjunctions *daß, jedoch*, as well as a number of prepositions *(zu, von, zur, mit, für ect.)*, which suggest a more complex as well as longer syntactic structure compared to interpreting. In addition, the words that are more distinctive for translations are on average much longer. On the other hand, features of spoken discourse are typical for the interpreted texts, such as hesitation markers (*euh, hum*), speech particles (*ja, mal, aber*) and contracted forms (*hab, ne*). Very few lexical items appear in the word cloud for interpreted speech, which points towards more simplification on the lexical level. Structurally, interpreted texts seem to prefer coordinating structures (*also, aber*) and impersonal forms (*man*).

For Spanish, apart from topical differences as seen for German above, differences in lexical choice are present (Figures 3c and 3d). Translated texts preserve forms of address (*Sr, Sra*), while interpreted texts drop them, probably due to time constraints and/or audience expectations; the receivers of the interpreted texts are present in the Parliament and have eye contact to the speaker, thus this information is not required. Another point worth mentioning is the choice of words more commonly used in written vs. spoken discourse, e.g. *debemos* instead of *tenemos* and *hay* (que). This shows that variation in translation should also be investigated on stylistic factors. For interpreted texts, similar distinctive features are shown as for German; speech markers (*ehm, pues*), coordinating conjunctions (*y*), but also



(a) T_EN→DE vs. SI_EN→DE

(b) SI_EN→DE vs. T_EN→DE

(c) T_EN→ES vs. SI_EN→ES

(d) SI_EN→ES vs. T_EN→ES

Figure 3. Differences in translation mode; translation vs. interpreting for two language pairs, English→German (a–b) and English→Spanish (c–d)

a large number of pronouns (*esto, eso, estos*). The latter indicates that interpreters substitute or avoid repetition of nouns, which can be again attributed to time constraints and management of cognitive effort.

5. Summary and discussion

We have proposed a data-driven method for the exploration of comparable corpora in the field of translation studies. The method is based on probabilistic language models (*n*-gram models) combined with an information-theoretic measure, relative entropy, for corpus/model comparison.

Compared to standard corpus-based accounts working on the basis of relative frequencies, the proposed method has the following benefits:

- it is not biased towards particular linguistic features but assists the analyst in detecting and selecting relevant and significant features;
- it is inherently comparative, i.e. comparison is not executed post-hoc (e.g. statistical test on a frequency distribution) but comparison (by relative entropy) is an integral part;

– it picks up low-frequency phenomena and weak signals (compared to traditional corpus-based approaches, which often favor high-frequency bands).

The latter is especially important when we investigate variation in translation which is even more subtle than the translationese signal itself. Also, the method can be combined with other established approaches for the analysis of translation variation and translationese, such as automatic clustering or classification. The method applied here was initially proposed for measuring corpus distinctiveness in general (Fankhauser et al. 2014) and has been shown to be effective in other kinds of corpus-oriented linguistic analysis where comparison plays a central role, including diachronic studies (see e.g. Degaetano-Ortlieb and Teich's (2019) account of the diachronic development of scientific English using KLD). The approach is therefore generic and allows incorporation of any variable of interest, provided it is encoded in the corpus metadata.

Here, we were able to show variation in translation according to two parameters, mode (translation vs. interpreting) and language pair/target language (here: German and Spanish as target languages, English as source) both at the lexical and grammatical level. Among the features we detected are known ones (e.g. interpreting showing strong signals of orality) and new ones (e.g. words distinctive for translation, compared to interpreting, being longer on average). Thus, our approach provides a suitable instrument both for checking scan-by-eye results as well as detecting additional significant features that are otherwise hard to obtain, especially in other than the high-frequency band.

Of course, the utility of the language models we have used here for the analysis of translation is necessarily limited. The simplest word-based *n*-gram models reduce language behavior to bags of words, i.e. they do not consider context. But we know that human language processing rests very much on use in context, not only the preceding context of *n* words of a given word but also the mutual information between words (as in collocations) as well as expectations raised through extra-linguistic context (situation, register). More sophisticated language models (e.g. word embeddings, neural models) may capture more subtle aspects of language use, but are often not immediately transparent. Making such models useful for linguistic research including contrastive linguistics, typology and translation is an ongoing endeavor, as witnessed by active workshop series such as 'Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature' (LaTeCH-CLfL)[6] or 'NLP for similar languages, varieties and dialects (VarDial)'.[7]

[6] https://sighum.wordpress.com/ (last accessed April 14, 2020)

[7] https://sites.google.com/view/vardial2020 (last accessed April 14, 2020)

In future work, we intend to use the features derived with the proposed method for text-based studies in automatic classification (cf. Lapshinova-Koltunski and Zampieri 2018; Rubino et al. 2016), which also offers the possibility of evaluating our models externally – here, by assessing how strong the detected features are in a classification task. In addition, selected features can be used for variationist studies to tease apart and weigh different factors; for a recent study see Szymor (2018) on the choice of perfective vs. imperfective aspect in Polish translations vs. originals.

We are also exploring other methods from information theory, notably Shannon's noisy channel model, which provides an adequate probabilistic basis for a deeper analysis of the effects of translational choices related to *shining through* vs. *normalization* and would enable testing concrete (and sometimes competing) hypotheses, such as e.g. *overrepresentation* vs. *underrepresentation* of target language features or the *gravitational pull hypothesis*, which tries to explain translationese effects on the basis of general cognitive processes (Halverson 2003). Furthermore, we currently explore approaches from distributional semantics for our larger corpora, which provide a stronger notion of contextual conditions of use than the language models we employed in the present study. For example, bilingual word embeddings can provide insights into the semantic similarity across languages (Östling and Tiedemann 2017; Zou et al. 2013), which can then be taken into account as another factor impacting on translation output. Using such techniques promises to shed light on semantic effects arising from translation, as described for instance for the case of inchoative verbs in Dutch by de Sutter et al. (2012), observing that semantic fields of translated language seem to be less differentiated than those of original language. Finally, from a computational perspective, the features detected by the proposed method can serve as a basis for tuning computational models and/or incorporating linguistic knowledge with a view to creating more comprehensive models of human translation.

Acknowledgements

We are grateful to Stefan Fischer for providing the word cloud visualizations.

Funding

This research is funded by the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies: Implications and Applications". In: *Text and Technology: In honour of John Sinclair*. Ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli. Amsterdam, Netherlands: John Benjamins Publishing Company, pp. 233–252. https://doi.org/10.1075/z.64.15bak

Bendazzoli, Claudio, and Annalisa Sandrelli. 2005. "An Approach to Corpus-Based Interpreting Studies: Developing EPIC (European Parliament Interpreting Corpus". MuTra2005 – Challenges of Multidimensional Translation: Conference Proceedings.

Baroni, Marco, and Silvia Bernardini. 2006. "A new approach to the study of Translationese: Machine-learning the difference between original and translated text". *Literary and Linguistic Computing*, 21(3):259–274. https://doi.org/10.1093/llc/fqi039

Bernardini, Silvia, Adriana Ferraresi and Maja Miličević. 2016. "From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective". *Target* 28: 61–86. https://doi.org/10.1075/target.28.1.03ber

Bernardini, Silvia, Adriana Ferraresi, Mariachiara Russo, Camille Collard and Bart Defrancq. 2018. "Building Interpreting and Intermodal Corpora: A How-to for a Formidable Task". In: *Making Way in Corpus-based Interpreting Studies*. Ed. by Mariachiara Russo, Claudio Bendazzoli and Bart Defrancq. Singapore: Springer. pp. 21–42. https://doi.org/10.1007/978-981-10-6199-8_2

Chesterman, Andrew. 2004. "Beyond the particular". In: *Translation Universals – Do they exist?* Ed. by Mauren, Anna and Kujamäki, Pekka. Benjamins Translation Library, 48(vi): 224. https://doi.org/10.1075/btl.48

Crocker, Matthew, Vera Demberg and Elke Teich. 2016. "Information Density and Linguistic Encoding (IDeaL)". *KI – Künstliche Intelligenz*, 30(1): 77–81. https://doi.org/10.1007/s13218-015-0391-y

Defrancq, Bart. 2015. "Corpus-based research into the presumed effects of short EVS". In: *Interpreting* 17.1: 26–45. https://doi.org/10.1075/intp.17.1.02def

Degaetano-Ortlieb, Stefania and Teich, Elke. 2018. "Using relative entropy for detection and analysis of periods of diachronic linguistic change". In: *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, COLING 2018, Santa Fe, NM, USA.

Degaetano-Ortlieb, Stefania and Elke Teich. 2019. "Toward an optimal code for communication: The case of scientific English". *Corpus Linguistics and Linguistic Theory 2019 aop*. https://doi.org/10.1515/cllt-2018-0088

De Sutter, Gert, Isabelle Delaere and Koen Plevoets. 2012. "Lexical Lectometry in Corpus-Based Translation Studies. Combining Profile-Based Correspondence Analysis and Logistic Regression Modeling." In: *Quantitative Methods in Translation Studies*. Ed. by Michael Oakes and Meng Ji, pp. 326–346. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.51.13sut

Fankhauser, Peter, Jörg Knappen and Elke Teich. 2014. "Exploring and Visualizing Variation in Language Resources". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Gellerstam, Martin. 1986. "Translationese in Swedish novels translated from English". In: *Translation Studies in Scandinavia: Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT)*. Ed. by Lars Wollin and Hans Lindquist. Lund, Sweden: CWK Gleerup, pp. 88–95.

Halverson, Sandra. 2003. "The Cognitive Basis of Translation Universals." *Target* 15(2): 197–241. https://doi.org/10.1075/target.15.2.02hal

Hareide, Lidun. 2019. "Comparable parallel corpora: A critical review of current practices in corpus-based translation studies". In: *Parallel Corpora for Contrastive and Translation Studies. New resources and applications*. Ed. by Doval, Irene and M. Teresa Sanchez Nieto. Benjamins, Amsterdam, pp. 19–38. https://doi.org/10.1075/scl.90.02har

Hughes, James M., Nicholas J. Foti, David C. Krakauer and Daniel N. Rockmore. 2012. "Quantitative patterns of stylistic influence in the evolution of literature". *Proceedings of the National Academy of Sciences* 109(20). 7682–7686. https://doi.org/10.1073/pnas.1115407109

Jurafsky, D. and Martin, J. H. 2008. *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing*. Upper Saddle River, NJ: Prentice Hall.

Kajzer-Wietrzny, Marta. 2012. "Interpreting universals and interpreting style". PhD thesis. Adam Mickiewicz University, Poznań, Poland.

Karakanta, Alina, Mihaela Vela and Elke Teich. 2018. "Preserving Metadata from Parliamentary Debates". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Klingenstein, Sara, Tim Hitchcock and Simon De Deo. 2014. "The civilizing process in London's Old Bailey". *Proceedings of the National Academy of Sciences* 111(26). 9419–9424. https://doi.org/10.1073/pnas.1405984111

Koehn, Philipp. 2005. "Europarl: a parallel corpus for statistical machine translation". In: *Proceedings of the Tenth Machine Translation Summit*. Phuket, Thailand: Asia-Pacific Association for Machine Translation, pp. 79–86.

Koppel, Moshe and Noam Ordan. 2011. "Translationese and its dialects", In: *Proceedings of Conference of the Association for Computational Linguistics (ACL)*, Portland, Oregon, pp. 1318–1326.

Lapshinova-Koltunski, Ekaterina and Marcos Zampieri. 2018. "Linguistic features of genre and method variation in translation: a computational perspective". In: *The Grammar of Genres and Styles: From Discrete to Non-Discrete Units*. Ed. by Dominique Legallois, Thierry Charnois, and Meri Larjavaara. Berlin, Boston: De Gruyter Mouton, pp. 92–117. https://doi.org/10.1515/9783110595864-005

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli and Mariachiara Russo. 2005. "Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus. *Meta*, 50 (4). https://doi.org/10.7202/019850ar

Östling, Robert and Jörg Tiedemann. 2017. "Continuous multilinguality with language vectors". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 644–649. https://doi.org/10.18653/v1/E17-2102

Rabinovich, Ella and Shuly Wintner. 2015. "Unsupervised Identification of Translationese". In: *Transactions of the Association for Computational Linguistics* 3: 419–432. https://doi.org/10.1162/tacl_a_00148

Rubino, Raphael, Ekaterina Lapshinova-Koltunski and Josef van Genabith. 2016. "Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. Association for Computational Linguistics. ACL. https://doi.org/10.18653/v1/N16-1110

Sandrelli, Annalisa and Claudio Bendazzoli. 2005. "Lexical Patterns in Simultaneous Interpreting: A Preliminary Investigation of EPIC (European Parliament Interpreting Corpus)". In: *Proceedings from the Corpus Linguistics Conference Series 1*. Birmingham, UK: University of Birmingham.

Shlesinger, Miriam and Noam Ordan. 2012. "More spoken or more translated? Exploring a known unknown of simultaneous interpreting". In: *Target* 24(1):43–60. https://doi.org/10.1075/target.24.1.04shl

Szymor, Nina. 2018. "Translation: universals or cognition? A usage-based perspective". In: *Target* 30(1):53–86. https://doi.org/10.1075/target.15155.szy

Teich, Elke. 2003. *Cross-linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter. https://doi.org/10.1515/9783110896541

Teich, Elke, José Martínez Martínez and Alina Karakanta. 2020. "Translation, information theory and cognition". In: *Routledge Handbook of Translation and Cognition*. Ed. by Fabio Alves and Arnt Lykke Jakobson. London: Routledge, pp. 360–375. https://doi.org/10.4324/9781315178127-24

Zou, Will Y., Richard Socher, Daniel Cer and Christopher D. Manning. 2013. "Bilingual Word Embeddings for Phrase-Based Machine Translation". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1393–1398.

Binomial adverbs in Germanic and Romance languages

A corpus-based study

Johannes Graën[1,2] and Martin Volk[3]
[1]Pompeu Fabra University / [2]University of Gothenburg /
[3]University of Zurich

As a special type of multiword expressions, binomials are a frequent phenomenon in many languages. We focus on binomial adverbs that are coordinations of two adverbial constituents. Their syntactic contribution to a sentence is adverbial as well and their semantic contribution is idiomatic. They have many uses, such as to intensify (*first and foremost*), express tendency (*more and more*), frequency (*over and over [again]*), vagueness (*more or less*), determination (*sooner or later*) etc.

In this work, we describe our approaches to identify binomial adverbs in a large multiparallel corpus. Alongside the well-known measure of reversibility, we also calculate measures of statistical association and look for single-word translation equivalents in other languages. Combining these features facilitates the identification of binomial adverbs.

Keywords: binomials, binomial adverbs, reversibility score, multi-word expressions, association measures

1. Introduction

Binomials have been studied in linguistics for many years. They are constructions of the type ⟨X conjunction Y⟩, where the words X and Y are typically of the same part of speech. Examples for English are *safe and sound, all or nothing, milk and honey, drag and drop*. Binomial adverbs are those binomials where both X and Y are adverbs, or where the full binomial is used as an adverb. Examples for English are *first and foremost, more or less, on and off*. Most other constructions with adverbial use are prepositional phrases (e.g. *à petit feu* 'slowly', *from tip to toe, en un abrir y cerrar de ojos* 'in the twinkling of an eye', *di punto in bianco* 'out of nowhere',

https://doi.org/10.1075/btl.158.13gra

a contrecœur 'unwillingly'), for a detailed overview of constructions in Italian see (Voghera 2004).

Binomial adverbs cover the whole spectrum from purely compositional (e.g. *sooner or later*) to fully idiomatic (e.g. *by and large*).[1] On the surface, binomials often display alliteration (e.g. *there and then*) or rhyme (e.g. *wear and tear*). Semantically they are often used for emphasis (e.g. *here and now*) or contrast (e.g. *within and outside*). A special subclass of binomials are repetitions (also called echoics) as e.g. *more and more, on and on, little by little*.

Our goal is to detect binomial adverbs automatically. We are particularly interested in idiomatic binomial adverbs so that we can recognize and treat them as units in natural language processing. We use small manually annotated corpora (treebanks) to get an overview. And we use large automatically tagged and lemmatized corpora for large-scale investigations. We exploit both monolingual corpora as well as cross-lingual features from translations. We focus on the Germanic languages (English, German, Swedish) and on the Romance languages (French, Italian, Spanish) and search for differences between these languages with regard to binomial adverbs.

We will describe our corpus preparation in Section 3 and our method for the identification and the ranking of the binomial candidates in Sections 4 and 5. This is followed by discussions on automatically detecting the boundaries of binomial expressions (Section 6) and using translation correspondences for determining idiomaticity (Section 7). But first let's get a first glimpse at binomial adverbs by investigating a French treebank.

Constituent structure treebanks mark binomial adverbs as adverbial phrases (here compound adverbs). An example from the French Le Monde treebank is depicted in Figure 1. This means we can target our search based on the phrase labels in the treebank.

This treebank comprises 20,500 sentences with 630,541 tokens. The most frequent candidate for a binomial adverb is *plus ou moins 'more or less'* with 22 occurrences. Among the top candidates we find idiomatic binomials like *bel et bien 'indeed', pure et simple 'purely and simply', tôt ou tard 'sooner or later'*.

[1] "[W]hich on the surface is a coordination of a preposition and an adjective" (Constant and Nivre 2016), but most closely resembles a coordination of adverbs that acts as an adverb.

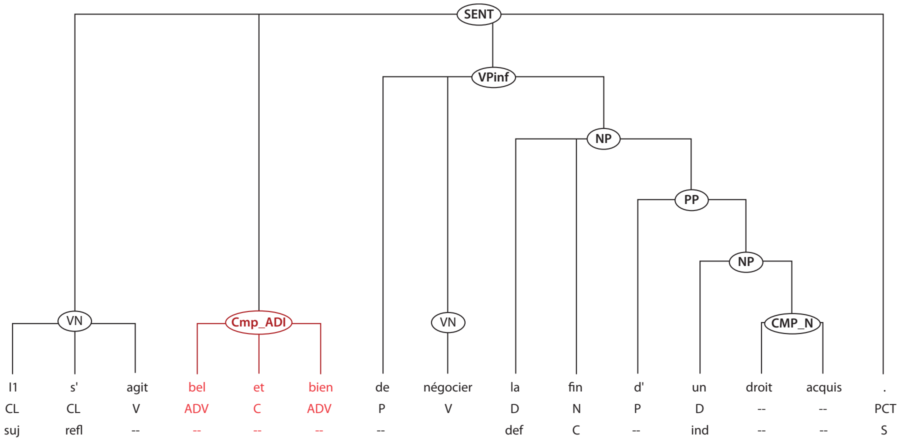


Figure 1. Syntax tree with French binomial (*bel et bien 'indeed'*) from the Le Monde treebank. The binomial is annotated as compound adverb (CMP_ADV). The sentence translates as: *Indeed, it is a matter of negotiating the end of an acquired right*

2. Related work

Early studies on binomials include (Malkiel 1959), an exploratory work on irreversibility of – predominantly English – binomials, and (Bendz 1965), a monograph on "word pairs" in Swedish with comparisons to Danish, English and German. Bendz deals not only with adverbs but all kinds of coordinated word pairs. He presents a semantic classification to distinguish the binomials into:

1. opposition pairs (e.g. English: *sooner or later, to and fro*, German: *dick und dünn, weit und breit*, Swedish: *tjockt och tunt, vitt och brett*)
2. enumeration pairs (e.g. Swedish: *män, kvinnor och barn; tid och rum*), but no examples for adverbs
3. synonym pairs (e.g. English: *first and foremost, simply and solely*, German: *frank und frei, ganz und gar*, Swedish: *blott och bart, helt och hållet*)

Bendz discusses many aspects of word pairs such as inheritance of the constructions from Latin, their prominence in the literature, but also formal properties such as alliteration, assonance and rhyme.

For German, Lambrecht (1984) investigates nominal binomials, i.e. conjoined nouns such as *Recht und Ordnung 'law and order', Kopf oder Zahl 'heads or tails', Vater und Sohn 'father and son'*. Müller (1997) presents a detailed study of German binomial constructions with *und 'and'* as conjunction (e.g. *Fug und Recht* *'justifiably', schalten und walten 'to bustle around', samt und sonders 'the whole lot'*). He is particularly interested in order constraints, which he regards as a defining feature of binomial constructions.

The ordering aspect is also discussed by Copestake and Herbelot (2011) for English binomials, with a text generation perspective. Masini (2006) discusses binomials in Italian retrieved from a corpus by means of pattern matching. The most comprehensive study of English binomials is arguably (Mollin 2014). Mollin performs a study on the British National Corpus (BNC), a 100-million-word corpus collected in the 1990s. She relies on automatic part-of-speech tagging and searches for pairs of the same parts of speech like ⟨noun *and* noun⟩, ⟨verb *and* verb⟩, ⟨adverb *and* adverb⟩ etc. She investigated all candidates where the more frequent sequence occurs 50 times or more. This resulted in 544 binomial types. Mollin judged the candidates by a so-called irreversibility score, which is a ratio of frequencies, the more frequent order ($f(X,C,Y)$) against the frequency of both orders:

$$\text{irr-score}(X, C, Y) = \frac{f(X, C, Y)}{f(X, C, Y) + f(Y, C, X)}$$

For example, *more and better* occurs 249 times in this order in our corpus, and it occurs 180 times in the opposite order *better and more*. This results in a low irreversibility score of 0.58. In contrast *by and large* occurs 161 times in our corpus, whereas the opposite order *large and by* does not occur at all. The irreversibility score is thus at its maximum 1, which is a first indicator for idiomatic usage. Mollin reports the highest irreversibility scores of 100% for the binomial adverbs *back and forth, out and about, today and tomorrow*, all of which were only found in the given order in the BNC.

Kopaczyk and Sauer (2017a) is a collection of papers on binomials in historical varieties of English. In (Kopaczyk and Sauer 2017b, Section 1.3.2), the editors provide an overview of literature on binomials. One of the interesting findings on binomials is that the preferred order can change over time. Another one is that no discrete set of properties has been identified that unequivocally determine the character of a binomial. It is rather a multitude of diverse factors from semantics and phonology that interact, but also cultural aspects and genre of a text play a role in defining the preferential order of coordinated pairs and the effect evoked by their reversal.

We have investigated multi-word adverbs for German and their impact on part-of-speech tagging accuracy and the re-combination of separated verb prefixes to their respective verbs (Volk et al. 2016). Since some separated verb prefixes are homographs with prepositions and also used in binomial adverbs (as e.g. *ab und zu 'from time to time', nach und nach 'gradually'*), it is important to identify the binomial adverbs in order to avoid confusion with separated verb prefixes and to prevent subsequent erroneous verb lemmas and syntax structures.

In (Volk and Graën 2017), we present an extended study on binomial adverbs for English, German, and Swedish and assessed their impact on syntactic parsing and machine translation. The results for machine translation of the binomials between these languages were surprisingly good. Obviously, modern machine translation systems can learn the translations of these fixed multi-word expressions well. On the contrary, parsing results were mixed and sometimes bad.

3. Corpus preparation

For our experiment, we use the Sparcling corpus (Graën et al. 2019), originally labeled FEP9 (Graën 2018). This corpus comprises parallel texts in 16 languages from CoStEP (Graën, Batinic, and Volk 2014), a cleaned version of the Europarl corpus (Koehn 2005) with the transcripts of European Parliament sittings for a time span of 15 years. CoStEP has around 40 million tokens each for the long-standing EU languages like English, French and German, and it has around 9 million each for the new member languages like Estonian, Polish, and Slovenian. Unlike legislative text collections, transcribed parliamentary debates bring about the advantage of being less restricted in terms of wording and lexical choice. The intermediary transcription step, however, may involve the replacement of colloquial expressions with a more formal alternative. For a comparison of the original speeches and the transcription see (Callegaro 2017, Section 4.2).

We first perform tokenization and sentence segmentation on all selected languages with our own adaptable tokenizer (Graën, Bertamini, and Volk 2018). The number of tokens in the Sparcling corpus totals more than 40 million for each language that we are looking at (English, French, German, Italian, Spanish), except for Swedish, which – for reasons of corpus design – has only 36 million tokens (see Graën 2018, Table 3.1). Further annotation consists of part-of-speech tagging and lemmatization with Stagger (Östling 2012) for Swedish and TreeTagger (Schmid 1994) for all other languages. As the available language models use different tagsets, we map all part-of-speech tags to the universal tagset (Petrov, Das, and McDonald 2012), so that we can use the same categories in all languages for our corpus queries. The following is an example sentence after part-of-speech (PoS) tagging and lemmatization. It has universal part-of-speech tags and English-specific part-of-speech tags (following the Penn treebank tag set). Note that *by* was erroneously tagged as a preposition which illustrates the need to recognize the binomial adverb as a unit in order to enable the correct processing in subsequent steps.

Token	**Lemma**	**Universal PoS**	**Penn PoS**
The	the	DET	DT
needs	need	NOUN	NN
of	of	ADP	IN
all	all	DET	PDT
those	those	DET	DT
requiring	require	VERB	VBG
international	international	ADJ	JJ
protection	protection	NOUN	NN
are	be	VERB	VBP
,	,	.	,
by	by	ADP	IN
and	and	CONJ	CC
large	large	ADJ	JJ
,	,	.	,
the	the	DET	DT
same	same	ADJ	JJ
.	.	.	SENT

We also analyze the sentences syntactically (with the help of a dependency parser), but for our investigation of binomial adverbs we abstain from using automatically obtained syntactic relations as we found them to be particularly unreliable for our phenomenon of interest (Volk and Graën 2017).

CoStEP comes with multilingual document alignment. We used string similarity of speaker names to identify matching speaker contributions (Graën, Batinic, and Volk 2014). This way, we aligned all speaker contributions, i.e. our documents, between all available languages. After the segmentation of documents into sentences, we aligned the respective sentence sequences multilingually (see Graën 2018, Section 4.3.1). The sentence alignment identifies the translation correspondences across the languages, for instance:

- English: The needs of all those requiring international protection are, by and large, the same.
- French: Les besoins de toutes les personnes nécessitant une protection internationale sont plus ou moins analogues.
- German: Die Bedürfnisse aller Personen, die internationalen Schutz benötigen, sind mehr oder weniger vergleichbar.
- Italian: Le esigenze di coloro che richiedono protezione internazionale sono in linea di massima le stesse.
- Spanish: Las necesidades de todos aquellos que requieren protección internacional son, en general, las mismas.

- Swedish: Behoven för alla människor som behöver internationellt skydd är mer eller mindre likartade.

Bilingual sentence alignment units, i.e. the entities of corresponding sentences between two languages, were then fed to four different word aligners. We regard a binary link between two tokens in a parallel sentence as reliable alignment if at least three of the aligners agree with this decision in both directions. With this algorithm we find that the English binomial expression *by and large* in the above example sentences has been translated as German *mehr oder weniger*, Spanish *en general*, French *plus ou moins*, Italian *in linea di massima*, and Swedish *mer eller mindre*.

4. Identification of candidates

To find candidates for binomial adverbs, one option is to search for the corresponding pattern of (universal) part-of-speech tags, i.e. ⟨ADV CONJ ADV⟩. Unlike manually annotated treebanks, however, corpora that are automatically tagged with statistical part-of-speech taggers have proven unreliable when it comes to correct part-of-speech tagging of binomial adverbs (Volk et al. 2016; Volk and Graën 2017). We have observed this already with *by and large* in the above example sentence.

In order to reduce the impact of incorrect part-of-speech tags in this pattern, we first collect all words that are tagged as adverbs anywhere in our corpus. For example, for English we find 2.38 million tokens in our corpus with the part-of-speech tag ADV (which is about 5.5% of all 43.1 million tokens in the English corpus). These adverb tokens correspond to 2354 unique adverbs (lower cased), the most frequent ones being *not, also, as, very, so*. Interestingly, 1812 of the 2354 English adverbs end in *-ly* which indicates that they have been derived from adjectives. One could argue that this leaves around 500 true adverbs. This count includes spelling variations (*therefore, therefor*), comparative forms (*closer, closest*), ordinal numbers (*sixth, seventh, ninth*), and a few annoying part-of-speech errors (*country, fist, plum*). Our count excludes hyphenated adverbs (such as *half-way, self-evidently, whole-heartedly*, 536 tokens in our corpus) that are unlikely to function in binomials. And of course, our count also misses adverbs that are consistently tagged as other part of speech.

For comparison, in our parallel German corpus we find 1.76 million tokens that are tagged as ADV. They amount to only 867 unique adverbs. This is due to the fact that German adjectives are not morphologically marked when they are used as adverbs. Thus, they are more difficult to identify by a part-of-speech tagger.

In a next step, we check whether these adverbs occur conjoined in the corpus. This means that we find candidates for binomial adverbs even if they are incorrectly tagged in the conjoined construction. This makes our search more robust and increases the recall. We do miss a binomial adverb candidate if one of its conjuncts is never tagged as an adverb in the whole corpus. This can happen when the words typically belong to a different part of speech, e.g. if they typically are prepositions, as e.g. *by, on, over*. We made sure that these are also considered.

Based on the bigram and trigram frequencies over the binomial candidate we compute different collocation scores in order to estimate the fixedness of the binomial as a proxy for idiomaticity.

5. Candidate ranking

A common way to approach the identification of lexical units for which no accurate binary classification is possible is to rank the items found such that the good ones appear at the top and the bad ones, that is typically a large number of false positives, somewhere in a long tail. We tried to obtain lists of good matches from native speakers and speakers with a high proficiency in one of the six languages, in order to evaluate our rating. This idea admittedly did not work out, as numerous cases turned out to be uncertain.

We may use the irreversibility score to rank candidates, but we may also use other collocation measures like the mutual information score. Mutual information measures the amount of information (in bits) that one of two parts contains about the other. For the pattern ⟨X C Y⟩ with X and Y being the adverbs and C the conjunction, the mutual information is calculated on the frequencies of the whole expression ($f(X,C,Y)$) and the two parts consisting of one adverb and the conjunction ($f(X,C)$ and $f(C,Y)$). N is the size of the corpus and renders the absolute frequencies (f) as relative ones:

$$\mathrm{MI}(X, C, Y) = \log_2 \frac{N \cdot f(X, C, Y)}{f(X, C) \cdot f(C, Y)}$$

For example, we use the corpus frequency of the triple *first and foremost* and contrast it to the tuple frequencies for *[first and]* and *[and foremost]*. If both tuple frequencies are close to the triple frequency, then this means that they frequently occur as a triple (and seldom alone) and therefore the triple is a good candidate for an idiomatic expression.

In this way we get high scores for candidates such as *inward and outward*, which means that *[inward and]* is a good predictor of *outward* as continuation and, the other way round, *[and outward]* predicts *inward* as preceding token.

The mutual information score, however, does not take into consideration the overall frequency of an expression, and thus tends to favor rare candidates such as *inward and outward, wittingly or unwittingly* and *sympathetically and favourably*. Figure 2 shows that rare candidates are among the top ranks in many languages,

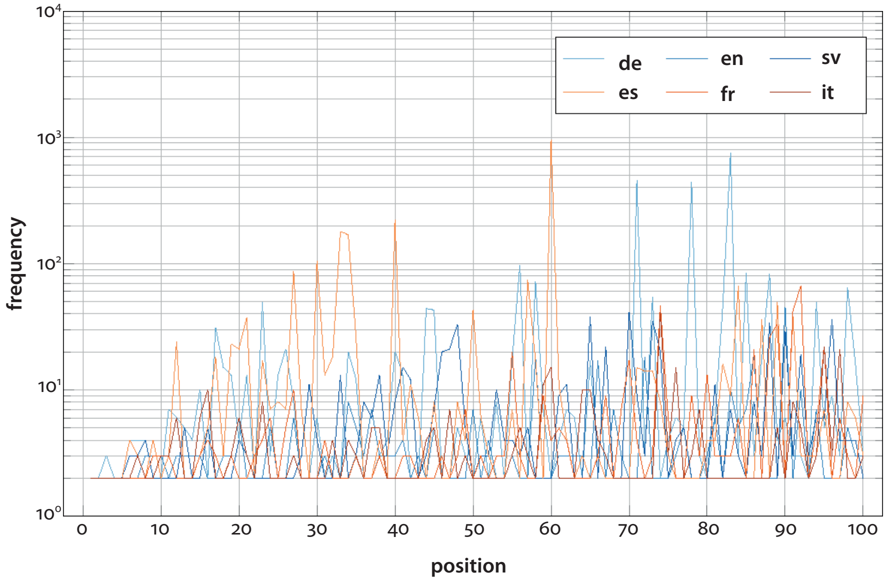


Figure 2. The top-100 candidates by mutual information ranking and their frequency in all six languages

whereas frequent candidates (the peaks in the diagram) are spread throughout the 100 positions.

An extension of this measure is the so-called local mutual information score, which is defined as the product of frequency and the mutual information measure:

$$\text{local-MI}(X, C, Y) = f(X, C, Y) \cdot \log_2 \frac{N \cdot f(X, C, Y)}{f(X, C) \cdot f(C, Y)}$$

The result is a list that combines the predictability of the expression and its frequency. We get a ranking that is similar at the top of the list, when we use another measure, the simple log-likelihood, to rank the candidates. Both association measures are described in detail in (Evert 2008). The simple log-likelihood measure comprises the main part of the local mutual information score (the base of the logarithm can be varied without affecting the ranking), which is altered by the difference of how often we observe the expression ⟨X C Y⟩ compared to how often we expect to observe it, based on the frequencies of ⟨X C⟩ and ⟨C Y⟩ and the assumption of independence:

$$\text{simple-loglikelihood}(X, C, Y) = 2\Big(f(X, C, Y) \cdot \log \frac{N \cdot f(X, C, Y)}{f(X, C) \cdot f(C, Y)} - \Big[f(X, C, Y) - \frac{f(X, C) \cdot f(C, Y)}{N}\Big]$$

The ranking of candidates according to the simple log-likelihood measure is shown in Figure 3. Very frequent candidates occupy the first ranks, but we also find some high-ranked candidates with few occurrences (e.g. *ora come ora* '*right now*' at position 27 with 33 occurrences).

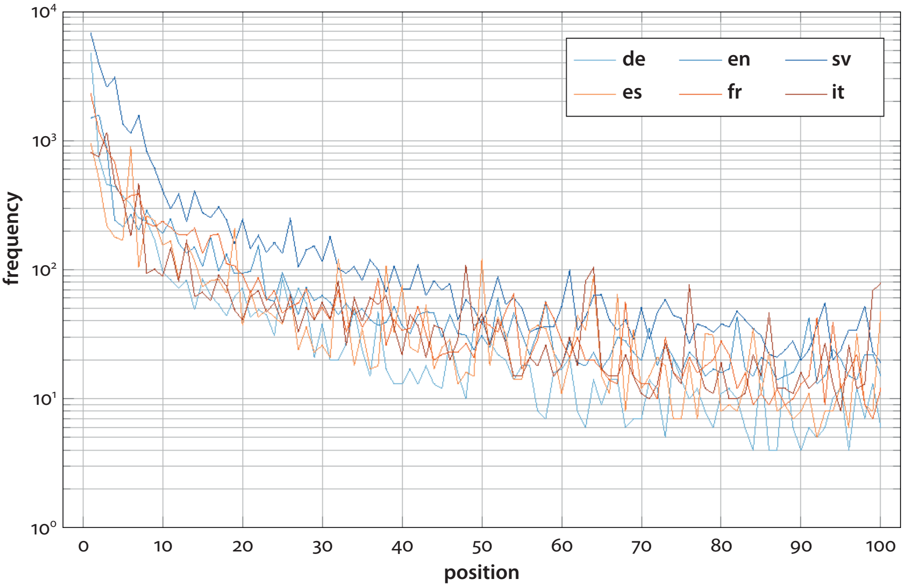


Figure 3. The top-100 candidates by log-likelihood ranking and their frequency in all six languages

6. Expression boundaries

Sometimes binomials are parts of larger fixed expressions. For example *over and over* is often part of the expression *over and over again* (in 127 out of 150 occurrences in our corpus). Therefore, we also check the boundaries of the binomial candidate by calculating the entropy to the immediate left and right of the identified binomial candidate. Entropy as a measure of how expected (or unexpected) a certain event is given other observed events has been defined by Shannon (1948) as:

$$H = -\sum_i p_i \cdot \log p_i$$

The probability *p* in our case is the conditional probability of a particular word W preceding or following the expression ⟨X C Y⟩, i.e.:

$$p = \frac{f(W, X, C, Y)}{f(X, C, Y)} \text{ or } p = \frac{f(X, C, Y, W)}{f(X, C, Y)}$$

If entropy at the left or right boundary is low (which means that there is little variation regarding the word preceding or following the binomial construction), then we take this an indication to extend the expression by one word to the left or to the right. In other words, if the binomial candidate (e.g. *over and over*) is often preceded or followed by the same word (here: followed by *again*), then this word must be included in the multi-word expression. Experiments with a window of two words suggest that most complete expressions can be captured with looking at a single word at the left and right context.

We experimentally determined that an entropy of 0.2 (the lowest value of 0 is obtained only for cases with a single alternative, i.e. without variation) is a good conservative threshold to tell apart words that belong to the expression from words that happen to frequently border the expression. In addition, we require both the preceding and following words to show an absolute frequency of 5 and a relative one of 25% to be eligible for extending the binomial adverb expression. That way, we find extensions for approximately 4% of the candidates.

One category of extended binomial adverbs comprehends those cases where the coordination is divided into two parts. This is the case for split negation particles, which occur predominantly in Romance languages. The most frequent example is *ni plus ni moins, né più né meno, ni más ni menos 'nothing more and nothing less'*. Similar discontinuous expressions include *tant … que … 'both …and …'* as in *tant bien que mal 'after a fashion'*, and *både … och … 'both … and …'* as in *både länge och väl 'for some time'*.

Other cases that are similar in several languages are *over and over again, immer und immer wieder, om och om igen* and *vaut tard que jamais, meglio tardi che mai, bättre sent än aldrig 'better late than never'*. There are many more cases of adverbial expressions that do not fit into the schema ⟨(something) adverb conjunction adverb (something)⟩ (see Masini 2006). We do not intent to find all those, but rather limit ourselves to pure binomial adverbs, indicating whenever they are potentially part of a larger expression.

7. Translation variants

With the irreversibility score, the mutual information scores and the simple log-likelihood measure we want to capture the idiomaticity of a candidate binomial. We are interested in whether the candidate is still compositional or rather must be regarded as a unit with a lexicalized meaning. The ranking from the above scores percolates many idiomatic binomials to the top, but among the true positives the ranked list still contains many unsuspicious expressions that are not idiomatic at all.

Therefore, we also checked the translation correspondences for our candidates. The idea is: If a binomial adverb has been translated with a single adverb in another language, then this is evidence for its idiomaticity.

Since our corpus comprises automatically computed word alignments between all languages, we can look up the translations of each binomial candidate. We are particularly interested in those cases, where the whole expression is translated to a single word in some other language. We know that content words (i.e. nouns, verbs, adjectives and adverbs) are more reliably aligned than function words (Graën 2018). This includes erroneous alignments of conjunctions, and so we search for cases where both adverbs in a candidate expression are aligned to the same word in the target language. In the case of the German candidate *nach und nach*, the alignment of both tokens *nach* to a single word in another language (e.g. English *gradually*, French *progressivement*, or Swedish *gradvis*) makes the expression a promising candidate.

We select the most frequent word if it represents at least 10% of the identified translation variants. Additionally, we indicate the most frequent part of speech of that word and its frequency. This gives us strong evidence for some of the candidates to be binomial adverbs, yet, that way, we only find a single-word translation variant for about 6% of the candidates. What is more, the method does not work well with repetitions (like *on and on*).

For the German candidate *nach wie vor*, we find the single-word translations *still* in English (in 99% of the occurrences tagged as adverb), *sigue 'continues'* in Spanish (in all cases tagged as a verb) and *fortfarande* in Swedish (in all cases tagged as an adverb). In Swedish, we find the English adverb *anywhere* and the Italian adverb *ovunque* for the candidate *var som helst*. In case the binomial adverb serves as intensifier, we frequently find the corresponding (regular) adverb in other languages, such as the Italian *semplicemente* for English *purely and simply* or French *purement et simplement*.

8. Conclusions

The identification and special processing of binomials is an important step for many language technology applications. This is particularly true for binomial adverbs that consist of homographs with other parts of speech.

In this paper, we have described our efforts to automatically identify binomial adverbs and their fixedness as an approximation of their idiomaticity. We employed established association measures for the identification of multi-word-units such as the mutual information score and log-likelihood ratio. In addition, we used the reversibility score proposed by Mollin (2014). The resulting ranking lists many interesting candidates on the top but does not lead to a fool-proof selection criterion.

An interesting finding was the application of entropy to automatically evaluate the boundaries of binomials as to whether they indicate that the binomial is part of a larger idiomatic expression.

Overall, we see the automatic ranking as a means to identify good candidates for manual inspection. For example, we found 18 out of 21 English binomial adverbs indicated by Mollin (2014), but also several hundreds more.

In the Appendix, we show the top 15 candidates for each language, together with the most important figures and single-word translation variants. The complete lists for all six languages (and several more) can be obtained from http://pub.cl.uzh.ch/purl/binomial_adverbs. The raw parallel corpus is available at http://pub.cl.uzh.ch/purl/PaCoCo.

Acknowledgements

The authors would like to thank Christine Konecny and Ana Bocanegra Valle for discussing binomial adverbs in Italian and Spanish, and Anne Göhring, Elisa Pellegrino, and Mónica Ballesteros for their help in assessing binomial candidates in the Romance languages.

Funding

This research is partly supported by the Swiss National Science Foundation under grant P2ZHP1_184212 through the project "From parallel corpora to multilingual exercises – Making use of large text collections and crowdsourcing techniques for innovative autonomous language learning applications", at Pompeu Fabra University in Barcelona (with Gr@el, Grup de Recerca en Aprenentatge i Ensenyament de Llengües) and at the University of Gothenburg (with Språkbanken Text).

References

Bendz, Gerhard. 1965. *Ordpar*. Stockholm: P. A. Nordstedt & Söners Förlag.

Callegaro, Elena. 2017. "Parallel Corpora for the Investigation of (Variable) Article Use in English: A Construction Grammar Approach." PhD diss., University of Zurich.

Constant, Matthieu, and Joakim Nivre. 2016. "A Transition-Based System for Joint Lexical and Syntactic Analysis." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 161–171. https://doi.org/10.18653/v1/P16-1016

Copestake, Ann, and Aurélie Herbelot. 2011. "Exciting and interesting: issues in the generation of binomials." In *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop*, 45–53.

Evert, Stefan. 2008. "Corpora and collocations." In *Corpus Linguistics. An International Handbook*, edited by A. Lüdeling and M. Kytö, 2:1212–1248. Berlin: Walter de Gruyter.

Graën, Johannes. 2018. "Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning." PhD diss., University of Zurich.

Graën, Johannes, Dolores Batinic, and Martin Volk. 2014. "Cleaning the Europarl Corpus for Linguistic Applications." In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, 222–227. Stiftung Universität Hildesheim, October. https://doi.org/10.5167/uzh-99005

Graën, Johannes, Mara Bertamini, and Martin Volk. 2018. "Cutter – a Universal Multilingual Tokenizer." In *Proceedings of the 3rd Swiss Text Analytics Conference*, edited by Mark Cieliebak, Don Tuggener, and Fernando Benites, 75–81. CEUR Workshop Proceedings 2226. CEUR-WS, June.

Graën, Johannes, Tannon Kew, Anastassia Shaitarova, and Martin Volk. 2019. "Modelling Large Parallel Corpora: The Zurich Parallel Corpus Collection." In *Proceedings of the 7th Workshop on Challenges in the Management of Large Corpora (CMLC)*, edited by Piotr Bański et al.

Koehn, Philipp. 2005. "Europarl: A parallel corpus for statistical machine translation." In *Proceedings of the 10th Machine Translation Summit*, 5:79–86. Asia-Pacific Association for Machine Translation.

Kopaczyk, Joanna, and Hans Sauer, eds. 2017a. *Binomials in the History of English: Fixed and Flexible*. Cambridge University Press. https://doi.org/10.1017/9781316339770

Kopaczyk, Joanna, and Hans Sauer. 2017b. "Defining and exploring binomials." *Binomials in the History of English: Fixed and Flexible*: 1–23. https://doi.org/10.1017/9781316339770.001

Lambrecht, Knud. 1984. "Formulaicity, frame semantics, and pragmatics in German binomial expressions." *Language* 60 (4): 753–796. https://doi.org/10.1353/lan.1984.0004

Malkiel, Yakov. 1959. "Studies in irreversible binomials." *Lingua* 8:113–160. https://doi.org/10.1016/0024-3841(59)90018-X

Masini, Francesca. 2006. "Binomi coordinati in italiano." In *Prospettive nello studio del lessico italiano: Atti del 9. congresso SILFI*, 2:563–571.

Mollin, Sandra. 2014. *The (Ir)reversibility of English Binomials. Corpus, constraints, developments*. Vol. 64. Studies in Corpus Linguistics. John Benjamins. https://doi.org/10.1075/scl.64

Müller, Gereon. 1997. "Beschränkungen für Binomialbildungen im Deutschen." *Zeitschrift für Sprachwissenschaft* 16 (1): 25–51. https://doi.org/10.1515/zfsw.1997.16.1-2.5

Östling, Robert. 2012. "Stagger: A modern POS tagger for Swedish." In *Proceedings of the 4th Swedish Language Technology Conference (SLTC)*.

Petrov, Slav, Dipanjan Das, and Ryan McDonald. 2012. "A Universal Part-of-Speech Tagset." In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, edited by Nicoletta Calzolari et al. Istanbul: European Language Resources Association (ELRA). isbn: 978-2-9517408-7-7.

Schmid, Helmut. 1994. "Probabilistic part-of-speech tagging using decision trees." In *Proceedings of International Conference on New Methods in Natural Language Processing*, 12:44–49.

Shannon, Claude Elwood. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Voghera, Miriam. 2004. "Polirematiche." In *La formazione delle parole in italiano*, 56–69. Max Niemeyer Verlag.

Volk, Martin, Simon Clematide, Johannes Graën, and Phillip Ströbel. 2016. "Bi-particle Adverbs, PoS-Tagging and the Recognition of German Separable Prefix Verbs." In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, 297–305. September.

Volk, Martin, and Johannes Graën. 2017. "Multi-word Adverbs – How well are they handled in Parsing and Machine Translation?" In *Proceedings of The 3rd Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT)*.

Appendix

Prominent left and right context words that potentially belong to the expressions are given in parentheses, frequent single-word translation variants are listed below the tables, and candidates and contexts that we consider questionable to incorrect are underlined.

English

Candidate	*f*	Irr-score	MI	Local-MI	Simple-II
first and foremost[a]	1496	100%	14.48	21667.31	10053.02
more and more	1575	100%	12.42	19560.32	8626.48
more or less	868	100%	15.39	13356.01	6305.12
directly or indirectly	242	98.8%	17.16	4151.52	2015.47
sooner or later	216	100%	17.59	3799.80	1855.71
again and again	267	100%	14.45	3858.79	1789.22
inside and outside	204	94.4%	16.35	3335.10	1599.93
here and now	289	100%	12.29	3552.32	1560.71
up and running	223	100%	14.52	3238.38	1503.70
less and less	191	100%	15.26	2914.86	1372.92
more and better	249	58.0%	12.30	3062.66	1345.91
by and large	161	100%	15.78	2540.55	1207.56
yes or no	135	100%	17.23	2325.43	1130.05
over and over (again)	150	100%	14.55	2182.92	1014.25
north and south	106	97.2%	17.49	1854.05	904.25

[a] (it) innanzitutto

German

Candidate	*f*	Irr-score	MI	Local-MI	Simple-II
nach wie vor[a]	4723	100%	13.04	61577.80	27627.53
mehr oder weniger	746	99.6%	15.45	11527.95	5448.52
ganz und gar[b]	456	100%	16.19	7381.16	3531.90
mehr denn je	442	100%	15.70	6940.83	3294.80

Candidate	*f*	Irr-score	MI	Local-MI	Simple-II
nach und nach[c]	373	100%	14.02	5230.73	2403.21
mehr und mehr[d]	321	100%	13.00	4171.31	1869.38
hier und heute	251	96.9%	14.11	3542.38	1630.72
mehr als nur	238	100%	11.59	2758.60	1184.84
hier und da	171	100%	12.78	2184.52	973.21
ganz oder teilweise	97	94.2%	17.50	1697.90	828.24
hin und wieder	84	100%	15.39	1292.53	610.18
hin und her	72	98.6%	17.38	1251.59	609.53
heute und morgen	83	97.6%	15.28	1268.37	597.63
drittens und letztens	49	100%	20.62	1010.24	510.22
hier und jetzt	85	89.5%	12.51	1062.91	469.94

[a] (en) still; (es) sigue; (sv) fortfarande

[b] (es) totalmente

[c] (en) gradually; (fr) progressivement; (it) gradualmente; (sv) gradvis

[d] (en) increasingly

Swedish

Candidate	*f*	Irr-score	MI	Local-MI	Simple-II
till och med[a]	6802	99.9%	10.78	73308.32	30538.01
först och främst[b]	3978	100%	13.84	55056.29	25191.19
helt och hållet[c]	2610	100%	13.36	34863.36	15769.83
i och med	3098	99.6%	10.95	33909.83	14221.75
hur som helst[d]	1349	100%	13.41	18087.96	8192.03
klart och tydligt[e]	1146	90.8%	14.20	16272.84	7505.23
från och med	1564	100%	10.87	17004.18	7109.54
mer eller mindre	824	99.9%	15.15	12484.70	5868.54
helt och fullt[f]	608	99.7%	13.12	7979.23	3587.98
sist men inte (minst)	412	100%	15.15	6241.90	2934.00
direkt eller indirekt	296	96.7%	16.66	4931.49	2377.05
mer och mer[g]	387	100%	12.67	4903.53	2178.22
förr eller senare	236	100%	17.14	4044.42	1962.98
mer än någonsin	410	100%	10.95	4488.38	1882.27
snabbt och effektivt	276	89.0%	13.09	3613.51	1623.55

[a] (de) sogar; (en) even; (es) incluso; (fr) même; (it) persino

[b] (de) zunächst; (fr) d'abord; (it) innanzitutto

[c] (en) fully; (es) totalmente; (fr) totalement

[d] (it) comunque

[e] (en) clearly; (es) claramente; (fr) clairement; (it) chiaramente

[f] (en) fully; (es) plenamente; (fr) pleinement; (it) plenamente

[g] (en) increasingly

French

Candidate	*f*	Irr-score	MI	Local-MI	Simple-ll
ne soit pas	2316	100%	13.45	31158.38	14127.22
(d') ores et déjà[a]	1187	100%	15.14	17975.58	8448.38
plus ou moins	859	99.8%	15.51	13326.54	6305.37
plus que jamais	683	99.9%	12.45	8499.70	3751.33
purement et simplement[b]	341	100%	16.93	5774.34	2794.50
(mais) aussi et surtout	374	100%	13.44	5026.26	2278.11
pas pour autant	387	100%	11.99	4638.33	2018.56
(ni) plus ni moins	229	100%	17.30	3962.67	1927.76
tôt ou tard	217	100%	17.69	3839.54	1877.63
haut et fort	237	100%	16.46	3902.07	1875.28
directement ou indirectement	213	99.5%	17.49	3725.13	1816.75
(un) tant soit peu	187	100%	17.82	3332.30	1632.24
encore et toujours[c]	185	97.4%	15.79	2920.30	1388.20
notamment parce que	211	100%	10.87	2293.38	958.75
rapidement et efficacement	134	91.8%	14.83	1987.31	928.48

[a] (de) bereits; (en) already; (es) ya; (it) già; (sv) redan

[b] (de) einfach; (en) simply; (it) semplicemente

[c] (en) still; (it) ancora; (sv) fortfarande

Italian

Candidate	*f*	Irr-score	MI	Local-MI	Simple-II
più o meno	809	99.9%	13.58	10981.86	4993.74
più che mai	754	100%	12.94	9753.42	4364.14
anche se non	1154	100%	9.45	10901.50	4257.36
prima o poi[a]	467	100%	16.37	7643.49	3667.84
(ma) anche e soprattutto	358	100%	13.33	4772.87	2157.56
direttamente o indirettamente	183	99.5%	17.52	3206.76	1564.66
meno che non	459	100%	8.79	4036.05	1513.94
(né) più né meno	94	100%	18.28	1718.43	846.60
oggi come oggi	101	100%	16.57	1673.33	805.44
qua e là	89	100%	18.25	1624.31	799.93
così come non	147	100%	11.31	1662.44	706.89
sempre e comunque	82	87.2%	15.47	1268.87	599.94
come se non	169	100%	9.15	1547.01	593.40
solo ed esclusivamente	62	100%	18.21	1129.25	555.88
sì o no	67	100%	16.32	1093.33	524.25

[a] (de) irgendwann

Spanish

Candidate	*f*	Irr-score	MI	Local-MI	Simple-II
más o menos	950	99.8%	15.36	14589.74	6883.90
más que nunca	511	100%	10.45	5340.15	2193.09
dentro y fuera (de)	220	96.1%	16.67	3666.98	1767.74
(ni) más ni menos	179	100%	17.57	3145.69	1535.89
(tanto) dentro como fuera (de)	169	96.0%	17.24	2914.28	1416.57
ya que no	885	100%	5.39	4768.01	1142.63
tarde o temprano	105	100%	18.15	1905.52	937.24
así como también	260	100%	9.24	2401.93	926.10
como si no	241	100%	9.64	2322.97	916.56
como si fuera	155	100%	12.87	1994.03	890.52
aquí y ahora	167	100%	12.069	2015.53	879.47
antes o después	86	100%	18.44	1586.15	782.96
más y más	163	100%	10.57	1723.04	711.37
sí o no	118	100%	13.20	1557.47	701.69
aquí y allá	74	100%	15.51	1147.73	543.00

Index

Corpus-based contrastive and translation research are areas that keep evolving in the digital age, as the range of new corpus resources and tools expands, opening up to different approaches and application contexts. The current book contains a selection of papers which focus on corpora and translation research in the digital age, outlining some recent advances and explorations. After an introductory chapter which outlines language technologies applied to translation and interpreting with a view to identifying challenges and research opportunities, the first part of the book is devoted to current advances in the creation of new parallel corpora for under-researched areas, the development of tools to manage parallel corpora or as an alternative to parallel corpora, and new methodologies to improve existing translation memory systems.

The contributions in the second part of the book address a number of cutting-edge linguistic issues in the area of contrastive discourse studies and translation analysis on the basis of comparable and parallel corpora in several languages such as English, German, Swedish, French, Italian, Spanish, Portuguese and Turkish, thus showcasing the richness of the linguistic diversity carried out in these recent investigations.

Given the multiplicity of topics, methodologies and languages studied in the different chapters, the book will be of interest to a wide audience working in the fields of translation studies, contrastive linguistics and the automatic processing of language.

ISBN 978 90 272 0918 4

9 789027 209184

John Benjamins Publishing Company