The use of corpora and other electronic tools in historical research on translation

'Every new wave of technology, directly or indirectly, affects the translation sector in some tangible way and in so doing alters the course of its history'
(Folaron 2018: 113)

Introduction

Translation history and historiographical approaches to translation have traditionally relied on the insight provided by the historical context and both contextual and paratextual features of the translated texts together with their reception. However, as Pym rightly indicates 'good historiography requires awareness of what translators actually did' (1996:2) and historical sources should be interpreted in an adequate manner. Only by correlating historiographical insights with empirical evidence obtained from the translated texts will it be possible to produce a coherent and sound translation history. In this line of work, technology and digital humanities offer tools to the translation historian which can complement non-computational methods and more traditional approaches to the sources and which can be very beneficial if implemented correctly. This is the main reason why in this chapter we advocate the use of tools such as corpora derived from linguistics to complement the research carried out from a historiographical point of view, while also signalling some of their possible drawbacks or limitations, which should be taken into consideration when undertaking any translation history research project.

Independently of the kind of historical approach the researcher adopts, the use of linguistic tools in researching translation is almost unavoidable nowadays. A socio-cultural approach to translation can benefit greatly from the use of catalogues, corpora and the several linguistic analyses that can be carried out thanks to these. Thus, this chapter will first examine the different tools available to researchers and evaluate their usefulness in historical research, from the compilation of a catalogue to the exploitation of a corpus in its distinct forms. We will also consider some of the issues that can arise when using these tools, starting with the need for the translation historian to develop a sensitivity to the digital environment in which s/he will work: some kind of data analysis training seems essential in order to avoid trivial interpretation of the data at his/her disposition and achieve adequate theorization (Wakabayashi 2019).

Catalogues/Bibliographies/Databases

For the purposes of this discussion (and setting aside their obvious differences) we will consider catalogues, bibliographies and databases to be essentially similar tools, as distinct from a corpus. In the search for solid bases for empirical research, the translation history scholar cannot overlook the existing bibliographies, databases, catalogues or lists. They constitute vast inventories of data which can help in the design of the research questions and that can also be used to compile entries in a corpus, as will be outlined in the next section. Like any other repertoire, they present limitations (see Pym 2009): a careful selection of the documents found in them constitutes a first step in the researcher's path but it must be complemented by a proper understanding and interpretation of the data.

Examples of international databases relevant to translation history include the famous *Index Translationum*, further explained below, and some others such as *HISTRA*, which will be briefly discussed as well.

Index Translationum (http://www.unesco.org/xtrans/)

It is the most complete catalogue of translations in the world and it was launched in 1932 by the Institute for Intellectual Cooperation in Paris,[1] under the auspices of the League of Nations in the form of a quarterly report. When publication ceased in 1941 because of the war, the Index was reporting the translations being published in 14 countries. The Index was revived in new series in 1948, thanks to the UNESCO and was published annually. The first issue of the new series included 8,570 translations published in 26 countries. The *Index* was converted into database format in 1979 (but not including earlier data), and the number of entries rose considerably to 54,447, featuring 54 countries (Pięta 2013: 301). Its database format and the fact that it is available online constitute its biggest advantages over other repertories, although the *Index* stopped being updated in 2012, the year of its 80th anniversary. In a matter of seconds, the researcher can obtain data on translations from 1979 to 2012 in any combination of languages of the participating countries. The entries are distributed by thematic area, in accordance with the Universal Decimal Classification (CDU). One important issue to consider is the different data gathering methods used by each country which leads to inconsistencies in the data (Poupad et al. 2009). It is therefore a good

[1] Ed: see Roig Sanz in this volume for a study of this institute.

starting point for working on an isolated hypothesis dealing with quantitative data on a big scale but it definitely falls short for a more complex analysis where a high degree of precision is required.

HISTRA (HIStory of TRAnslation) catalogue (http://histra.unileon.es/)

The HISTRA (HIStory of TRAnslation) online catalogue aims to transfer to a digital resource a bulky archive of indexed cards which was compiled by Dr. Julio César Santoyo, an expert in the field of Translation History in the Hispanic world. Santoyo indexed the bibliographical references of Spanish translations of works originally written in English from the 16th century until the 1980s. The data it contains is varied and practically impossible to obtain from any other resource, which makes the archive all the more valuable. This project is a work in progress being developed by researchers at the University of León (Spain) and once completed it will offer translation history researchers a rich and versatile database on Hispanic translations and translators. The methodology of HISTRA guarantees exhaustive bibliographic references and the normalization and standardization of the entries, using the international MARC21 (*Machine Readable Cataloguing*) format. The bibliographic entries in the HISTRA database always provide the source author and title of the work in English (information which is sometimes missing in the *Index Translationum*), together with the Spanish title, the place of publication, publishing house and year of publication. The entries also include the physical details of the publication (with photos of the cover when possible), the collection or series in which the book was published (if relevant), the topic/field of the work, the translator's name. All the additional information concerning the entries, either provided by the index cards or discovered during the cataloguing process, is archived in the 'Notes' field. Any relevant external resources, such as open access full text copies, are also linked/attached to the entry. Examples of possible studies based on HISTRA include the analysis of interconnections between authors of the same genre and their translators or that of translators included in the database who also feature as authors of their own works. HISTRA provides a robust platform for the retrieval of information about translators and translated works, thus facilitating research in translation history and studies on literary reception of works translated from English into Spanish.

Apart from the *Index* and HISTRA, there are other databases which are relevant for translation history research, such as: the *Perso-Indica database* for Persian works on

Indian learned traditions (www.perso-indica.net); the *Renaissance Cultural Crossroads Catalogue*, which is a list of all translations out of and into all languages printed in England, Scotland, and Ireland before 1641 hosted by the University of Warwick (UK) (www.dhi.ac.uk/rcc/) (Wakabayashi 2019: 133); the *Intercultural Literature in Portugal 1930-2000: A Critical Bibliography*, a critical bibliography of translated literature published in book-form in Portugal launched in 2007 by the University of Lisbon Centre for English Studies and the Centre for Communication and Culture – Catholic University of Lisbon. Coordinated by Teresa Seruya, Alexandra Assis Rosa and Maria Lin Moniz, it has led to several publications by its researchers and the publication in 2010 of its free online database which covers data from 1930-86, and includes 24,260 entries (translatedliteratureportugal.org).

Ad hoc catalogues

The previously mentioned databases are resources or repertories researchers can resort to when starting a project, but they can also serve as a starting point for the construction of other catalogues. From a methodological point of view, the construction of catalogues in Translation Studies serves to constitute and organize evidence from which to build and support the later textual study. Sometimes the translation history researcher needs to build an *ad hoc* repertoire for the purposes of their research since the topic, period or object of study chosen may require it. In that case, relying on already existing databases like the ones described above can be of great help. Once compiled, catalogues can be examined in order to establish regularities 'which will in later stages be useful to establish criteria for further corpus selection' (Merino 2005: 89).

Corpora

The corpus approach to translation studies, especially in its descriptive branch, has proven to be a fruitful methodology which is now considered to be a research paradigm in its own right, whose main aims are 'the empirical study of the product and process of translation, the elaboration of theoretical constructs, and the training of translators' (Laviosa 2003:45). Before analysing in detail the kinds of corpora which can help the researcher in the historiographic mapping of translation, it is necessary to clarify what we mean by a corpus. We follow Bosseux when she states that

> a corpus now primarily refers to a collection of texts that is held in machine-readable form and can consequently be analysed in a variety of ways, both automatically and/or semi-automatically (2007:80)

The main difference therefore between a catalogue and a corpus is that the latter includes the actual texts in digital format, and not just the bibliographical information, or metadata. Because of this, corpora are normally the subsequent step after the compilation of a catalogue in any given research, since they complement each other. Corpus-based approaches, which were first used in the 1980s, have now become widely-used in translation studies, especially in its descriptive branch since they offer quick access to empirical evidence and, if analyzed properly, they offer immediate feedback. These two assets make them a very valuable tool for the translation history researcher, who can use them to reconstruct the process of translation or the translators' *modus operandi*. As Granger rightly acknowledges, 'it was Mona Baker who pioneered the corpus-based trend in the early 90s' (2003:18). Different kinds of corpora can be compiled and used in translation research, depending on its purpose and scope. It is not our aim to be exhaustive here, but to offer a brief explanation of the several possibilities available to the researcher of the field thanks to this approach, complementing them with practical examples of their use, in the hope they can be an aid when undertaking research in translation history.

The most common kinds of corpus used in translation studies are the following (quoting Ramón 2002: 401):

- *Translation corpora or parallel corpora*: they consist of texts in one particular source language and their corresponding translations into one particular target language.
- *Comparable corpora* (multilingual corpora in Baker's terminology, 1995): corpora of original texts in two different languages. The texts are not translations of each other, but they deal with the same topic and share features such as length, date and intended audience that make them comparable.
- *Corpora of translated language* (comparable corpora in Baker's terminology, 1995): complex kind of monolingual corpus that includes texts translated in one particular language, for example English, from a variety of other languages, for example French, Spanish or German.

The use of different kinds of corpora like the ones described above can help the researcher obtain information from primary sources which is not available from secondary and tertiary ones. Thus, the use of parallel corpora enables the comparison of source and target texts, more quickly and with a much greater level of detail than is possible with manual analysis, thanks to the several possibilities that technology has to offer when dealing with texts (see below). Comparable corpora can allow the translation history researcher to gain knowledge about the style of texts in a determined period while corpora of translated language can serve to check if the language of translated texts differs from the one of texts originally written in that language, thus opening the field to studies of style marks, translation authorship or reception studies on a more general basis.

We agree with Tahir-Gürçağlar when she states that 'the major milestone for historical translation research has been the emergence of DTS' (2013:138).[2] Many studies carried out in the descriptive branch of Translation Studies use translation or parallel corpora as a main source of data. The corpora used on a historical project will need to be tailored to its objectives and will, therefore, often have to be created *ad hoc*. This is mainly due to the fact that ready-made corpora are scarce, especially if we are interested in texts from specific periods or contexts. On the other hand, creating new corpora can be quite expensive since they require the use of digital tools and it may be necessary to pay copyright on the material being used. The main advantage of *ad hoc* corpora is that, at least in theory, any notion that is expressed in the source language should have an equivalent in the target language, which opens up the possibility of interesting research.

Having the complete texts in digital format, however, does not *per se* guarantee any additional understanding of the translation phenomena compared to a manual analysis. What is more, one could argue that the risk of losing information in the process of digitization is always present, such as the dimensions of a document, which are important signifiers in their own right in manuscripts and printed media (Wakabayashi, 2019). This is where text analysis tools come to the fore,[3] allowing for the establishment of relationships, statistics or indexes in a way that facilitates shifts between a micro and a macro scale such as would be difficult to achieve by means of a

[2] Ed: see Vandaele in this volume for a different perspective on DTS in relation to translation history.
[3] 'A corpus is only as good as the querying system you have to consult it' (Roberts 1996, quoted in Rabadán y Nistal 2002:69).

non-computerized study. Before it can be used, however, this kind of corpus needs to be prepared for analysis by means of annotating and/or aligning tools. There are several tools used in the preparation of corpora: here we offer a brief description of the most important, explaining what they are used for in reference to doing research in translation history. We will refer to the tools according to their functions, independently of their actual name, as this is what interests us for research purposes.

Preparing the texts for use in a corpus

Annotation Tools

Apart from the actual text, a corpus can also be provided with additional linguistic information, called 'annotation'. This meta information can be of different types, but the most common are grammatical tags which are a very useful tool for research on language. Annotation tools substantially extend the range of research questions that a corpus can be used for – which, in the case of translation history, can focus on context-sensitive features that reflect the changes that may have taken place in a language over time, or stylistic features related to the translator's style.

Text Translation Alignment Programs

Once the texts and their translations are in digital format and are annotated, the source and target segments need to be aligned so that the translation can be analyzed: for this operation the researcher uses an alignment tool. The unit of segmentation and the way the results are displayed continue to be controversial points when using these tools, but their usefulness is undeniable in any study of translated texts today.

Querying the texts

Concordance Generators

The term concordance has evolved over time[4] and it is usually understood now as a collection of the occurrences of a word form, including the word's immediate textual context, up to a predefined number of words to the left and right within the text. The concordance is displayed with the search term in the centre of the screen, so one can very easily and intuitively perceive patterns in its use. Thus, a concordance generator is basically a program for looking up words and expressions in their context, within a

[4] See Rabadán y Nistal (2002: 69) for a brief account on its development.

corpus of texts. In the case of a bilingual matching generator, it can be used to search in a parallel corpus of already translated segments, which can be very useful in revealing usage patterns (Wakabayashi 2019) and also the evolution over time of the meaning of key words and concepts (see the *Genealogies of Knowledge* project described below). However, one of the main limitations of concordancers is the relative lack of contextual information they offer: for example, research on context-specific text would require more than a few concordance lines for the investigation to be accurate.

Anchor Words

As Rabadán indicates (2008: 107), in a parallel corpus, 'the anchor words are specific words that are defined for the two languages involved': these words are normally related by some type of cross-linguistic equivalence and their main use has to do with the identification of specific examples which can be illustrative of the phenomenon the researcher is looking for in his/her analysis.[5] We should keep in mind that the selection of the anchor words can slant the results, as can the selection of the texts to analyze in the first place. An example of how anchor words can be used is the study of point of view in literary texts carried out by Bosseaux (2007): in her search for certain linguistic and narratological features of two Virginia Woolf novels, *To the Lighthouse* and *The Waves,* she turned narratological concepts into linguistic entities in the form of anchor words which could subsequently be analyzed by the software, something which proved to be a complex process. Thus, in her search for the point of view in her study, she paid attention to the use of deixis, among other aspects, by searching for it in in the form of words such as 'now' or the first pronoun 'I'. This allowed her to gain further insight on how far a translator's choices affect the novel's point of view. The potential of tools such as this one is therefore large, but they are only as useful as the skillfulness of the analysis being made.

Visualizing the results

The way in which the results obtained from a corpus study are visualized has an impact on how they are presented to the community and therefore on how effectively the research is disseminated. The possibilities for visualizing the results of studies carried

[5] See the section on TRACE below for an example of this kind of word and their usability in translation history research.

out with corpora are almost endless and in a way which fosters not only a more active presentation and questioning of results, but also facilitates public engagement (Wakabayashi 2019). Researchers today have easy and free access to visualization packages such as Wordle or Phrase Net, but since software can change overtime it is not our purpose here to offer a comprehensive list of the resources available but mainly to comment on their usefulness for the translation history researcher.[6] Among the functions these tools offer we find the creation of animated maps, the presentation of historical networks or animated timelines, all of which help the translation scholar to explore the underlying causes of the relationships s/he has established using corpora (ibid). Some might argue that the only advantage these tools offer lies in the powerful way they display results and that they do not add any essential insight compared to a more traditional presentation of the results. As with the rest of the tools or resources presented in this chapter, these visualization packages do present some shortcomings, such as the need for textual explanations in some cases and also the lack of contextualization of some of the connections established, which can be a great hindrance in a translation history study. According to Theibault, the key issues of these visualizations are the density and the transparency of their information which is not always adequate (Theibault, J. 2013). An example of the use of these visualization resources can be seen in the project entitled *Mapping the Republic of Letters,* developed by Stanford University, in the USA, which creates sophisticated, interactive tools in order to address questions about the scholarship networks which were the lifelines of learning, succeeding in digitalizing and visualizing early modern correspondence in innovative ways. The project is made up of a wide range of case studies which give the researcher multiple points of intersection and which are based on different information sources, each one presenting particular information visualization challenges (http://republicofletters.stanford.edu/index.html).

Currently, most of these tools are freely available online and lend themselves to interactive use and online collaboration, including across disciplines, a fact which is particularly advantageous in producing relational translation histories (Wakabayashi 2019). Some efforts have been made in this direction, and one of the most interesting to

[6] For a comprehensive list of useful software the DIRT Directory is a wiki which offers tools for carrying out many different forms of text analysis and visualization.It has not been updated since 2012, but it is still one of the best compilations to date:
https://digitalresearchtools.pbworks.com/w/page/17801672/FrontPage

date is the project entitled *Corpusnet* (http://corpusnet.unileon.es/). This consists of a hub of bilingual and multilingual corpora and related resources featuring any of the languages of Spain alongside other languages (mainly English, French, German, Italian and Portuguese). The project is run by ten researchers from eight different Spanish Universities who are led by Rosa Rabadán at the University of León, and its main objective is to promote more ambitious and more visible research by facilitating easier access to existing resources and encouraging the cooperation of users. As such, the project offers researchers a really comprehensive and useful compilation of both parallel and comparable corpora of different types, compiled by the research groups involved and for either linguistic or translation history research. This is complemented by a set of *ad hoc* tools. Such a freely available online compilation is exceptional and invaluable. The network was published online in June 2019, allowing any researcher free access to its vast repertoire.[7]

Now that we've looked at some of the most widely used tools and resources, it is the time to think more specifically about the application of this corpus-informed approach to translation history and outline possible avenues of research which could benefit from it, apart from the ones already mentioned. Thus, potential research paths where the use of these repertoires and corpora are particularly useful include:

- the study of a specific period of history
- the study of translation history in a particular country
- the study of a particular text genre and its development throughout history
- the study of translation norms in a specific period and/or context
- the study of a translator's style related to a specific period

These are just some examples of possible approaches which could be tackled by means of this alliance between technology, linguistics and history and which still need to be covered in more depth in the scholarly field. As Santoyo states, 'if we think of the history of translation as a *mosaic*, there can be little doubt that there are still many small pieces or tesserae missing, as well as empty spaces yet to be filled in' (2006: 13). The next section sets out to outline some important attempts at completing that mosaic with the aid of corpora.

[7] Project Name: *Corpus y networking: consorcio de proyectos para la gestión de recursos bi/multilingües y sus aplicaciones*. Reference: FFI2016-81934-REDT.Funding body: MINECO (Ministerio de Economía y Competitividad).Period: 2017-2019.

Some examples of research projects using corpora

Due to space constraints, we cannot offer a comprehensive account of studies carried out using analytical tools from linguistics to do research in translation history. Therefore, I have chosen two examples: one that was launched 30 years ago and another more recent project.

The TRACE Project: Mapping the History of Translation in Spain under the Franco Dictatorship and beyond (http://trace.unileon.es)

Ideology and translation are concepts which are often intertwined in the study of translation history. The TRACE (TRAnslations CEnsored) project tackles these two variables in a very specific context: Spain under the Franco dictatorship (1939-1975 and beyond).[8] The initial hypothesis is that every cultural product imported during this period was controlled by the ideological, linguistic and cultural expectations of the target culture and, therefore, both the official censorship and self-censorship were instrumental in shaping their translations.

From its very beginning in the late 1990s, TRACE has always been a joint venture of several researchers in two important Universities in Spain (León and the Basque Country, and Cantabria at one stage). The aim of this project is to construct a map of what actually got translated in Spain and how during that period, 'not from what could have been, or could nowadays be, but from empirical evidence drawn systematically from rich documentation sources' (Merino 2005:87). So far, there have been numerous studies that have employed the TRACE methodology: the studies were organized according to period (the dictatorship lasted almost forty years and it was more practical to establish subperiods), genre (poetry, narrative, theatre, audiovisual media) and combination of languages, the most studied being English-Spanish.[9]

In TRACE's research, the use of an electronic corpus assists the researcher in the search for the problematic passages of the texts. The TRACE methodology follows several steps, being the first one the compilation of a catalogue or corpus 0 (since it does not contain any text yet) which is the basis for further textual study (Merino 2003: 644). For example, in the case of my own research inside the project, I designed a

[8] Even though Franco died in 1975 and with him the official period of dictatorship, the mechanisms for book control continued operating until the establishment of a Constitution in 1978 and with it freedom of expression.

[9] For two visual representations of the several studies carried out inside the project so far, see Merino Álvarez's tables in 2017: 143.

catalogue of narrative works composed of more than 9,000 entries containing information of translated narrative from English into Spanish during the last years of the Francoist control system (1970-1978) which could be exploited in many ways. In order to compile it, like the rest of the members of the project, I resorted to several databases and information from the censorship files.[10] The metadata included in this catalogue contained information about the most relevant aspects of each entry, such as publishing house, translator, year of publication, etc. Given that it is such an extensive catalogue, further analysis of the texts is made considerably quicker and more effective thanks to the use of tools such as the ones described above.

The transition to an actual textual corpus from the catalogue of metadata (or corpus 0) is carried out according to a process which can be quite critical due to the wide range of possibilities that can present themselves to the researcher: thus, the entries can be studied paying attention to aspects such as the most translated authors, the most significant publishing houses or the most prolific translators, to name a few. This narrows down the analysis according to the purpose of the research in questions and the hypothesis to be tested. Thus, for example, the study I carried out on translated fiction focused on those entries which had undergone some changes in their translated text; changes that were either imposed by the censorship board (and therefore present in the censorship file) or by the publishing house, or implemented by the translator him/herself (what is commonly known as self-censorship).

Once the textual corpus (corpus 1) has been selected according to the criteria the researcher wants to investigate,[11] the texts need to be prepared. This is an *ad hoc* designed corpus, and, as such, it is time-consuming for the researcher and normally designed for a specific purpose, but if it is formatted according to international standards, this helps to expand its usefulness and future usability.[12] The texts need to be aligned, thus displaying at the same time the English source text and the Spanish target

[10] One of the most reliable sources for understanding the cultural landscape of the period is the AGA (General Archive of the Administration), in which the records of the censorship procedures during the Franco regime are held. The lists of authors, national and foreign alike, and titles of plays, original and translated, which were duly filed when submitted to the censor by producers, editors or exhibitors, have become a sort of archaeological site to be excavated and studied -- something which is accomplished in the project (Merino Álvarez 2016:37).

[11] The representativeness of the texts chosen in the first place to compile corpus 1 is one of the most controversial issues in works of this kind and if it is not done following quantitative criteria, it can be done by considering the question of the significance of the material chosen instead of its statistical representativeness, which would be perfectly valid as well.

[12] To this respect, and advocating for a further life-span of *ad hoc* corpora, Rabadán (2019) proposes several possibilities, among them adding layers of annotation to them or combining parallel data with those provided by comparable corpora.

text (or texts, if there was more than one translation available). Having experienced first-hand the problems that a descriptive research like the one carried out under the TRACE group can encounter when dealing with texts of a different nature (theatre, narrative, poetry, cinema) when aligning and then analysing them, I believe this step in the research has greatly benefited from the creation of a tool such as the TAligner 3.0. Among its main features, it includes an aligner and a tool for the analysis and export of results. It has been designed taking into account the different textual types a TRACE researcher might deal with and, unlike other products, it makes it possible to consider the paragraph or the sentence in the case of narrative texts, or the "replica" in the case of audiovisual or theatrical texts as the unit of alignment. One of the main advantages of the program is that it can be used for labelling and aligning texts in the same language or in various language combinations, which broadens the possibilities of study.

The last step in the methodology is to choose the concrete passages from these texts that will serve to test a hypothesis, thus compiling a corpus 2, which is a corpus formed by segments from the texts of our corpus 1. These kind of corpora are particularly useful in descriptive research because they provide evidence of how translators actually perform, which can help to explore norms of translation in specific socio-cultural and historical contexts (Bosseaux 2007: 80). In the TRACE project, the selection of the segments from the texts which are object of a comparative analysis has traditionally been carried out according to the four thematic areas which coincide with the categories most used by the censors operating at the time, namely: sexual morals, politics, religion and profane language. These categories had already been indicated as the most contentious ones during the regime by one of the most prestigious researchers of censorship during the Francoist period in Spain, the sociologist Manuel Abellán (1980). Furthermore, they are the ones normally reflected in the censorship file templates that can be accessed in the AGA. These templates specifically referred to morality, attacks against Catholicism and the politics. The use of profane language was frequently added by the comments of the censors as well.

Having identified the segments which could be problematic in the Source Text and their counterparts in the Target Texts, where there might be evidence of (self) censorship, also using *anchor words*, these segments are collected to form a corpus 2. In my own research on narrative texts, I compiled a list of terms which can be considered as signs of the issues that the authorities of Franco's regime were looking out

for and which are present in the AGA files.[13]I identified key terms which allowed me to search the aligned texts with the software available for that means and identify instances of censorship. Table 1 below lists some *anchor terms* that were used to identify instances of profane language:

ANCHOR TERMS- PROFANE LANGUAGE			
ENGLISH	***ESPAÑOL***	***ENGLISH***	***ESPAÑOL***
Bastard	Cabrón; Hijo de puta	Fucking…	… de mierda; jodido
Bitch	Puta; Zorra	Hell (what the hell; who the hell, etc; like hell; hell of a lot,etc.)	Coño, qué coño, quién coño, etc;
Bloody…	Maldito/a; puñetero/a	Jesus (Christ)!	¡Hostia!¡Coño! ¡Joder! ¡Mierda!
Christ!	¡Hostia! ¡Joder!	Shit; Not to give a shit	Mierda; No importar un carajo; un bledo
Damn* /darn /damm*; Give a damn	Maldito; puñetero Importar un bledo	Whore	Puta

Table 1. Examples of Anchor terms: Profane Language

The list was compiled so that a scan of the texts using these terms would retrieve around 95% of those segments that might be considered controversial. Once the segments have been identified, they are compared to see if there are any significant changes between the English and the Spanish version(s). And on the basis of any changes that may be found, the researcher can deduce whether or not there had been any official or self- censorship.

Up to now, there have been numerous studies that have employed the TRACE methodology: English-Spanish translations have already been mapped out for the most part, whilst new avenues of research involving French and German as source languages and Basque as target language have been opened up. As the research continues, we get a clearer picture of the real progress, both from a quantitative and a qualitative standpoint, experienced in the translation and censorship of narrative texts (the TRACEn corpus in TRACE terminology), poetry (TRACEp corpus), theatre plays (TRACEt corpus), and audiovisual materials (cinema: TRACEc and television: TRACEtv). Having access to

[13] I did this in the same way that Bosseaux put 'narratological concepts into linguistic entities that the software would be able to analyse' in her study of the point of view in literary texts via computerized means and applied to two novels of Virginia Woolf (2007: 11). For a complete reference to the list of anchor words used see Gómez Castro (2009).

thousands of censorship files, it is possible to explain which sets of texts were more representative and which had a greater impact on the target culture. Software tools were used at various stages during the project. First a catalogue was created in order to select the texts for the textual corpora, and, considering the very high number of entries, software tools allowed us to identify quantitative trends and compile statistics. Second, a textual corpus was compiled containing the full texts of the works selected, which could be queried both in terms of the censorship categories we'd identified, as well as for more linguistic-oriented research such as grammatical and lexical analysis, for example, thus extending the life-span and usability of the corpora. Third, the texts were scanned using a list of *anchor words*, which allowed us to examine the text in a way which would have been extremely difficult to carry out manually even in an annotated text. However, there were difficulties, as with any approach. Sometimes the texts, due to their date of publication, are not available in digital format and therefore it is necessary to scan them in advance and prepare them for use with the software tools. This is tedious and time-consuming work, but is unavoidable when adopting this methodology. Also, some consider this approach to lack contextual evidence, but this issue can be addressed by integrating a study of the historical context and of the polysystemic relations inside the literary scene of the period under scrutiny. Thanks to this integrated mode, the researcher can maintain a perspective of the historical moment in which the translations were carried out and how it affected the work of the translators.

By carrying out studies in this way, the project members have contributed historical accounts of the translations done during this very specific period, thus shedding light on obscure or unknown areas of the Spanish translated culture under a dictatorship which made use of censorship for cultural control.

The Genealogies of Knowledge Project (http://genealogiesofknowledge.net/)

> The intellectual history of translation examines the production, changes, and migration of discourses (theoretical, philosophical, critical, literary, academic, social, institutional, methodological, popular) on translation across time, space, and contacts with other disciplines or under the effect of external constraints.
>
> (Wakabayashi 2012: 2539)

Genealogies of Knowledge is also a large scale project which involves the use of diverse electronic corpora and software tools. Led by Mona Baker at the University of Manchester and funded by the Arts and Humanities Research Council in the UK, this

ambitious project, which officially come to an end in the Spring of 2020 but which continues producing studies, focuses on translation phenomena and other sites of mediation involving ancient Greek as a starting point, plus three distinct lingua francas: medieval Arabic, Latin and modern English. Its research centers on a series of key cultural concepts such as democracy, equality, truth or nation and how translation has contributed to their transformation in history and as they travelled through cultures, languages and epochs over the last 2,500 years. The project focuses on two constellations of concepts: one related to the body politic and the other related to scientific, expert discourse. Rather than focusing on a series of unconnected individual concepts, the project studies these two constellations that 'have been central to Anglophone and European societies since the medieval period and are usually traced back to the ancient Greeks' (Baker et al., 2020). These concepts are expressed by different lexical items in different languages, and, concentrating on these two constellations, the project outlines two threads of analysis: one centered on evolution and another one on contestation. The first involves tracing shifts in the meaning and the use of a given lexical item like *democracy* or *nation* in English over time and across different geographical spaces (ibid). The second implies examining how these concepts, or specific interpretations of them, are contested by various individuals or groups, especially in digital space (ibid). The most interesting for us here is the first one: the historical evolution and transformation through translation of the two constellations, focusing on seminal moments of change in the reception and reproduction of translated texts and their meanings by subsequent readerships. This entails examining commentaries and (re)translations from/into Greek, Latin, medieval Arabic and modern English (Genealogies of Knowledge Webpage, 2019).

Such a challenging study requires large corpora in the three languages studied, and a range of open-source software applications to interrogate the corpora and assist with the presentation of findings. Instead of advocating for the use of strictly parallel corpora including just translations, the project includes 'a series of non-parallel but carefully interlinked corpora' (Baker et al. 2020) which, as the members of the project recognize, can preclude certain kinds of analysis since there is no alignment of the texts (as happened in TRACE, for example), but which, on the other hand, can offer a view of the translated text without being constantly compared with a source in a search for supposed inaccuracies.

Since the corpora compiled for the project are not primarily designed for researching on linguistic features or to delve into translator style but with thematic criteria in mind, issues which are commonly considered very relevant for these kinds of studies such as size, balance or representativeness are not likely to be encountered here. Besides, even though the researchers working on the project come from various disciplines, they each specialize and use just one or two corpora, but they all use the same tools, which are open source, thus maximizing the resources developed by the project and establishing a long-lasting legacy. The Genealogies of Knowledge project aims to provide free, restricted access to the corpora and the software through a specially designed interface, an aspect which contributes to the advance of the digital humanities and to collaboration among colleagues from the same or neighboring disciplines. Furthermore, the researchers of the project highlight the importance given to visualization techniques[14], which enhance interaction with the corpus (Baker et al., 2020).

The range of studies that can be carried out using the resources of this project is wide, and the research avenues that can be followed include studies on retranslation, patterns in collocation or reference strategies, to name a few. Up to now, some works have been published dealing with political discourse and statesmanship (Jones 2019), community and authority (Buts 2020), migrants and exile (Baker 2020) or Aristotle's works (Karimullah 2020), among others. With this kind of research, the project hopes to yield novel insights into how translations and related forms of mediation generate and transform knowledge, as stated in its main aims. Given that it is such a large scale project, it involves a significant step forward in the kind of corpora studies normally related to translation and linguistics, broadening their scope to include other disciplines and languages. Nonetheless, like any other study, it does have some limitations: it could be argued that its results suffer from a lack of historical context, which is however partially counteracted by the presence of commentaries and critical editions in the corpora, as well as translations. At the risk of working 'on a scale that elides individual historical actors' (and factors, I would add) as Robertson and Mullen rightly indicated (2017:18-19), the results provided so far are promising and show an emerging conceptual sphere across very different environments, something worth exploring further (Baker et al., 2020).

[14] See the work by Luz and Sheehan (2020) for a thorough explanation of the development of visualization tools for the Genealogies project.

Studies to identify typical features of translated texts and/or translator's style

Some other analyses, albeit on a smaller scale, have been carried out in the field of translation history with the aid of corpora and technology, particularly those which use these tools to analyse the textual, grammatical and stylistic features of translated texts. This kind of study uses a corpus-based approach to uncover ideology in texts by comparing lexical features of translations and original (non-translated) texts of the target culture. Thus, for example Kemppanen (2004) and Laviosa (2000) use this approach with different corpora but with a similar methodology, by means of keywords. Kempannen looks at Russian-Finnish translations and original Finnish texts within the discourse of Finnish political history and does so by combining a quantitative analysis of keywords, ideologically functional words which can be isolated from the texts, and a qualitative one where texts are examined from a narrative point of view (2004:90). He chooses the keywords by investigating the context in which they are used, which allows him to perceive lexical patterns that can be interpreted as the expression of an ideology (concordance generators can be handy at this stage). Examples of keywords in his study are 'Suomi (Finland)', 'sota (war)' and 'ystävyys (friendship)'. The study can focus on one or more keywords at the same time, and the features of translated language are identified by relating them to the properties of non-translations (104). Laviosa did something similar with five frequent and semantically related words (Europe, European, European Union, Union, and EU) looking at them in the newspaper subcorpus of the Translational English Corpus (TEC), which is a monolingual corpus of English texts translated from a variety of source languages, covering four text genres: fiction, biography, newspaper articles, inflight magazines (2000: 161). By carrying out this study, she wanted to show that a corpus-based methodology had reasonable potential to support qualitative research beyond mere linguistic description (ibid) and that translated language can be investigated *per se,* without the need to refer to other corpora.

Bosseaux (2007, see above) is another example of a study carried out to demonstrate that corpus-based tools can greatly facilitate and enhance the comparison of source and target texts beyond a manual analysis, in this case to examine how point of view in a work of fiction is created in the original and adapted in translation. Along similar lines, Ruano San Segundo (2017)[15] analyzes how speech verbs in Dicken's

[15] Although Dickens is the author he has most profusely studied, Ruano San Segundo has also carried out analysis on other authors such as Tennessee Williams (for a comprehensive list of his publications, see his profile in Google Scholar: https://scholar.google.es/citations?user=_9F6w_0AAAAJ&hl=es&oi=ao).

Hard Times are rendered in four Spanish translations using concordancing software and parallel corpora in his aim to illustrate how these verbs also contribute to characterization. The way they are translated, he argues, has a significant impact on the impression readers form in their minds, and a study of this kind is intended to show that literary translation studies can benefit from electronic tools in the still emerging field of Corpus Literary Translation Studies (CLTS).

Stylometry is another area of research which can benefit from a corpus-based approach and, concerning translation history, can help ascribe authorship to anonymous translations or track a translator's stylistic habits through different translations of various authors. It is a question of assuming, as Crisafulli rightfully acknowledges, that 'the translator's outlook will surface at specific sensitive points of the target text' (2002:40). Baker (2000) was one of the first researchers to make use of this methodology for researching the style of a literary translator and she acknowledged possible drawbacks that should be taken into account for future studies, such as the difficulty in distinguishing between stylistic elements that are attributable to the translator and those which simply reflect the source author's style (2000: 261). Studies of this kind seem to work best when researchers consider translation not only as a mere reproductive activity but as a creative one. It is important to remember that the ideal procedure is to always historically contextualize the results obtained through the use of corpora and their tools in reference to historical research using primary historical sources so that we can have both the linguistic microscope and the cultural telescope, following Tymoczko's metaphor (2002).

Conclusion

Technology evolves at a very quick pace and the tools and resources used today can become obsolete very quickly. This forces the researcher to be constantly up to date and alert. It is undeniable that this is demanding, but, as Tahir-Gürçağlar indicated, 'the availability of electronic resources has made historical research on translation history both easier and more challenging' (2013: 140). Nonetheless, it is important to note that these techniques are not suitable for all research avenues, and to consider the reticence that some researchers feel about the use of a corpus approach in doing historical research on translation, because the historical context can sometimes be lost in the process (Rundle 2012: 236).

Notwithstanding these reservations, with this chapter we wish to embrace the positive aspects that linguistic tools and digital humanities have to offer for the study of translation history. Given that translation history is largely text-based rather than event-based (Wakabayashi 2012), electronic resources and tools can prove useful in this field. A careful, meditated and aware use of these tools is the key to a fruitful compilation of data and cross-fertilization of disciplines. Eventually it's all a question of the results they facilitate, since 'quantitative methods cannot in themselves write good history. But they can help us head in the right direction' (Pym 1996: 14).

Related topics

Translation History; translation historiography; method in translation history; digital tools; corpora; Digital Humanities.

Further reading

Pym, Anthony (1998) *Method in Translation History.* Manchester, St. Jerome.
A pioneering work where Pym offers a set of methodological tools for the empirical study of translation history.

Rabadán, Rosa y Purificación Fernández Nistal (2002) *La traducción inglés-español: fundamentos, herramientas, aplicaciones.* León, Universidad de León.
An excellent comprehensive study combining both translation research and professional practice with clear explanations of tools and their applications in the field.

Tahir-Gürçağlar, Şenaz (2013) "Translation History" in *The Routledge Handbook of Translation Studies*, Carmen Millán and Francesca Bartrina (eds). New York, Routledge: 131-143.
A magnificent entry in the Routledge series focused on key differences in the field of translation history and on its relevance.

Wakabayashi, Judy (2019) "Digital Approaches to Translation History", *The International Journal for Translation and Interpreting Research* 11(2): 132-145.
A very accomplished account of the latest advances in digital approaches to translation history.

Bibliography:

Abellán, Manuel Luis (1980) *Censura y creación literaria en España (1939-1976).* Barcelona: Península.

Baker, Mona (1995) "Corpora in Translation Studies: an Overview and Some Suggestions for Future Research", *Target,* 7: 223-43.

Baker, Mona (2000) "Towards a Methodology for Investigating the Style of a Literary Translator", *Target,* 12(2)*:* 241-266.

Baker, Mona (2020) "Rehumanizing the migrant: the translated past as a resource for refashioning the contemporary discourse of the (radical) left", in *Genealogies of Knowledge*, Mona Baker and Henry Jones (eds). Special collection for *Palgrave*

Communications 6 (12), URL: https://www.nature.com/articles/s41599-019-0386-7 (accessed 20 September 2020)

Baker, Mona et al. (2020) "Using Corpora to Trace the Cross-Cultural Mediation of Concepts through Time: An interview with the coordinators of the Genealogies of Knowledge Research Network". Interview and translation by *Zhao Wenjing.* Available online at: https://genealogiesofknowledge.net/2020/04/29/using-corpora-to-trace-the-cross-cultural-mediation-of-concepts-through-time-an-interview-with-the-coordinators-of-the-genealogies-of-knowledge-research-network/. (accessed 10 September 2020).

Bosseaux, Charlotte (2007) *How Does it Feel? Point of View in Translation. The Case of Virginia Woolf into French.* Amsterdam/New York: Rodopi.

Buts, Jan (2020) "Community and Authority in *ROAR Magazine*", in *Genealogies of Knowledge*, Mona Baker and Henry Jones (eds). Special collection for *Palgrave Communications* 6 (16), URL: https://www.nature.com/articles/s41599-020-0392-9 (accessed 20 September 2020)

Crisafully, Edoardo (2002) "The quest for an eclectic methodology of translation description" in *Crosscultural transgressions: Research models in translation studies II: Historical and ideological issues,* Theo Hermans (ed.), Manchester, England: St. Jerome: 26-43.

Folaron, Debbie. A (2018) "Technology" in *A History of Modern Translation Knowledge,* Lieven D'hulst & Yves Gambier (eds). Amsterdam: John Benjamins: 113-116.

Genealogies of Knowledge Webpage (2019) Available at: https://genealogiesofknowledge.net/about/ (accessed 1 August 2020)

Gómez Castro, Cristina (2009) *Traducción y censura de textos narrativos inglés-español en la España franquista y de transición TRACEni (1970-1978)*. León: Universidad de León. Unpublished PhD dissertation. Available online: https://buleria.unileon.es/handle/10612/1413

Granger, Sylviane (2003) "The Corpus Approach: a Common Way Forward for Contrastive Linguistics and Translation Studies?" in *Corpus-Based Approaches to Contrastive Linguistics and Translation Studies,* Sylviane Granger, Jacques Lerot & Stephanie Petch-Tyson (eds) Amsterdam/New York: Rodopi: 17-29.

Jones, Henry (2019) "Searching for Statesmanship: A corpus-based analysis of a translated political discourse", *Polis: The Journal for Ancient Greek and Roman Political Thought* 36(2): 216-241.

Karimullah, Kamran (2020) "Editions, translations, transformations: refashioning the Arabic Aristotle in Egypt and metropolitan Europe, 1940–1980", in *Genealogies of Knowledge*, Mona Baker and Henry Jones (eds). Special collection for *Palgrave Communications* 6 (3), URL: https://www.nature.com/articles/s41599-019-0376-9 (accessed 20 September 2020)

Kempannen, Hannu (2004) "Keywords and Ideology in Translated History Texts: a Corpus-Based Analysis", *Across Languages and Cultures,* 5 (1): 89-106.

Laviosa, Sara (2000) "*TEC:* a Resource for Studying what is "in" and "off" Translational English", *Across Languages and Cultures,* 1(2): 159-177.

Laviosa, Sara (2003) "Corpora and Translation Studies", in *Corpus-based Approaches to Contrastive Linguistics and Translation Studies, ,* Sylviane Granger, Jacques Lerot & Stephanie Petch-Tyson (eds) Amsterdam/New York: Rodopi: 45-54.

Luz, Saturnino and Shane Sheehan (2020) "Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge" in *Genealogies of Knowledge*, Mona Baker and Henry Jones (eds).

Special collection for *Palgrave Communications* 6 (49), URL: https://www.nature.com/articles/s41599-020-0423-6 (accessed 20 September 2020)

Merino Álvarez, Raquel (2003) "TRAducciones CEnsuradas inglés-español: del catálogo al corpus TRACE (teatro)", en *I AIETI. Actas del Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación. Granada, 12-14 de febrero de 2003,* Ricardo Muñoz Martín (ed). Granada: AIETI: 641-670.

Merino Álvarez, Raquel (2005) "From Catalogue to Corpus in DTS: Translations Censored under Franco. The TRACE Project", *Revista Canaria de Estudios Ingleses,* 51: 85-103.

Merino Álvarez, Raquel (2016) "The censorship of theatre translations under Franco: the 1960s", *Perspectives*, 24(1): 36–47.

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.

Pieta, Hanna (2013) "Fontes bibliográficas na história da tradução em Portugal e sua aplicação na identificação de traduções da literatura polaca" in *A Scholar for all Seasons: Homenagem a João de Almeida Flor*, J. Carlos Viana Ferreira et al. (eds). Lisboa: CEAUL: 297-309.

Poupad, Sandra, Anthony Pym & Esther Torres-Simón (2009) "Finding Translations. On the use of Bibliographical Databases in Translation History", *META: journal des traducteurs/META:Translators' Journal,* 54:2: 264-278.

Pym, Anthony (1996) "Catalogues and Corpora in Translation History", in *The Knoweldge of the Translator: From Literary Interpretation to Machine Translation,* Malcolm Coulthard and Patricia Odber de Baubeta (eds). Lewiston/Queenston/Lampeter: Edwin Mellen Press: 167-190. URL: http://usuaris.tinet.cat/apym/on-line/research_methods/1996_catalogs.pdf (accessed 1 August 2019).

Pym, Anthony (2009) "Humanizing Translation History", *Hermes: Journal of Language and Communication Studies*, 42:23-49.

Rabadán, Rosa y Purificación Fernández Nistal (2002) *La traducción inglés-español: fundamentos, herramientas, aplicaciones.* León: Universidad de León.

Rabadán, Rosa (2008) "Refining the idea of «applied extensions»", in *Beyond Descriptive Translation Studies: Investigations in homage to Gideon Toury,* Anthony Pym, Miriam Shlesinger & Daniel Simeoni (eds). Amsterdam/Philadelphia: John Benjamins: 103-117.

Rabadán, Rosa (2019) "Working with parallel corpora: Usefulness and usability", in *Parallel Corpora for Contrastive and Translation Studies. New resources and applications,* Irene Doval & M. Teresa Sánchez Nieto (eds). Amsterdam; Philadelphia: John Benjamins: Studies in corpus linguistics, v. 90: 57-77.

Ramón, Noelia (2002) "Contrastive Linguistics and Translation Studies Interconnected: The Corpus-Based Approach", *Linguistica Antverpiensia, New Series-Themes in Translation Studies,* 1: 393-406.

Roberts, Roda P (1996) "The Use of Bilingual Corpora in Translation", unpublished conference given at the University of Valladolid, Spain (18th April 1996).

Robertson, Stephen and Lincoln Mullen (2017) "Digital history and argument", *Roy Rosenzweig Center for History and New Media.* Retrieved from:

https://rrchnm.org/wordpress/wp-content/uploads/2017/11/digital-history-and-argument.RRCHNM.pdf

Ruano Sansegundo, Pablo (2017) "Reporting Verbs as a Stylistic Device in the Creation of Fictional Personalities in Literary Texts", *Atlantis,* 39 (2): 105-124.

Rundle, Christopher (2012) "Translation as an approach to history", *Translation Studies,* 5 (2): 232-240.

Santoyo, Julio César (2006) "Blank Spaces in the History of Translation", in *Charting the Future of Translation History*, George Bastin and Paul Bandia (eds). Ottawa, University of Ottawa Press: 11–43.

Tahir-Gürçağlar, Şenaz (2013) "Translation History" in *The Routledge Handbook of Translation Studies*, Carmen Millán and Francesca Bartrina (eds). New York, Routledge: 131-143.

Theibault, John (2013) "Visualizations and historical arguments", in *Writing History in the Digital Age,* Jack Dougherty & Kristen Nawrotzki (eds). Ann Arbor: The University of Michigan Press: 173-185.

Tymoczko, Maria (2002) "Connecting the Two Infinite Orders. Research Methods in Translation Studies", in *Crosscultural Transgressions: Research Models in Translation Studies II: Historical and Ideological Issues,* Theo Hermans (ed). Manchester: St. Jerome: 9-25.

Wakabayashi, Judi (2012) "History of Translation", *The Encyclopedia of Applied Linguistics* Vol. 4. Wiley-Blackwell. 2012: 2535–2542.

Wakabayashi, Judy (2019) "Digital Approaches to Translation History", *The International Journal for Translation and Interpreting Research* 11(2): 132-145.