
Using corpora and corpus analysis tools in the practice of literary translation

Josep Marco Borillo

 <http://interfas.univ-tlse2.fr/lamaindethot/1034>

Référence électronique

Josep Marco Borillo, « Using corpora and corpus analysis tools in the practice of literary translation », *La main de Thôt* [En ligne], 9 | 2021, mis en ligne le 26 août 2024, consulté le 26 août 2024. URL : <http://interfas.univ-tlse2.fr/lamaindethot/1034>

Using corpora and corpus analysis tools in the practice of literary translation

Josep Marco Borillo

PLAN

Introduction

1. Corpora and professional translators
2. Parallel corpora and literary translation
3. A systematic approach to the use of corpora and corpus analysis tools in literary translation
4. Concluding remarks

TEXTE

Introduction

- 1 The work of professional translators nowadays is more computer-bound than ever. Translators tend to look up unfamiliar words in electronic or online dictionaries, and most of their searches for factual or encyclopedic information take place in the framework of the internet. Most legal and technical translators use translation memories on a regular basis, often as a job requirement, and more and more translation commissions every day no longer involve translating from scratch but post-editing the output of a machine translation (MT) system. Literary translation practice stands out against this backdrop as particularly resistant to some forms of technology. Of course literary translators have long been using word processors and the aid provided by the internet, but they usually regard translation memories as irrelevant to their task, as terminological and formulaic repetition does not play an important role in literary texts, and MT as an outlandish invention. Even though recent research has remarkably qualified the alleged incompatibility between literary translation and these technologies (see e.g. ROTHWELL, 2021 on the use of a translation memory to retranslate a French classic, and TORAL and WAY, 2018 on reader assessment of post-edited MT), it is still true to say that their use by literary translators is far from common practice.

- 2 But how about corpora? What role do corpora play in literary translation? What sort of assistance can they provide at the different stages of the translation process? In this paper I will try to explain and illustrate some of the uses that literary translators can make of electronic corpora through the tools offered by Sketch Engine (<https://www.sketchengine.eu/>). In section 1 an overview will be presented of surveys on the use of corpora by professional translators. In section 2 the potential of one particular kind of corpora, parallel corpora, will be explored as a source of actual equivalents and solution types. Section 3 will draw on Youdale's (2020) proposal to provide a systematic approach to the use of corpus analysis and management tools at the different stages of the literary translation process and apply it to my own Catalan translation of Jane Austen's *Pride and Prejudice*. Section 4 will offer some concluding remarks.

1. Corpora and professional translators

- 3 Available surveys show that professional translators do not seem to be very keen on using corpora. Frérot (2016, 46 and ff.) reviews a number of these surveys, one of which was conducted within the MeLLANGE Project, with 740 respondents. According to Frérot (ibid.), "the data collected brought to light that corpora and concordance use were not common practice among professional translators". Gallego-Hernández (2015) distributed a questionnaire that was completed by 581 professional translators, most of whom (526) were Spanish. The author reports that "[n]early 50% of respondents [*he refers to the 526 Spanish translators*] stated that they 'never/almost never' use corpora to translate. Barely 30% answered 'sometimes', and only 18% indicated that they use corpora 'often/very often'" (GALLEGO-HERNÁNDEZ, 2015). Various reasons have been put forward to account for this state of affairs. Frankenberg-Garcia (2015, 354) ascribes it to lack of competence ("they simply do not know how to use them well enough to understand their potential benefits"), which might just stem from lack of adequate training, or, to put it in Peraldi's (2019, 270) words, to "the uneven and disparate teaching of corpus-based TS in master's degrees". But, according to Frankenberg-Garcia (2015, 355), lack of training is a direct consequence of the fact

that “there is no particular pressure from the market to train translators to use corpora”, as opposed to translation memories and, more recently, MT. Peraldi (2019, 270) quotes from a 2018 study of the European language industry which reveals that over 50% of professional translators nowadays use MT and over 80% “have integrated CAT tools in their work environment”, but the technological trend does not seem to encompass corpora. Peraldi (2019, 271) further refers to “the strong resistance to technology that has long characterised the translating community” and has led them to see technology as a threat and an imposition – even if eventually they have had no choice but to avail themselves of it. This attitude would be much more deeply entrenched in literary translators. Horenberg (2019, 10) claims that “[o]pportunities and benefits of TTech are generally easily disregarded by literary translators and researchers, as they commonly say that they fear technologies will limit translators’ creativity, or that their skills and therefore jobs will become superfluous”, and the same point is made by Youdale (2020). The latter author hastens to add that “there have in recent years been a few interesting experiments designed to see whether these technologies might in fact have a role to play in literary translation” (ibid.). Most of these experiments focus on machine translation, as seen in the previous section, and are research- rather than practice-driven. It remains a fact that CAT tools and MT play a very peripheral role in literary translation practice; but so do corpora, and that is much more intriguing.

2. Parallel corpora and literary translation

- 4 It falls outside the scope of this paper to offer an account of any length of what corpora have to offer translators. Several authors have highlighted the various uses that monolingual, comparable and parallel corpora can be put to in both translator training (e.g. ZANETTIN et al., 2003; OSTER, 2007; BEEBY et al., 2009) and translation practice (BOWKER, 1998; BERNARDINI, 2006; ZANETTIN, 2002, 2012). In this section I will concentrate on the potential of a particular kind of corpora, parallel corpora, for literary translators. Why parallel corpora? I beg leave to quote myself for an answer (MARCO, 2019, 40):

Translators are sure to welcome parallel corpora wherever available, as they provide a wealth of *actual* translation solutions at the touch of a button. The advantages of parallel corpora over bilingual dictionaries are obvious: the translation equivalents they offer have actually been used by someone (usually a professional translator) in a specific context, and there are as many of them as matches for a given query. Not for nothing are parallel corpora also known as *translation corpora* in the literature (e.g. Johansson 2007: 5).

- 5 In addition to actual equivalents in context, parallel corpora also offer, on a more abstract level, “repertoires of strategies deployed by past translators” (ZANETTIN, 2002, 11), that is of ways in which a particular problem has been solved. Patterns observed in particular areas (e.g. certain kinds of culture-specific items) may help translators make better-informed decisions. If a translator, for example, wavers between rendering a unit of measure (e.g. *inch*) as its lexical *prima facie* equivalent in the target language or as its equivalent in the metric system, corpus data may help them choose one or the other.
- 6 There are certain areas of language contrast that can turn out to be particularly problematic in literary translation because they are apt to raise stylistically relevant issues. Luis Magrinyà’s (2015) *Estilo rico, estilo pobre* (‘Rich style, poor style’) faces a number of such issues on the basis of the author’s long experience as an editor, writer and translator. The book is not specifically focused on translation, but many of the examples provided come from translated texts. *Rich style* is an ironic way of referring to instances of embellished, ornate writing aiming to fulfil an ideal of “richness, variety, beauty, precision, nuance, functionality, intensity, even energy” (2015: 28).¹ The problem is that “those noble aims are not always pursued along the right paths”. *Poor style*, the opposite of beautiful writing, stems from lack of attention, of reflection, from “indolence, automatism, lazy ignorance of the resources offered by the language” (2015, 29). My point in this section is that parallel corpora can help translators find out how the problems mentioned by Magrinyà are solved in translations and to what extent style enrichment or impoverishment actually occurs.
- 7 Let us first look at an instance of poor style. Magrinyà (2015, 121) argues that in Spanish literary texts we often come across instances of characters *sitting* where the sitting position draws attention to itself.

The same would apply to other positions, such as standing or lying. Magrinyà argues that specification of those positions is unusual in Spanish, which favours the position-neutral verb *estar* ('to be', in a circumstantial sense), and often derives from literal translation of the English verbs *sit*, *lie* or *stand*, which involve specification and are quite frequent. However, he goes on to add (2015, 124), "[m]any translators have long been aware of this semantic parasitism, make good use of the verb *estar* and rightly spare us a number of strange details". Which attitude prevails in actual translation? A search was conducted for the lemma *sit* in the fiction component of P-ACTRES, an English-Spanish parallel bi-directional corpus compiled at the Universidad de León in Spain (see SANJURJO-GONZÁLEZ and IZQUIERDO, 2019 for details). Out of 100 matches, 76 had to be discarded because they were irrelevant for different reasons – they designated the action of sitting, the sitting position was relevant, *sit* co-occurred with an inanimate subject, it was part of a phrasal verb (*sit back*, *sit for somebody*, *sit up*) or was used transitively (*sit an exam*). The remaining 24 parallel concordances might be claimed to refer to situations where the sitting position was either superfluous or irrelevant. In 15 of those instances, reference to the sitting position was retained all the same, whereas in 9 it was not. That is how the balance stands. As to the range of solutions involving non-specification of the sitting position, Table 1 shows some of them, with the verbs matching *sit* highlighted in red. The verbs used in the first and third examples (*quedarse* and *permanecer*) are similar in meaning and could be paraphrased as 'remain' in English. The second uses the copulative verb *estar*, similar to circumstantial *be*. And the fourth omits not only the verb *sit* but the whole clause "I'll sit at the bus-stop", probably because the information can easily be retrieved from context. Access to data of this kind enables a translator to fulfil the double aim of determining where translation practice stands with regard to Magrinyà's argument and finding translation solutions which involve non-specification of the sitting position, provided they were looking for them.

ST	TT
1523337: Here my mother's face would glow softly from within, and she would sit quietly, remembering. (FHL1E.s155)	Al llegar a ese punto la cara de mi madre resplandecía levemente y se quedaba callada recordando. (FHL1S.s157)
1412530: Are we supposed to sit here and wait for you all night? " (FGO1E.s1754)	¿O es que vamos a estar esperándolo toda la noche? (FGO1S.s1736)

1522771: It was so much harder for me to sit quietly than it was for her, and I do n't believe it had anything to do with my age. (FHL1E.s134)	A mí me costaba mucho más que a ella permanecer en silencio, y no creo que se debiera a mi edad. (FHL1S.s135)
2347084: The next one will come soon, and in the meantime I'll sit at the bus stop and have an ice cream. (FSP1E.s886)	Pasará otro dentro de unos minutos, y mientras tanto, me comeré mi helado. " (FSP1S.s879)

Table 1. Translation solutions for *sit* involving non-specification of the sitting position in the fiction component of P-ACTRES

- 8 The same operation was carried out for *stand* and *lie* on the P-ACTRES corpus. Out of 100 parallel concordances, the search for *stand* yielded 46 relevant cases (i.e. cases where the standing position could be regarded as superfluous), 7 involving specification, 39 non-specification of the standing position. The query for *lie* yielded 23 relevant matches, 8 including specification, 15 non-specification of the lying position. Of these three verbs, then, *sit* is the most prone to position specification in Spanish translation.
- 9 On the other hand, a very interesting case of *rich* style can be found in what Magrinyà refers to as “the talkative verbs” (“los verbos parlanchines”, 2015, 47), i.e. reporting verbs used to designate speech acts in direct, indirect or free indirect discourse. Magrinyà (2015, 47-48) claims that reporting verbs fulfil a phatic function – they serve to keep the communication channel open and more often than not they are not properly *read*, or at least they are not read with the same level of awareness as other parts of the text. They are almost taken for granted, so they should not draw attention to themselves. English tolerates repetition quite well and tends to use *say* in most cases. Spanish is not so tolerant and avails itself of a larger number of options, as witnessed by Magrinyà’s long list (2015, 48-50). But – the argument goes on – lexical variety may become a pitfall, as reporting verbs are not interchangeable, each one possessing its own nuances. And – one is tempted to add – repetition in an English original may be deliberate and, consequently, a feature of style.
- 10 Quite independently of Magrinyà’s claims (in fact, prior to them), a lively debate took place on ACeTt’s mailing list in May 2012 on the translation of *say* as a reporting verb in dialogues. ACeTt is the main professional association of book translators in Spain, so this debate can be regarded as representative of attitudes towards a stylistic issue that has ceased to be minor, to judge by the amount and tone of

the contributions. A selection was published in dialogue form in issue number 43 of *Vasos Comunicantes*, ACeTt's journal. The basic claim, made by some, which can be taken as a point of departure, was that replicating repetition of *say* in Spanish sounds simplistic and had better be avoided. Along similar lines, other translators said that the generalised use of *say* is a feature of the English language, not of any particular author, so that rendering it invariably as *decir* would turn an unmarked feature in English into a marked one in Spanish. Finally, it was also argued that repetition of *say* is a feature of contemporary English-language fiction, not of the English literature as a whole or of a particular author. Some translators further claimed that the general use of *decir* is not so rare in Spanish-language letters, and gave some examples. And it must also be borne in mind that a translator's decision to render *say* consistently as *decir* may clash with editorial practices and preferences. The translator does not always have the last word on this.

- 11 Again, as in the case of *sit/lie/stand*, the relevant question here is what happens in reality. In this case, two parallel corpora were queried for the form "said" and its target text matches: P-ACTRES for English-Spanish and COVALT (Valencian Corpus of Translated Literature, compiled by the eponymous group at Universitat Jaume I in Castelló, Spain)² for English-Catalan. After manual removal of cases where *said* did not occur in dialogue, P-ACTRES yielded 82 relevant examples (out of 100 matches), with 52 rendered as *decir*, 27 as other verbs and 3 as zero (not translated). The English-Catalan sub-corpus of COVALT yielded 91 relevant matches (again, out of 100), 75 of which were translated as *dir* (the *prima facie* equivalent of *say* in Catalan), 14 as other verbs and 2 as zero (not translated). As to alternatives to *decir/dir*, respectively, they included *exclamar* ('exclaim'), *observar* ('remark, observe'), *asegurar* ('assure, guarantee') or *mentir* ('lie') in Spanish and *respondre* ('answer'), *insistir* ('insist'), *preguntar* ('ask') or *contestar* ('answer') in Catalan. These results show that repetition of the *prima facie* equivalent of "said" is favoured in Spanish in over 63% of the cases and in Catalan in over 82%. As to the verbs used as alternatives, Spanish shows more variation than Catalan, with more semantic nuances added to the neutral meaning of English *say*. In other words: Spanish shows a richer style.

3. A systematic approach to the use of corpora and corpus analysis tools in literary translation

- 12 The previous section just hints at ways in which parallel corpora can help literary translators make better-informed decisions on the basis of data provided by published translations. The number of issues that could be settled in this way is virtually limitless, so the whole previous section may strike the reader as being rather anecdotal in nature. A couple of issues were illustrated, but a comprehensive list is out of the question. Moreover, other types of corpora could be brought into the picture in addition to parallel corpora. Can we be more systematic in our approach, though? Is it possible to comprehensively map the kinds of services that corpora can render literary translators?
- 13 In recent years a couple of such maps have been drawn taking the stages of the translation process as their starting-point. Peraldi (2019) conducted an experiment with a number of translators at the French Ministry of Economy and Finance with a view to raising awareness of the potential of corpora as translation aids. Three monolingual corpora were made available to the translators through Sketch Engine. Results allowed the author to identify four “key functionalities” (2019, 280 and ff.) in corpus use. The first is understanding key concepts in the source language, especially through definitions in context. The second is exploring phraseology in the target language. The third is finding equivalents, mainly in an inductive way, as no parallel corpora are available. The fourth is choosing between synonyms through the Word Sketch Difference utility in Sketch Engine, in an attempt to enhance idiomaticity. The first functionality clearly pertains to the comprehension, or source text analysis, stage of translation; the third is directly linked to the central reformulation stage; and the second and fourth could either be placed in the reformulation or in the revision phase.
- 14 Youdale’s (2020) alignment with the different stages of the translation process is still more explicit. He suggests using an approach that blends close and distant (MORETTI, 2013) reading techniques in order

to maximise the amount of stylistic information the translator has at their disposal. By means of corpus-based (e.g. Sketch Engine) and visualisation (e.g. CATMA and Voyant) tools, distant reading (2020)

can reveal the existence of potentially stylistically relevant linguistic patterns which close reading is either unlikely to reveal or is incapable of detecting, such as lexical variety and the use of words and phrases which are repeated, but are spread over a number of chapters.

- 15 Close and distant reading can be used in four specific ways, the first three of which are closely related to the three stages of the literary translation process, whereas the last one focuses on the translator's style (YOUDALE, 2020):
 1. In analysing the ST after (and possibly even during) initial reading, and in helping the translator to formulate specific translation goals.
 2. In undertaking the first draft of the translation.
 3. In comparing the ST and draft translation, with a view to seeing if certain translation goals have been achieved, and what has actually happened in the translation process.
 4. In the auto-analysis of translator style. This can be done at any time, but is more likely to yield meaningful results after a translator has completed several translations.
- 16 In what follows, I will draw on the first three ways mentioned by Youdale and try to work out their implications for the analysis of my own Catalan translation of Jane Austen's *Pride and Prejudice*, *Orgull i prejudici*, published by the Valencian publisher Bromera in 2018. No distant reading techniques were used during the translation process, but they can be used now to determine to what extent my overall stylistic goals were achieved and what areas could have benefited from the use of that kind of techniques.
- 17 Sketch Engine is a very popular online tool for compiling and querying corpora (see e.g. KILGARRIFF et al., 2014). Users can interrogate existing corpora or build their own (whether monolingual or multilingual), and any corpus can be analysed through a number of utilities, such as wordlists, concordancing, n-grams, keywords, word sketches (main collocates of the search word classified according to grammatical criteria) and word sketch difference (collocations of two words

compared). Therefore, due to its potential, I decided to use Sketch Engine to compile and analyse my corpora. I compiled three very simple corpora: one with the source text, one with the target text and a parallel one made up of both ST and TT aligned. Alignment was performed with LF Aligner, which allows the user to edit the automatic alignment and to generate a .tmx file, which I then uploaded to Sketch Engine. Youdale (2020) raises the issue of the cost/benefit relationship of alignment and reports a duration time of about six hours to manually revise the output of the aligner. My experience was very similar: it took me about five hours to align a 122,018-word long text and its translation. Manual editing is certainly a time-consuming task, but, when five hours are seen against the background of the whole amount of time invested in translating a book, the benefits to be drawn from analysing the parallel corpus may well be worth the effort.

- 18 Youdale (2020) mentions a set of “standard” CDR (close and distant reading) analyses that can be performed on a corpus in order to shed light on relevant aspects of style. Two of these analyses are shown to be particularly revealing because they are linked to the translator’s priorities (both in Youdale’s case and mine): average sentence length and lexical richness. Sentence length is not identical with sentence complexity, as a long sentence may be made up of coordinated or juxtaposed clauses that create no sense of complexity because there is no hierarchy; but both features will normally tend to overlap. In general terms, the longer a sentence, the likelier it is to show elements of complexity. One of my stylistic aims while translating Austen’s novel was to preserve its degree of syntactic complexity by respecting sentence boundaries whenever possible. This was no hard-and-fast rule: adjustments had to be made when this aim bumped against target language norms, such as typographical conventions in dialogue. But it was certainly a stylistic priority. As argued by Whitfield (2000, 114) and Marco (2018, 75), there is more to the sentence than just grammar. If grammar is perceived as selection and combination of constituents, word order matters in semantic terms, as it lays emphasis on some constituents at the expense of others; and, in addition to that, there is rhythm, which stems from the subtle interaction of content and the sequence and length of the elements making up the sentence. The rhythmic flow of a text may be substantially altered

by introducing large-scale (as opposed to local) changes in the number of sentences and their structure, hence my priority to respect sentence boundaries. Table 2 shows the results of my analysis as regards number of sentences and average sentence length in both source and target text. The TT contains 48 more sentences than the ST, which represents an increase of 0.82% – a very slight increase indeed. As to sentence length, the TT sentences are 1.64 words longer than the ST sentences on average, which represents an increase of 7.81%. Translations tend to be longer than source texts, as established by Frankenberg-Garcia (2009, 57), due to explicitation or other factors. In the present case, the ST has 122,018 tokens, whereas the TT comprises 132,604. If the number of sentences remains very similar, as just seen, a higher number of tokens is bound to result in a higher average sentence length. But all these differences, whether in number of total tokens or in average sentence length, operate within a reasonable margin, which never exceeds 10%.

	ST	TT
Number of sentences	5,811	5,859
Average sentence length	20.99	22.63

Table 2. Number of sentences and average sentence length in *Pride and Prejudice* and Marco's Catalan translation

- 19 Lexical richness or variety is usually measured through the type/token ratio, or number of different words (types) in a text or corpus in proportion to the total number of words (tokens). The problem with this measure is that it does not work with languages displaying wide differences in morphological variation. Catalan is much more highly inflected than English, and each inflected form is counted as a type; therefore, any Catalan translation from English is bound to have a significantly higher number of types than its corresponding source text. This problem can be solved by grouping together different inflected forms of the same lemma, and use the lemma/token ratio as an alternative to the type/token ratio. The lemma/token ratio is expressed as a percentage: number of different lemmas x 100 / number of tokens. Results for a comparison based on this measure can be seen in Table 3. The main thing about those results is that lexical richness has not decreased as a consequence of the translation process, i.e. no lexical simplification has occurred. The

number of lemmas is higher for the Catalan translation, but so is the overall number of tokens; therefore, the lemma/token ratio is almost identical for both texts. A significantly lower ratio for the target text would have signalled a (presumably unwanted) lexical impoverishment.

	ST	TT
Number of tokens	122,018	132,604
Number of lemmas	4,737	5,145
Lemma/token ratio	3.88	3.89

Table 3. Lemma/token ratio in *Pride and Prejudice* and Marco's Catalan translation

- 20 Youdale argues (2020) that a large proportion of types or lemmas in a corpus are hapax legomena, i.e. lemmas that occur only once, whereas only a (relatively) few occur very often. Ranking lemmas according to their absolute frequency in a text or corpus can provide an overall picture of the balance between diversity and repetition, both of which can be stylistically significant. A larger number of hapaxes in the target text together with a lower score for lemmas occurring quite frequently might signal a tendency to avoid repetition, and vice versa. Table 4 shows such a ranking for the source and target texts under scrutiny here. Differences across texts are not very remarkable. They are perhaps more marked as regards the number of hapaxes and of lemmas occurring twice, which indicates more lexical variety in the translated text for those two ranks; but the figures are quite similar for the remaining ranks. If these figures are expressed not as raw but relative frequencies (percentages), differences tend to be further neutralised, as the total number of different lemmas is higher for the target than the source text.

Absolute frequency of lemma	Number of lemmas ST	Number of lemmas TT	Relative frequency ST (%)	Relative frequency TT (%)
1	1,680	1,918	35.46	37.27
2	660	736	13.93	14.30
3	424	421	8.95	8.18
4	248	289	5.23	5.61
5	179	219	3.77	4.25
6	137	172	2.89	3.34

7	104	124	2.19	2.41
8	99	107	2.08	2.07
9	90	81	1.89	1.57
10	76	79	1.60	1.53
>10	1040	999	21.95	19.41
Total	4,737	5,145		

Table 4. Frequency distribution of lemmas in *Pride and Prejudice* and Marco's Catalan translation

- 21 The potential uses of corpus analysis through Sketch Engine illustrated so far are just quantitative. They deal with figures providing a general overview of particular aspects, also known as *document-level metrics* (e.g. LYNCH, 2014, 376). But other utilities in the suite allow for more qualitative inquiries. The keywords utility, for example, provides a list of the most distinctive words (both single words and multi-word strings) in the corpus, i.e. of those lexical items that are significantly either more or less frequent in the corpus in question than in a reference corpus – in the present case, English Web 2013. Keyness is often important in terms of content, and it may have important implications for characterisation or style. The 50 top-ranking items on the list of key single words include many character and place names, which may be said to be rather predictable, but also such other terms as *civility*, *amiable*, *agreeable*, *politeness* and *imprudent*. What all these words have in common is that they often feature in character description, character assessment (by the narrator or other characters), etc. In short, they are closely linked to characterisation, but also to themes that cut across all Jane Austen novels. As argued by Page (1972, 56-60) and emphasised by Alsina (2008, 89), words describing the characters' social attitude or intellectual and moral features are particularly relevant in those novels, not only because they occur frequently (as witnessed by their presence on the keyword list) but also because they often have precise meanings, to the extent that apparent synonyms cannot be used interchangeably. This raises two problems for translators. At the reformulation stage, they need to convey those subtle nuances of meaning by teasing apart the implications of a given word from those of its semantic neighbours. Both at the reformulation and the revision stages, they may consider checking for consistency in the translation of these es-

pecially sensitive words. That could be achieved by means of a bilingual glossary developed by the translator themselves, but being able to look up words and expressions in a parallel corpus may prove to be very helpful for this kind of check.

- 22 To illustrate this, I will focus on two of the words on the above list, *civility* and *amiable*. The lemma *civility* occurs 49 times in *Pride and Prejudice* with two distinct meanings: in the singular it means “formal politeness and courtesy in behavior or speech” (LEXICO, 2020), whereas the plural form refers to “polite remarks used in formal conversation” (ibid.). The former denotes a (relatively abstract, or general) social attitude, whereas the latter is applied to specific speech acts. The plural form occurs 7 times in the ST and is always translated as “compliments”. The singular form occurs 42 times and is rendered as “cortesia” (‘politeness, courtesy’) 35 times, “educació” (‘politeness’) 3 times, “gentilesa” (‘kindness’) twice, “bona educació” (‘politeness, good breeding’) once, and “compliments” (‘civilities’) once. Thus, translation shows absolute consistency in the plural form and a high degree of consistency in the singular form (35 out of 42 times), but also some leeway for variation, which can be justified by the nuances of meaning the word takes on in different contexts. On the other hand, the lemma *amiable* occurs 36 times in the ST, but its meaning is more difficult to pin down, which is reflected in a relatively high number of different translation solutions. Oxford Dictionaries (LEXICO, 2020) defines *amiable* as “having or displaying a friendly and pleasant manner”, and then goes on to cite “kind” and “lovely, lovable” as the original senses of the word. So *amiability* seems to be related either to the manners or to the deeper qualities of a person, which make him or her worthy of love or appreciation. Moreover, in Austen’s novel (as can be gleaned from the information provided by Word Sketch), *amiable* co-occurs with people (“Charlotte”, “companion”, “neighbour”, “woman”, “daughter”, “man”) but also with abstract nouns (“qualifications”, “light” in a metaphorical sense, “quality”, “feeling”). It follows from this that consistency would not be a desirable goal in the translation of *amiable*. In fact, it is variously translated as “amable” (‘kind’) 7 times, “bo” (‘good’) 7 times, “bona persona” (‘good person, good-natured’) 4 times, “bondadós” (‘good-natured’) 3 times, “bondat” (‘goodness’) 3 times, “gentil” (‘kind’) twice, and once as “amablement” (‘kindly’), “amistós” (‘friendly’), “aptitud”

(‘qualification’), “ben intencionat” (‘well-meaning’), “cordial” (‘warm-hearted’), “estimable” (‘lovable’), “gentilesa” (‘kindness’), “lloable” (‘laudable’), “simpàtic” (‘agreeable’) and “favourable” (‘favourable’). At least 20 of these translation solutions can be grouped under the heading “good/good-natured/worthy of love or affection”, the rest either displaying the sense of “kind/friendly” or having more neutral implications, especially when *amiable* is applied to an abstract entity.

- 23 The width of the word’s semantic range and its resulting vagueness may well justify variability in the translation solutions. However, variation is not always justified, and my performance could certainly have benefited from the use of distant reading tools such as those provided by Word Sketch. There are two cases in which a particular collocation is repeated twice, the second occurrence being an ironic echo of the first. These cases were not detected by my close reading radar and I translated them differently, when consistency was clearly needed. The first is “amiable Charlotte”, used by Mr Collins to refer to his intended, Charlotte Lucas, either in free indirect discourse or plain narrative. Under the narrator’s control, then, this sequence is used to insist on Mr Collins’ pompousness and predictability, i.e. for characterisation purposes. “Amiable” in these two cases was variously translated as “amable” and “gentil”, which are certainly synonyms; but synonymy here misses the point, as repetition was needed instead. The same holds true for the second case, “amiable light”. After having been informed by Mr Darcy of the real nature of the dealings between himself and Mr Wickham, Elizabeth Bennet concludes, during a conversation with her sister Jane, that it would not be wise to acquaint the neighbourhood with this information: “The general prejudice against Mr. Darcy is so violent, that it would be the death of half the good people in Meryton to attempt to place him in an amiable light”. Later on, in the course of her visit to Derbyshire, Elizabeth hears Mr Darcy’s housekeeper praise her master warmly and thinks: “In what an amiable light does this place him!”. Here the irony is aimed at herself, since she echoes her own words when her opinion of Mr Darcy was much lower than it is now. “[T]o place him in an amiable light”, in the first occurrence, is translated as “per mirar-lo amb bons ulls” (literally, ‘to look at him with good eyes’); “[i]n what amiable light”, in the second, as “Quina manera més amable” (‘What an amiable/kind way’). Both renderings have to do with positive attitude, but, as in the previ-

ous example, synonymy misses the point and the ironic effect is lost through avoidance of repetition. The point here is that avoidance of repetition is not a deliberate choice but the outcome of lack of awareness, which may often accompany exclusively close reading.

- 24 As mentioned above, the keywords utility in Sketch Engine also affords the possibility to identify multi-words, i.e. strings of two or more words that stand out when the focus corpus (or corpus under scrutiny) is compared to a reference corpus. As in the case of single words, the first 50 key multi-words on the list include predictable combinations entailing little difficulty or relevance from a thematic point of view, such as “whole party”, “dear sister”, “young man” or “short silence”. But there are others that are likely to ring a bell in the translator/analyst’s mind, either because they are aware of having come across them several times while translating or because they relate to matters of thematic concern, characterisation or style. One of these combinations, “humble abode” (5 times), illustrates the former possibility. It is invariably used by pompous Mr Collins to refer to his home with an air of false humility. It is part of Mr Collins’ idiolect, then, and contributes to his characterisation. I was fully aware of this and consequently translated it in a consistent way, always as “modesta llar” (‘modest/humble home’). But the case was otherwise with “real character”, which may be said to tie in with one of the themes of the novel – the interplay between reality and appearance and the deceptiveness of appearances in social life. This collocation occurs 5 times in the ST, and it is variously translated as “veritable caràcter” (‘true character’) twice and once as “reputació” (‘reputation’), “tipus de persona” (‘kind of person’) and “de veritat” (‘truly, in truth’). Even though it may not be strictly necessary in terms of characterisation or style to be consistent here, as no irony is intended and the words are placed in the mouth or the consciousness of different characters, more could have been done by way of homogenisation had the translator been aware of the recurrent use of this expression.
- 25 The latter example may serve to introduce another potential use of Sketch Engine through its Word Sketch Difference utility, that of discriminating between the collocational profiles of near-synonyms. A translator may reasonably waver between two words with almost identical meanings, not on the ground of their particular meaning but of the degree of typicality of their collocation with another word. In

the case just referred to (“real character”), Catalan *caràcter* (‘character’) might be modified either by *veritable* or *autèntic*, both meaning ‘true, real’. Drawing on a general web-based Catalan corpus, Word Sketch Difference tells us that *veritable* typically collocates with *naturallesa* (‘nature’), *democràcia* (‘democracy’), *revolució* (‘revolution’), *essència* (‘essence’), and *autèntic* with *referent* (‘referent, reference’) and *sabor* (‘flavour’). Apart from the fact that *veritable* is rather more frequent than *autèntic* in raw terms (7,075 vs. 4,365 matches), the former seems to attract abstract nouns more strongly than the latter, so it would be a more likely collocate for *caràcter*. If this inference is true, my choice of adjective seems to be borne out by Word Sketch Difference data. Therefore, this tool may prove useful when checking for idiomaticity in the target language either during the reformulation or the revision stage.

4. Concluding remarks

- 26 It makes no sense for literary translators not to be using corpora more often in their professional practice. One can understand their reservations about CAT tools or MT, as these tools (especially the latter) may perhaps be seen as not exactly enhancing creativity. But it is more difficult to understand why corpora are not exploited on a regular basis, as they pose no threat either to creativity or to job availability.
- 27 This paper has illustrated two possible ways of using corpora in literary translation practice. The first way (section 2) concerns parallel corpora and their potential to provide actual translation solutions and repertoires of strategies. More particularly, parallel corpora were used to check to what extent some stylistic claims made by Magrinyà (2015) were borne out by corpus data and what alternatives were provided to the solutions criticised by that author. In this respect, the main difficulty might lie in the fact that parallel corpora of literary texts are not easily available, often for copyright reasons. Some of these corpora are lodged in servers belonging to universities or other research institutions and can only be accessed on demand. Professional literary translators may not have the time or indeed the patience to go through the motions and may understandably settle for more readily available resources. Free access to those parallel cor-

pora would serve the needs and interests of all concerned, both university-based research groups and professionals, with no damage to the legitimate interests of copyright holders (source texts authors, translators and publishers), as nobody uses corpus concordances to reconstruct a whole text from its fragments, patchwork-wise. This is a crucial aim, and there should be more people involved in achieving it, even through legal amendments if necessary.

- 28 The second way (section 3) concerns the integrated use of corpus analysis tools during the translation process through a combination of close and distant reading techniques, as suggested by Youdale (2020). It would involve analysing the source text by means of the utilities provided by Sketch Engine before, during and after the reformulation stage, as well as analysing the target text and the source and target texts aligned after that stage. In section 3 a number of ways to exploit those utilities were illustrated. They may be used to determine whether translation goals (e.g. with regard to sentence average length and lexical richness) have been attained and to check for the consistency or idiomaticity of certain translation options through Keywords, Word Sketch or Word Sketch Difference. The main issue in this respect is whether practising translators will consider the benefits to be drawn from the parallel corpus analysis worth the effort of aligning source and target, with the manual revision of the alignment output it all implies. It is a time-consuming task, but the time investment may become diluted in the sustained effort of translating, say, a 300-page long text, which often spans months. It is basically a matter, then, of raising translators' awareness of the benefits of analysing their own translations once aligned with the source text with a view to uncovering patterns that may remain hidden in any close reading of a text – no matter how close. The aim thus envisaged is not to replace human by machine output, as in other computer-based tools, but to join human and machine forces.

BIBLIOGRAPHIE

ALSINA, Victòria, 2008, *Llengua i estilística en la narrativa de Jane Austen. Les traduccions al català*, Vic, Eumo.

BEEBY, Allison, Patricia RODRÍGUEZ INÉS & Pilar SÁNCHEZ-GIJÓN (eds.), 2009, *Corpus Use and Translating: Cor-*

pus Use for Learning to Translate and Learning Corpus Use to Translate, Amsterdam and Philadelphia, John Benjamins.

BERNARDINI, Silvia, "Corpora for translation education and translation practice: Achievements and challenges", in Proceedings of the Third International Workshop on Language Resources for Translation Work, Research and Training (LR4Trans-III). Available online at: http://mellange.eila.univ-paris-diderot.fr/bernardini_lrec06.pdf.

BOWKER, Lynne, "Using specialized monolingual native-language corpora as a Translation resource: A pilot study", in *Meta*, 1998, n° 43(4), p. 631-651.

FRANKENBERG-GARCIA, Ana, "Are translations longer than source texts? A corpus-based study of explicitation", in Beeby, Allison, Patricia Rodríguez Inés & Pilar Sánchez-Gijón (eds.), *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, 2009, Amsterdam and Philadelphia, John Benjamins, p. 47-58.

FRANKENBERG-GARCIA, Ana, "Training translators to use corpora hands-on: Challenges and reactions by a group of thirteen students at a UK university", in *Corpora*, 2015, n° 10(3), p. 351-380.

FRÉROT, Cécile, "Corpora and corpus technology for translation purposes in professional and academic environments. Major achievements and new perspectives", in *Cadernos de Tradução*, 2016, n° 36(1), p. 36-61.

GALLEGO-HERNÁNDEZ, Daniel, "The use of corpora as translation resources: A study based on a survey of Spanish professional translators", in *Perspectives*, 2015, n° 23(3), p. 375-391. doi:10.1080/0907676X.2014.964269

HORENBERG, Lisa W., 2019, *Using CAT in Literary Translation - How Tools May Support Translators in Source-Text Analysis, Translation and Retranslation: A Case Study of "Mr Loveday's Little Outing"*, Utrecht University (master's thesis).

JOHANSSON, Stig, 2007, *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies*, Amsterdam, John Benjamins.

KILGARRIFF, Adam, Vít BAISA, Jan BUŠTA, Miloš JAKUBÍČEK, Vojtěch KOVÁŘ, Jan MICHELFEIT, Pavel RYCHLÝ & Vít SUCHOMEL, "The Sketch Engine: ten years on", in *Lexicography ASIALEX*, 2014, n° 1, p. 7-36.

LYNCH, Gerard, "A supervised learning approach towards profiling the preservation of authorial style in literary translations", in Proceedings of the 25th International Conference on Computational Linguistics (COLING), Dublin, Ireland, Association for Computational Linguistics. <https://www.aclweb.org/anthology/C14-1037.pdf>

MAGRINYÀ, Luis, 2015, *Estilo rico, estilo pobre*, Barcelona, Debate.

MARCO, Josep, "Els patrons sintàctics en les novel·les de Jane Austen traduïdes per Jordi Arbonès", in *Quaderns. Revista de traducció*, 2018, n° 25, p. 71-85.

MARCO, Josep, "Living with parallel corpora: The potentials and limitations of their use in translation research", in Doval Reixa, Irene & M. Teresa Sánchez-Nieto (eds.), *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*, 2019, Amsterdam and Philadelphia, John Benjamins, p. 39-56.

MORETTI, Franco, 2013, *Distant Reading*, London and New York, Verso.

OSTER, Ulrike, “Working towards autonomy: Corpora in the translation classroom”, in Wotjak, Gerd (ed.), *Quo vadis Translatologie? Ein halbes Jahrhundert universitäre Ausbildung von Dolmetschern und Übersetzern in Leipzig. Rückschau, Zwischenbilanz und Perspektiven aus der Aussensicht*, 2007, Berlin, Timme, p. 311-325.

LEXICO, «Oxford English and Spanish Dictionary, Thesaurus, and Spanish to English Translator», 28 December 2020, <URL: <http://www.lexico.com>>.

PAGE, Norman, 1972, *The Language of Jane Austen*, Oxford, Basil Blackwell.

PERALDI, Sandrine, “Integrating corpus-based tools in translators’ work environment: Cognitive and professional implications”, in *Revista Internacional de Organizaciones*, 2019, n° 23, p. 265-292. <https://doi.org/10.17345/rio23.265-292>

ROTHWELL, Andrew, “Retranslating Zola in stereo with CAT tools”, paper presented at the #Translating Europe Workshop 2020 – La traduction littéraire et SHS à la rencontre des technologies de la traduction: enjeux, pratiques et perspectives, Université Toulouse Jean Jaurès, 7 February 2020.

SANJURJO-GONZÁLEZ, Hugo & Marlén IZQUIERDO, “P-ACTRES 2.0: A parallel corpus for cross-linguistic research”, in Doval Reixa, Irene & M. Teresa

Sánchez-Nieto (eds.), *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*, 2019, Amsterdam and Philadelphia, John Benjamins, p. 215-231. <https://doi.org/10.1075/scl.90.13san>

TORAL, Antonio & Andy WAY, “What level of quality can Neural Machine Translation attain on literary text?”, in Moorkens, Joss, Sheila Castilho, Federico Gaspari & Stephen Doherty (eds.), *Translation Quality Assessment: From Principles to Practice*, 2018, Cham, Springer, p. 263-287.

WHITFIELD, Agnes, “Lost in Syntax: Translating Voice in the Literary Essay”, in *Meta*, 2000, n° 45(1), p. 113-126.

YOUDALE, Roy, 2020, *Using Computers in the Translation of Literary Style: Challenges and Opportunities*, New York and London, Routledge.

ZANETTIN, Federico, “Corpora in translation practice”, in Yuste-Rodrigo, Elia (ed.), *LREC Workshop No. 8: Language Resources in Translation Work*, 2002, Las Palmas de Gran Canaria, LREC, p. 10-14.

ZANETTIN, Federico, 2012, *Translation-driven Corpora*, Manchester, St. Jerome.

ZANETTIN, Federico, Silvia BERNARDINI & Dominic STEWART (eds.), 2003, *Corpora in Translator Education*, Manchester: St Jerome.

ANNEXE

All translations from languages other than English are my own.

Please visit <http://www.covalt.uji.es/> for further details.

AUTEUR

Josep Marco Borillo

Universitat Jaume I (Castelló, Spain) jmarco@uji.es