(DRAFT SUBMITTED TO CORPORA)

A corpus study of splitting and joining sentences in translation

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

The use of corpora in translation studies has risen dramatically over the past years, contributing towards a growing body of empirical research focusing not only on what differentiates translated from non-translated language, but also on the changes or shifts that translators make from source to target texts. Most of the latter studies are centred on sub-sentential elements, such as the contrastive use of particular lexis or grammar. However, translation shifts can transcend the level of the sentence. For example, sentences can be split or joined in translation, or there can be complex shifts that combine the two. While there is some research on sentence splitting, there do not seem to be many studies about sentence joining, or indeed sentence splitting and joining together. The present study seeks to address this gap. Using a bidirectional parallel corpus of Portuguese and English fiction, over 90 thousand source-text sentences and their corresponding text in translation were analysed from a quantitative perspective, and a closer look was taken at a sample of over one thousand parallel text segments involving sentence joining and splitting. The main findings were that in both translation directions (1) there was a strong tendency for sentence preservation, (2) the differences between sentence splitting and joining were not significant, and (3) changes in sentence boundaries were predominantly associated with the standardization or normalization of syntax and a tendency for explicitation.

Keywords translation, parallel corpora, sentence splitting, sentence joining, standardization, normalization, explicitation

1. Introduction

Corpus linguistics has had a great impact on translation studies over the past years. From a more applied perspective, comparable corpora in different languages have been used to explore L1–L2 equivalence and inform translation practice and education (e.g. Zanettin et al. 2003, Beeby et al. 2009, Frankenberg-Garcia 2015). From a more theoretical standpoint, corpora of translations have been studied alongside corpora of texts that are not translations in order to understand what sets translated language apart (e.g. Maurenen and Kujamäki 2004, Frankenberg-Garcia 2008, Delaere et al. 2012). Parallel corpora, in turn, have been used not only for very applied purposes – such as training machine-translation programs on data from human translations, or simply looking up how translators have dealt with particular words or phrases – but also to develop translation theories, by investigating whether there are any trends in the shifts that occur from source to target texts (e.g. Johansson and Hofland 2000, Johansson 2007, Pérez-Blanco 2009, Xiao and Dai 2014, Frankenberg-Garcia 2014, 2016). In the present paper, a bidirectional parallel corpus of English and Portuguese fiction is used to investigate sentence boundaries in translation. Its aim is to explore the more well-known phenomenon of sentence splitting in conjunction with sentence joining, and to examine how translators actually split and join sentences.

1.1 Previous studies

If we take a closer look at previous studies in translation involving parallel corpora, it becomes clear that most analyses reported in the literature are either purely lexical or

constrained by sentence boundaries. For example, Johansson and Hofland (2000) investigate modal auxiliaries, Pérez-Blanco (2009) looks at stance adverbials, Frankenberg-Garcia (2016) examines loan words, and Johansson (2007), Frankenberg-Garcia (2014) and Xiao and Dai (2014) analyse a variety of sub-sentential features.

While it is true that scholars may have been more interested in specific lexical and grammatical shifts, another reason why research involving the use of parallel corpora tends to focus on phenomena that occur within sentence limits is the fact that, when scrutinizing parallel concordances, it is simpler to focus on sub-sentential elements. This is because of how parallel text alignment is normally carried out. Thanks to the relative ease with which sentence boundaries can be identified automatically by means of punctuation marks and language-specific segmentation rules, it is common practice for source texts and translations to be segmented sentence by sentence, and then aligned such that whenever source-text and translation sentences do not coincide, alignment is carried out on a one-to-many or a many-to-one basis (Gale and Church 1993, Danielsson and Ridings 1997, Hofland and Johansson 1998, Véronis 2000, Barlow 2002). This process is illustrated in figure 1, where source-text sentence A and translation sentence A_1 are aligned one-to-one, but source-text sentences B and C are aligned on a many-to-one basis with translation sentence BC_1, and source-text sentence D is aligned on a one-to-many basis with translation sentences D_1, D_2 and D_3. While this procedure enables one to retrieve parallel concordances with matching text segments (and inspect sub-sentential elements within them), it does not automatically distinguish translation sentence A_1 (equivalent to one source-text sentence), from translation sentence BC_1 (equivalent to two source-text sentences joined together), from translation sentences D_1, D_2 and D_3 (equivalent to one third of a source-text sentence each). Although it is also possible to establish alignment links at the level of the word or clause (for example, see Hansen-Schirra et al. 2006 and Macken et al. 2008), this does not directly help understanding shifts that occur beyond those levels. Thus while it is fairly straightforward to retrieve parallel concordances to inspect shifts that occur within aligned segments, it is not as simple to gather information on sentence splitting and joining.

Figure 1 Alignment at sentence level

SOURCE TEXT	TRANSLATION	Alignment type
Sentence A	Sentence A_1	One-to-one
Sentence B Sentence C	Sentence BC_1	Many-to-one
Sentence D	Sentence D_1 Sentence D_2 Sentence D_3	One-to-many

It is nevertheless quite straightforward to compute the total number of sentences in source texts and translations separately. Serbina (2014) compared the number of source-and target-text sentences in a bidirectional German-English parallel corpus of one million words and found that there were more sentences in the German translations than in the English source texts, and fewer sentences in the English translations than in the German source texts. While this enables one to assert that there must have been some sentence splitting in English-German translations and some sentence joining in German-English translations, overall sentence counts only describe the combined effect of splitting and

joining sentences. They do not allow one to quantify exactly how much sentence splitting, joining and preservation there has actually been, or systematically locate what prompts translators to split or join sentences[1].

Another method used to investigate changes in sentence boundaries in parallel corpora is to carry out lexical searches for specific source-text words that signal the use of complex syntax, and then inspect the resulting parallel concordances to find out whether there has been any syntactic simplification in the translation. Using a parallel corpus of popular science and the Oslo Multilingual Corpus (OMC), Ramm (2004) looked up a selection of adverbs that mark relative clauses in German. She found that 6/42 relative clauses analysed in the OMC and 8/18 in the popular science corpus were upgraded to independent sentences when translated into Norwegian.[2]

Using a similar methodology and a one-million-word corpus of English-German business texts, Bisiada (2013) focused on the retrieval of parallel concordances containing concessive and clausal conjunctions such as *although*, *while*, *because*, *since*, *for* and *as* to investigate how German translators dealt with the more complex syntax entailed by them. Unlike most studies on sentence-boundary shifts, Bisiada also investigated German concessive and causal conjunctions introduced by translators without any prompt from equivalent words in source texts, thus enabling him to obtain some measure of sentence joining. Bisiada (2013:127) concluded that 'the tendency among translators towards sententialisation [i.e., sentence-splitting] is much stronger than the opposite tendency towards combining them'. However, since both Ramm (2004) and Bisiada (2013) used lexical queries as a starting point, they were not able to observe changes that were not motivated by the selected lexis under analysis. These studies are thus only able to offer a partial view of sentence-boundary shifts.

There are also a few earlier, non-corpus-based studies about sentence boundaries in translation. For example, in her qualitative analysis of 50 Dutch novels from the late fifties to 1980 translated into English, Vanderauwera (1985) observed that there seemed to be a tendency for translators to break up long sentences. Fabricius-Hansen (1999) analysed cases of sentence splitting in the English and Norwegian translations of Lorenz's *Das Sogennante Böse*, and noticed that the German source text preferred a more complex syntactic structure than both translation languages.

1.2 Factors affecting sentence boundaries

Fabricius-Hansen (1999) attributed sentence splitting to differences in structural norms between languages, and the generally higher informational density of German. Decades earlier, Kaplan (1972) had already discussed the contrastive rhetoric of different languages, where some languages appear to exhibit greater tolerance for long and complex sentences. This could explain the need for splitting or joining sentences in translation.

Another possible factor behind changes in sentence boundaries could be the cognitive demands of the translation process itself. Translators need to unpack complex

[1]However, Serbina (2014) was able to further inspect certain instances of sentence splitting by examining cases where clause alignment and sentence alignment did not match.

[2] Note that the German relative adverb *was* ['which, what'], arguably one of the main subordination markers in German, was excluded from the analysis "because it was not possible to filter out the instances of *was* not corresponding to a relative marker" (Ramm 2008:143).

grammatical structures in a source text in order to make sense of them, and may not re-pack them when rendering the translation. Or they may forge mental connections between source-text sentences, and then transpose such connections to the translation. However, despite significant advances in translation process research (e.g. Lörscher 2005, O'Brien 2013), there does not seem to be any study that specifically informs why translators split or join sentences.

Sentence boundaries can also be examined from the perspective of translation theories. Toury (1987:95) put forward the idea that translators are guided by similar norms, 'irrespective of the translator's identity, language, genre, period and the like'. This possibility began to be explored further after Baker (1993, 1996) proposed using corpora to investigate translation trends which could be universal. Three of such so-called universals can have direct, albeit conflicting, implications for the ways translators deal with sentence boundaries. The first one is Blum-Kulka's (1986) explicitation hypothesis, which Baker (1996:180) glossed as 'an overall tendency to spell things out rather than leave them implicit'. This can include not only the tendency for translators to add extra adverbs, which do not affect sentence structure, but also extra conjunctions that make the relationship between clauses more explicit and foster joining independent clauses together. The second alleged universal is simplification, or 'the tendency to simplify the language used in translation' (Baker 1996:181). One of the various mechanisms of simplification anticipated by Baker is breaking up long and complex source-text sentences into shorter translation sentences, as reported in Vanderauwera's (1985) previously mentioned study. Note, however, that producing shorter sentences for the sake of simplification contradicts the tendency for explicitation via the addition of coordinating and subordinating conjunctions to link independent sentences together, whose effect would be the opposite. The third universal discussed by Baker (1996) that could have consequences for changes in sentence boundaries is normalization, whereby it is hypothesized that translators tend to use language more conservatively. If this holds true, then it could work either way with regard to sentence boundaries – translators could split or join syntactically unusual source-text sentences into whatever resulted in more conventional units in the translation.

Baker's (1996) concept of normalization seems to draw on what Toury (1995:267) had referred to one year earlier as the law of 'growing standardization', which states that distinctive source-text features tend to be replaced by more standard or conventional options available in the target language. Malmkjaer (1997), for example, observed that English translators tended to normalize the unusual punctuation employed in stories by Danish author Hans Christian Anderson.

Another aspect of translation that needs to be taken into account when examining what translators do with sentences is Toury's (1995:275) law of interference, whereby 'phenomena pertaining to the make-up of the source text tend to be transferred to the target text'. When source and target language conventions coincide, the interference of source-text features will not be noticed in translations. However, translators can also transfer characteristics of source texts that clash with target language conventions. At first sight, this negative type of interference is incompatible with the law of growing standardization: either translators let negative interference follow its course or they try to curb it by standardizing the translation according to target-language norms. However, there is an extra layer of complexity involved here. For Toury (1995:278), interference is more likely than standardization 'when translation is carried out from a "major" or highly

prestigious language/culture, especially if the target language/culture is 'minor' or "weak"'. For example, as discussed in Frankenberg-Garcia (2016:21), in the particular case of English and Portuguese, Portuguese could be said to be less hegemonic and more prone to interference from English because '[w]hile most Portuguese speakers are exposed to the English language and culture in their everyday lives, the number speakers of English in the world who are familiar with Portuguese is comparatively very scant.' Professional translators aware of this imbalance are bound to take it into account when assessing the extent to which source language/culture interference will be tolerated by target language/culture readers, and how much standardization is required of them.

Apart from the status of the source language/culture in relation to the target language/culture, text genre may also affect translators' choices with regard to promoting standardization or permitting interference: a prestigious genre like literature, which also has a markedly strong expressive function, is less likely to suffer major changes in translation than less celebrated, predominantly informative genres like instruction manuals, where translators may alter the text more freely to facilitate target-language comprehension. Thus, in addition to the source language/culture dominance factor discussed by Toury, translators may be more open to transferring foreign traits of prestigious genres than of genres of lower status. Indeed, in Ramm's (2004) previously mentioned study, there was less sentence splitting in the Oslo Multilingual Corpus (containing a substantial component of literature) than in a corpus of popular science.

It is also relevant that different genres are characterized by specific lexical and grammatical features. Recipes in English, for example, tend to consist of lists of ingredients followed by simple imperative sentences instructing people what to do with them. Thus unlike other genres, like literary or academic, in recipes there are comparatively fewer compound and complex sentences with the potential to be split.

Finally, it is important to note that certain text types are nowadays frequently translated in a computer-assisted translation (CAT) environment, which probably affects the way translators deal with sentence boundaries in no small way. This is because the software pre-segments source texts at the level of the sentence in preparation for translation, and translators are prompted to translate segment by segment, in a platform that discourages sentence splitting, and makes it particularly challenging to join sentences together.

None of what translators actually do with source-text sentences can be regarded as a so-called universal on the basis of the evidence available so far. In Vanderauwera's (1985) qualitative study comparing Dutch novels and their English translations, we only know what happened in a single translation direction, so it is not possible to tell whether the changes observed had more to do with linguistic differences between Dutch and English, or were due to translation itself. Likewise, Fabricius-Hansen's (1999) and Bisiada's (2013) previously mentioned studies were also unidirectional. Although Bisiada looked at English translated into German and Fabricius-Hansen looked at German translated into English, the former corpus-based study is not directly comparable with the mostly qualitative observations reported in the latter. Moreover, the genre of the texts under analysis in the two studies differed, so it is not possible to factor in their effect: sentence splitting in the translation of a syntactically complex argumentative text from German into English cannot be compared with sentence splitting in English to German business texts. And in the only quantitative study that explored sentence splitting in more than one translation direction that seems to be available to date, Serbina (2014) only reported on

the total number of sentences in English and German source-texts and translations, which does not take into account the effect of sentence joining.

In order to explore translation universals, Baker (1996) proposed comparing non-translated texts with similar texts translated from a variety of source languages. This would act as a control for the interference of language-pair-specific translation shifts. However, although this method may uncover certain traits that differentiate translated from non-translated language, looking at translations in the absence of source texts will not allow one to observe sentence splitting and sentence joining.

Another way of teasing out potential translation universals is to carry out bidirectional analyses, which, as noted by Frankenberg-Garcia (2009), can also serve to filter out language-dependent variables. Thus if translating from language A to language B results in splitting sentences, and translating from language B to language A results in joining comparable sentences together, then it is likely that language A tolerates longer sentences better than language B. However, if translators split certain sentences apart and join others together regardless of language direction, then the phenomenon is likely to be intrinsic to translation.

Yet there does not seem to be any systematic study of how translators deal with sentence splitting *and* sentence joining in different language directions, or indeed of the extent to which original source-text sentence boundaries are actually preserved in translation. Using a 3-million word bidirectional parallel corpus of Portuguese and English fiction, the present study thus sets out to seek answers to the following questions:

1) To what extent do Portuguese and English translators preserve sentence boundaries? Given the conditions surrounding Toury's (1995) law of interference, do translators working into English, which is regarded as a more dominant language/culture than Portuguese, alter sentence boundaries more freely than their counterparts working into Portuguese?

2) Does the more well-known phenomenon of sentence splitting prevail over sentence joining in translation? Does the performance of translators working into English differ from the performance of translators working into Portuguese with regard to the extent to which they split and join sentences?

(3) How exactly do translators split and join sentences? Does the performance of translators working into English differ from the performance of translators working into Portuguese with regard to the ways they go about splitting and joining sentences?

Answers to the above questions will hopefully shed some light on whether there are any trends regarding sentence-boundary shifts in translation regardless of translation language direction. If what applies to the English-Portuguese direction also applies to the opposite, Portuguese-English direction, this could be an indication of a translation universal.

2. Defining sentence boundaries

For an analysis of sentence-boundary shifts to be systematic, it is essential to define what is meant by sentence. For Quirk et al. (1985), a sentence is a grammatical unit that is

composed of one or more clauses. This can include elliptical material, such that [1] and [2] are also regarded as sentences.

[1] Not today.
[2] Silêncio! [Silence!]

In order to analyse thousands of sentences, however, it is necessary to translate this into an operational, machine-readable concept. The working definition of sentence for the purpose of the present study is the one employed in the segmentation of source texts in the COMPARA corpus (Frankenberg-Garcia and Santos 2003), according to which a sentence begins with a capitalized word and ends with a hard punctuation mark (full-stop, ellipsis, exclamation mark and question mark), and is followed by another capitalized word or no text at all. For example, extract [3] from the corpus contains three sentences (marked <s>):

[3]<s>The Maire, who had been so anxious to deliver his story, dithered over this direct question. <s>`It's necessary to understand, Madame... <s>The Sauvy brothers saw this with their own eyes, through the window... and we heard later that this also used to happen at the interrogation centres in Lyon and Paris.

Note that the application of this sentence-separation criterion means there is a new sentence after the first ellipsis, but not after the second one, regardless of the author's or translator's intention, or indeed of the reader's or researcher's interpretation. Thus, in the present study, even a minor change like capitalizing *and* after the second ellipsis in [3] would count as a sentence split.[3]

The colon is only considered a sentence separator if it marks the end of a paragraph, irrespective of whether the word after the colon begins with a capital letter. Thus both [4] and [5] are regarded as one sentence.

[4]<s>That's what Maureen used to do: look adoringly at me.
[5]<s>Edward: Hang on!

Segmentation does not occur when there are full-stops in abbreviations such as *Mrs.* or points separating decimal numbers, thanks to segmentation rules that recognize such exceptions. Hard punctuation marks within direct speech followed by strings that include reporting verbs such as *say*, *tell* and *suggest* are also not considered sentence separators. Thus excerpt [6] is classified as one sentence, even though the word after the question mark begins with a capital letter.[4]

[6]<s>'You OK?' Robin's daughter said, standing close to him, but not touching.

3. The corpus

The data used in the present study comes from the same COMPARA corpus – a three-million-word bidirectional parallel corpus of English and Portuguese fiction

[3] Although this shift may not have comparable consequences in terms of information packaging as upgrading a subordinate clause to an independent sentence, it is nevertheless important to record it, as it can have implications from the viewpoint of Toury's (1995) previously mentioned law of growing standardization.
[4] Further examples of the application of these segmentation criteria are available at http://dinis2.linguateca.pt/COMPARA/construcao_compara.php#sepfrase [27/02/2016].

(Frankenberg-Garcia and Santos 2003)[5]. It is a relatively balanced corpus, with 1.44 million words in Portuguese and 1.54 million words in English. The texts in the corpus were published between 1837 and 2002, and although a few source texts are in the public domain, all but one translation are protected by copyright law. Because of copyright restrictions, the bitexts in COMPARA are 10% to 30% extracts of unequal length taken randomly from the beginning, middle or end of books.

Only published source texts and translations, and only direct Portuguese-English and English-Portuguese translations were admitted in the corpus. The 75 bitexts in COMPARA represent the work of 36 original fiction writers from Angola, Brazil, Mozambique, Portugal, the United States, Britain, Ireland and South Africa, and 48 professional translators from Brazil, Portugal, Britain and the United States. In the present study, no distinction will be made with regard to different varieties of English and Portuguese. Although there are well-known lexical, spelling and sub-sentential grammatical differences (such as the use of prepositions) between different varieties of Portuguese and different varieties of English, there does not seem to be any evidence of discourse differences affecting sentence structure reported in the literature to justify such a distinction. The term Portuguese translators will be used henceforth to refer to translators working into both Brazilian and European Portuguese, and the term English translators will refer to translators working into both British and American English.

The translations in COMPARA are unlikely to have been produced using computer-assisted-translation (CAT) tools (and hence are unlikely to have been affected by them) because even today such tools are not normally used by literary translators (Granell 2015). Moreover, the most recent translation in COMPARA was published in 2002, which is before the use of CAT tools became widespread.[6]

Figure 2. Source-text-based alignment in COMPARA

SOURCE TEXT	TRANSLATION	Alignment annotation
Sentence A	Sentence A_1	
Sentence B	Sentence B_1	Split
	Sentence B_2	
Sentence C	Sentence $CD_{1st\ half}$	Joined
Sentence D	Sentence $CD_{2nd\ half}$	Joined
Sentence E	∅	Deleted
Sentence F	Sentence F_1	
	Sentence F_{2add}	Added

As discussed in section 1.1, one of the problems of using parallel corpora to analyse sentence boundaries is the way texts are normally aligned. In COMPARA, alignment was initially carried out on the usual basis, using the EasyAlign 1.0 tool incorporated in the IMS workbench (Christ et al. 1999). The automatic alignment output was then manually post-edited to ensure the alignment unit was always one single source-text sentence and the corresponding text in the translation, whether it was one, more than one, or even only part of a sentence. Source-text sentences that were left out of the translation were aligned with blank units and sentences that were added to the translation were fitted into the immediately preceding alignment unit (Frankenberg-Garcia et al. 2006). Alignment

[5] Open-access at www.linguateca.pt/COMPARA [25/04/2016]
[6] See Chan (2014) for a historical overview of the commercialization of CAT tools.

annotation was then inserted to enable one to run queries that automatically locate source-text sentences that were split, joined, deleted, and added to the translation (see figure 2).

The present study did not use the entire COMPARA corpus. Two bitexts were purposefully excluded from the analysis because Portuguese Nobel prize winner José Saramago practically does not use full-stops, and, as detailed in Frankenberg-Garcia (2006), the segmentation of his texts follows a different, customized set of rules. Another five bitexts also had to be excluded because their alignment had not been post-edited to conform to the alignment criteria established. Figure 3 summarizes the sub-corpus of COMPARA upon which the present analysis is based. Note that although there are fewer English-Portuguese bitexts, their combined length is actually greater than that of the Portuguese-English bitexts.

Figure 3. Sub-corpus used in the analysis[7]

	N	Source-text words	Source-text sentences
PT-EN bitexts	39	575,604	42,927
EN-PT bitexts	29	757,958	48,591
TOTAL	68	1,333,562	91,518

4. Data analysis

This section begins by examining the overall proportion of sentence boundaries preserved in translation. Next, sentence joining and sentence splitting will be compared. This will be followed by a closer inspection of how translators split and joined sentences.

4.1 Sentence preservation

Of the 91,518 source-text sentences in figure 3, a total of 83,118 were preserved in translation, 45,291 by the Portuguese translators and 37,827 by the English translators. As the bitexts in the corpus are of unequal length, instead of using these raw values, where longer source texts would have been assigned more weight, it made sense to analyse the percentage of sentences preserved per bitext.

Given the prestige normally associated with literary fiction, it was predicted that sentence preservation would prevail in both translation directions. However, since English is taken to be a more hegemonic language/culture than Portuguese in the sense that Portuguese readers are more familiar with English than English readers are with Portuguese, the analysis also assessed whether English translators altered sentence boundaries with more freedom than their Portuguese counterparts.

The results obtained for sentence preservation are summarized in table 1. As the data was not normally distributed, central tendencies are expressed in terms of medians. The results indicate that both Portuguese and English literary translators have a very strong tendency to preserve source-text sentence boundaries: in both directions the data point towards an over 90% rate of sentence preservation.

Table 1. Sentence boundaries preserved in translation by PT and EN translators

[7] Further details about the exact text size and the individual authors and translators represented in COMPARA are available at http://www.linguateca.pt/COMPARA/ [29/02/2016]

% of source-text sentences preserved	Portuguese translators (29 bitexts)	English translators (39 bitexts)
Median	94.6	90.1
High	98.2	98.9
Low	70.9	52.2

A one-tailed Mann-Whitney (non-parametric) test was then applied to compare each bitext in the two independent datasets in order to find out whether the English translators preserved fewer sentence boundaries. The Z-score was 2.1081, which is significant at $p \leq 0.05$. This leads one to reject the null hypothesis that the English translators preserve as many sentence boundaries as their Portuguese counterparts.

4.2 Sentence splitting and joining

This section examines whether the more frequently researched phenomenon of sentence splitting is more common than sentence joining, and whether the performance of translators working into English differs from the performance of translators working into Portuguese in this respect.

Overall, 3600 sentences were split and 4282 sentences were joined in the corpus.[8] Thus at first glance sentence joining appears to be actually more common than sentence splitting. However, these figures are too general, for they do not take into account differences between the two translation directions or indeed between each bitext. Table 2 summarizes the more detailed results regarding sentence splitting and joining.

Table 2. Sentence splitting and joining

	Portuguese translators (N=29)		English translators (N=39)	
	% joined/bitext	% split/bitext	% joined/bitext	% split/bitext
Median	2.71	1.81	2.77	3.23
High	10.70	15.70	23.90	26.40
Low	0.39	0.23	0.00	0.24

In the previous section, we saw that on average over 90% of source-text sentences in both translation directions had been preserved, thus the proportion of sentences left to be split and joined was comparatively small. The median values in table 2 indicate that, in both translation directions, the differences between sentence joining and splitting did not seem to be very substantial, although apparently the Portuguese translators joined more than split while the English translators split more than joined.

In order to check whether the slight differences between splitting and joining were significant, a two-tailed Wilcoxon matched-pairs signed-ranks test for comparing related samples was applied to the Portuguese translators' data, and then to the English translators' data.[9] The Z-score for Portuguese translators was 1.4801, and the one for the English translators was 0.8484, neither of which is significant at $p \leq 0.05$. Thus it cannot be concluded that the performance of literary translators working into English differed from

[8] Sentence deletion (467 sentences) and addition (51 sentences) were comparatively marginal phenomena.
[9] The Wilcoxon matched-pairs test focuses on the differences between paired data – in this case, the differences between sentence splitting and sentence joining for each separate bitext – but, unlike the matched t-test, it does not assume normal distribution.

the performance of literary translators working into Portuguese with regard to the extent to which they split and joined sentences.

4.3 How translators split and joined sentences

The quantitative results described in the previous section did not disclose any significant differences with regard to sentence splitting and joining. However, the Portuguese and English translators could still differ with regard to how they actually went about splitting and joining sentences. This section therefore zooms in on a sample of over 1000 parallel text segments involving sentence splitting and joining.

As already mentioned, text size in COMPARA varied considerably. In order to obtain a sample that was as balanced as possible for this part of the analysis, for each bitext (39 PT-EN and 29 EN-PT), a random set of parallel concordances containing up to 10 source-text sentences that were split and another one with up to 20 source-text sentences that were joined were selected for a more fine-grained analysis. If a source-text sentence was split into more than two translation sentences, only the first two parts of the split were considered. Similarly, if more than two source-text sentences were joined in the translation, only the first two that were joined were taken into account. Following the same principle, for complex shifts involving splitting and joining, only the first change (whether a split or a join) was inspected.

The above procedure generated a sample of 1133 parallel text segments: 603 containing source-text sentences that were split in translation (266 by Portuguese translators and 337 by English translators) and 530 segments with source-text sentences that were joined in translation (248 by Portuguese translators and 282 by English translators).

4.3.1 Sentence splitting

After inspecting the concordances involving sentence splitting, six categories were developed to account for different types of split. These are described below and apply to both language directions, as can be seen from the corpus examples supplied[10].

A. *Hard Punctuation Inserted*. This category accounts for sentence splitting by inserting a full-stop, exclamation mark, question mark or ellipsis where there was no previous punctuation mark, with the next word being capitalized:

ST**:**<s>Really I just hope I won't be intruding.
TT:<s>Palavra. <s>Só espero não estar a intrometer-me.
BT:<s**>**I swear. <s>I just hope not to be intruding.

ST:<s>Sim, isso já disse mas ainda não disse que para o fim da vida só não fez filhos na papisa de Roma porque não a apanhou no consultório.
LT:<s>Yes, I've said that but I still haven't said that towards the end of his life he only did not make children with the papess from Rome because he did not catch her at the surgery.
TT:<s>Right, I've said that. <s>But I haven't said that toward the end of his life if he didn't get a Roman papess pregnant it was only because he didn't find one in his office.

B. *Soft to Hard Punctuation***.** This category describes sentence splitting by changing commas, semi-colons, colons or dashes into full-stops, exclamation marks, question marks or ellipsis, with the next word being capitalized:

[10] ST = source text; TT = target text; BT = back translation; LT = literal translation

ST**:**<s>Oh yes, Mr Wilcox, these are ever so hard.
TT:<s>Oh, sim, Sr. Wilcox. <s>Estas são tão duras.
BT:<s**>**Oh, yes, Mr Wilcox. <s>These are so hard.

ST:<s>Voltou a comer, a mãe procurava agora observar melhor o filho.
LT:<s>He started to eat again, the mother tried now to better observe the son.
TT:<s>He started eating again. <s>His mother was watching him more carefully now.

C. *Capitalization***.** This category involves sentence splitting by changing words beginning with small letters after full-stops, exclamation marks, question marks or ellipsis into words beginning with capital letters:[11]

ST**:**<s>A voice! a voice!
TT:<s>A voz! <s>A voz!
BT:<s**>**The voice! <s>The voice!

ST:<s>Vamos... fala!
LT:<s>Come on… speak!
TT:<s>Come now... <s>Speak up!

D. *Minus Coordination***.** This category captures sentence splitting by removing a coordinating conjunction (underlined in the examples), such that what was previously a coordinate clause becomes an independent sentence: [12]

ST**:**<s>An hour later he was at the Opera, and Lord Henry was leaning over his chair.
TT:<s>Uma hora depois achava-se na Ópera. <s>Sobre a sua cadeira apoiava-se Lord Henry.
BT:<s**>**One hour later he found himself at the Opera. <s>On his chair was leaning Lord Henry.

ST:<s>Você nunca me perguntou como eu pressenti a chegada de Legião, e agora eu vou lhe dizer: pela audição.
LT:<s>You never asked me how I sensed the arrival of the Legion, and now I will tell you: by hearing.
TT:<s>You have never asked me how I knew that Legion was about to arrive. <s>Now I will tell you how: by listening.

E. *Minus Subordination.* This category accounts for sentence splitting by upgrading a subordinate clause into an independent sentence via the deletion of a subordinating conjunction (underlined in the examples) or the replacement of a gerund with a subject-verb construction:

ST**:**<s>They need sleep when they're growing, Vic.
TT:<s>Eles estão a crescer. <s>Precisam de dormir, Vic.
BT:<s**>**They are growing. <s>They need to sleep, Vic.

ST:<s>Ou é melhor estar com os livros, que contam estórias incríveis sempre nas horas que a gente quer ouvir.
LT:<s>Or it is better to be with the books, which tell incredible stories always at the hours that people want to hear.

[11] As discussed in section 2, while this change is so minor that it will not affect information density, it could have implications in terms of Toury's (1995) law of growing standardization.
[12] Note that although the punctuation changes that accompany this transformation will overlap with categories A or B, there is no ambiguity in the classification because those categories consist of punctuation changes alone, without involving coordinating conjunctions. The same principle applies to the Minus Subordination and Major Reformulation categories described further on, and to the reverse sentence-joining categories described in 4.3.2.

TT:<s>And better still to be alone with one's books. <s>They tell their incredible stories at the time when you want to hear them.

F. *Major Reformulation*. This category accounts for sentence splitting by reformulating the source-text sentence completely, such that the translation is rendered as two separate sentences:

ST**:**<s>I'm having a minor operation.
TT:<s>Vou ser operado. <s>Uma operação simples.
BT:<s>I will be operated on. <s>A simple operation.

ST:<s>Que desejava? disse enfim o dono da casa.
LT:<s>What did he want? said finally the owner of the house.
TT:<s>Finally the master of the house spoke. <s>`Is there something I can do for you?´ he asked.

The above classification system was straightforward to apply. A random sample of 50 parallel concordances involving sentence splitting in both language directions (25 EN-PT and 25 PT-EN) was given to a second coder, with 96% inter-rater agreement.[13]

The distribution of different types of sentence splitting is summarized in tables 3 and 4. As the data was not consistently normally distributed, central tendencies are expressed as medians. The results indicate that the performance of the English and the Portuguese translators was remarkably similar in this respect. In both translation directions, most cases of sentence splitting were due to Soft to Hard changes. All other types of sentence splitting looked negligible, with medians equal to zero.

Table 3. Types of sentence splitting by Portuguese translators

N=29	Hard Inserted	Soft to Hard	Capitali-zation	Minus Coord	Minus Subord	Major Reform	Total
Frequency	5	194	21	21	14	11	266
Median	0	7	0	0	0	0	
High	2	10	6	3	3	4	
Low	0	0	0	0	0	0	

Table 4. Types of sentence splitting by English translators

N=39	Hard Inserted	Soft to Hard	Capitali-zation	Minus coord	Minus subord	Major reform	Total
Frequency	2	251	24	24	29	7	337
Median	0	7	0	0	0	0	
High	1	10	5	7	3	1	
Low	0	0	0	0	0	0	

4.3.2 Sentence joining

The categories used to describe the sample of 530 parallel text segments used to inspect sentence joining were the exact reverse of those used in the description of sentence splitting. These are explained below, with bidirectional examples from the corpus.

[13] The only discrepancy was the transformation of a gerund into a subject-verb construction being interpreted as Major Reformulation, when it should have been classified as Minus Subordination. The second coder was briefed on the difference between the two and the issue was resolved.

G. *Hard Punctuation Deleted.* This category captures sentence joining by deleting a full-stop, exclamation mark, question mark or ellipsis, with the next word after the punctuation mark removed being rewritten in lower-case:

ST**:**<s>So there was no trouble? <s> On Monday?
TT:<s>Então não houve problema na segunda-feira?
BT:<s**>**So there was no problem on Monday?

ST:<s>A velha assustou-se: qual o fogo que o homem vira? <s>Se nenhum não haviam acendido?
LT:<s>The old woman was frightened: what fire had the man seen? <s>If they had lit none?
TT:<s>The old woman got alarmed: what was this fire the man had seen if they hadn't even lit one?

H. *Hard to Soft Punctuation.* This category describes sentence joining by changing full-stops, exclamation marks, question marks or ellipsis into commas, semi-colons, colons or dashes, and rewriting the word that follows in lower-case:

ST**:**<s>Excuse me, sir. <s> I must attend to him straight away
TT:<s>Com licença, sir, tenho de ir atendê-lo imediatamente.
BT:<s**>**Excuse me, sir, I have attend to him immediately

ST:<s>Engraçado. <s> Eu não me lembrava.
LT:<s>Funny. <s>I didn’t remember.
TT:<s> Funny, I didn't remember.

I. *Minus Capitalization.* This category accounts for sentence joining by changing words beginning with capital letters after full-stops, exclamation marks, question marks or ellipsis into words in lower-case:

ST**:**<s>And saying everything they'd got out of her, dirtying herself... <s>All in front of him.
TT:<s>E a dizer tudo quanto lhe arrancavam, a sujar-se... tudo na frente dele.
BT:<s**>**And saying everything they extracted from her, getting dirty... all in front of him.

ST:<s>Tudo a mesma gente: púnicos, mouros... <s>Farinha do mesmo saco.
LT:<s>All the same people: Carthaginians, Moors… <s>Flour from the same sack.
TT:<s>They're all the same: Carthaginians, Moors... flour out of the same sack.

J. *Plus Coordination.* This category encompasses sentence joining by inserting a coordinating conjunction (underlined in the examples), such that what was previously an independent sentence is converted into a coordinate clause:

ST**:**<s>She stopped. <s>She straightened up.
TT:<s>Deteve-se e endireitou-se.
BT:<s**>**She stopped and straightened up.

ST:<s>Ela buscou apoio no marido. <s>O marido parecia indiferente à inquietação da mulher.
LT:<s>She seeked support from the husband. The husband seemed indifferent to the agitation of the wife.
TT:<s>She looked to her husband for support, but he seemed indifferent to his wife's disquiet.

K. *Plus Subordination.* This category describes joining sentences by subordination by adding a subordinating conjunction (underlined in the examples), or converting subject-verb constructions into gerunds, such that what was previously an independent sentence becomes nested inside another sentence:

ST:<s>I collected it myself at a very great personal risk. <s>I am afraid they will try to claim it as theirs though.
TT:<s>Eu próprio fui buscá-lo com enormes riscos pessoais, embora receie que eles pretendam reclamá-lo como seu.
BT:<s>I myself went to fetch it with enormous personal risk, although I fear they intend to claim it as theirs.

ST:<s>Choveram murros, pontapés, bofetadas. <s>A assistência, em volta, aplaudia.
LT:<s>Blows, kicks, punches rained. <s>Those watching around clapped.
TT:<s>Blows, kicks, punches rained down on him while the bystanders cheered.

L. *Major Reformulation.* This category captures sentence joining by reformulating two source-text sentences completely, such that the translation is rendered as a single sentence:

ST:<s>«She would never know,» said Louise. <s>«How could it hurt her?»
TT:<s>Ela nunca o saberia; por isso, não estarias a fazê-la sofrer.
BT:<s>She would never know it; therefore, you wouldn't be making her suffer.

ST:<s>Havia perto uma banca de jornais. <s>Comprei uma revista de cinema.
LT:<s>There was a news-stand nearby. <s>I bought a film magazine.
TT:<s>On the way in I bought a film magazine at the news-stand.

As with the sentence-splitting categories, the sentence-joining ones were unambiguous to apply in practice. Inter-rater agreement reached 100% for a random sample of 25 EN-PT and 25 PT-EN segments involving sentence joining. The distribution of different types of sentence joining is summarized in tables 5 and 6.

Table 5. Types of sentence joining by Portuguese translators

N=29	Hard Deleted	Hard to Soft	Minus Capitaliz.	Plus Coord.	Plus Subord.	Major Reform.	Total
Frequency	12	116	9	33	53	25	248
Median	0	3	0	1	1	0	
High	3	9	3	5	6	6	
Low	0	0	0	0	0	0	

Table 6. Types of sentence joining by English translators

N=39	Hard Deleted	Hard to Soft	Minus Capitaliz.	Plus Coord.	Plus Subord.	Major Reform.	Total
Frequency	8	177	13	43	33	8	282
Median	0	4	0	1	1	0	
High	2	9	3	5	4	3	
Low	0	0	0	0	0	0	

It can be observed from the medians in tables 5 and 6 that the overall performance of the English and the Portuguese translators with regard to sentence joining was also quite similar. In both translation directions, the most frequent type of joining occurred due to Hard to Soft Punctuation changes. The median of four obtained for the English translators was however slightly higher than the three obtained for the Portuguese translators. To find out whether this difference could be significant, the non-parametric Mann-Whitney two-tailed test for independent samples was applied to further compare the 116 cases of Hard to Soft Punctuation in the 29 EN-PT bitexts and the 177 cases in the 39 PT-EN

bitexts. The Z-score of 1.2138 did not reach significance at $p \leq 0.05$, thus it cannot be asserted that the English and Portuguese translators behaved differently in this respect.

The medians in tables 5 and 6 also indicate that in both translation directions there seemed to be a slight tendency for sentence joining by subordination and coordination. Although in terms of totals there was more subordination than coordination in the Portuguese translator data and more coordination than subordination in the English translator data, from the medians obtained one cannot say the two behaved differently in this respect. This question will however be further investigated in 4.3.3. The remaining types of sentence joining – Hard Punctuation Deleted, Minus Capitalization and Major Reformulation – seemed negligible, with medians equal to zero.

4.3.3 Contrasting splitting and joining strategies

The analyses in the two previous sections focused on detailing the ways in which Portuguese and English translators went about first splitting and then joining sentences. To obtain a more complete picture of their performance, this final section zooms in on contrasting sentence-splitting and sentence-joining strategies. In particular, the aim of this section is to determine whether there were any trends regarding:

i. Inserting versus deleting hard punctuation marks;
ii. Changing soft punctuation marks into hard ones versus changing hard punctuation marks into soft ones;
iii. Introducing capitalization versus removing it;
iv. Transforming coordinate clauses into independent sentences versus transforming independent sentences into coordinate clauses;
v. Transforming subordinate clauses into independent sentences versus transforming independent sentences into subordinate clauses;
vi. Making major reformulations to split sentences versus making major reformulations to join sentences.

As seen in 4.3.1 and 4.3.2, contrasts (i), (iii) and (vi) above were very marginal. In both translation directions, the medians for the number of sentences split and joined as a result of inserting or deleting hard punctuation marks, capitalizing and not capitalizing words, and major reformulations were equal to zero. It was therefore felt that there was not enough data regarding these phenomena to allow further investigation. Instead, focus will be given to the more substantial soft/hard, coordination and subordination contrasts summarized in table 7.

Table 7. Contrasting splitting and joining strategies (*significant at $p \leq 0.05$)

	PT translators	EN translators
Soft to Hard : Hard to Soft	194:116*	251:177*
Minus Coordination : Plus Coordination	21:33	24:43*
Minus Subordination : Plus Subordination	14:53*	29:33

As can be seen from the ratios in table 7, in both language directions, it was more common for the translators to change punctuation marks from soft to hard to split sentences than to change punctuation from hard to soft to join sentences. Also in both language directions, it was more common for the translators to increase rather than decrease coordination and subordination, resulting in more sentence joining.

In order to check whether these results reflected actual translation trends rather than the idiosyncratic behaviour of just a handful of translators, one-tailed Wilcoxon matched-pairs signed-ranks (non-parametric) tests for comparing related samples were applied to each of these contrastive dimensions.

For the differences between Soft to Hard and Hard to Soft, the Z-scores were 2.8464 and 2.5142 respectively for the Portuguese and English translator data, both of which are significant at $p \leq 0.05$. This suggests that, in both translation directions, the tendency to split sentences by changing soft punctuation marks into hard ones is greater than the tendency to join sentences by changing hard punctuation marks into soft ones.

For the differences between Minus Coordination and Plus Coordination, the Z-score obtained for the Portuguese translator data was 1.1991, which is not significant at $p \leq 0.05$. For the English translators, however, Z was 1.7407, which is significant at $p \leq 0.05$. This suggests that Portuguese and English translators may behave differently in this respect: while the differences between increasing and decreasing clause coordination were not significant among the Portuguese translators, the English translators showed a greater tendency to join independent sentences by coordination than to split coordinate clauses into separate sentences.

Finally, for the differences in subordination, the Z-scores were 2.9544 for the Portuguese translator data and 0.2571 for the English translator data. Only the Portuguese Z-score was significant at $p \leq 0.05$. Thus the Portuguese and English translators also appear to behave differently with regard to subordination. While there was a tendency for the Portuguese translators to increase more than reduce clause subordination, there was no marked difference between linking independent sentences by subordination and separating subordinate clauses among English translators.

5. Discussion

The first part of the analysis (section 4.1) focused on the proportion of source-text sentence boundaries preserved in translation. Previous studies like Serbina (2014) compared the total number of sentences in source texts and translations, but did not provide a systematic analysis of the extent to which source-text sentence boundaries were actually preserved in translation. In the present study, as predicted in the introduction, sentence preservation was very high among translators working with literary texts, with a median of over 90% of the sentence boundaries of source texts remaining intact in both translation directions.

However, this should not be interpreted as evidence that the potential effect of the source-text-driven pre-segmentation of texts imposed by today's widely used CAT software is negligible. As discussed in the introduction, the prestige associated with literary texts is bound to make them less prone to changes, including changes in sentence boundaries, than other genres. In the future, apart from looking at more language pairs, it would therefore be especially important to find out how the rates of sentence preservation for translated literature compare with those of other, less prestigious genres, where translators may feel more at liberty to alter the source-text author's style. Such analyses may not be simple to carry out outside a controlled experimental setting, however. It would be

essential to ensure that the translations were not produced under the intervening influence of CAT in the first place.

The results presented in section 4.1 also showed that in addition to the high rates of sentence preservation for the literary genre, the English translators tended to adhere a little less frequently to source-text sentence boundaries than their Portuguese counterparts. These findings seem to tie in with the conditions surrounding Toury's (1995) law of interference discussed in the introduction. In other words, the data suggests that translators working from a less hegemonic language/culture to a more dominant one (in this case from Portuguese to English) seem less susceptible to transferring characteristics of the source text (in this case sentence boundaries) to the target text than translators working in the reverse direction. As Portuguese readers are more familiar with English than English readers are with Portuguese, English sentence structures seem to have been more easily transferred to Portuguese than Portuguese sentence structures to English.

The next part of this study focused on sentence splitting and sentence joining from a quantitative, bidirectional, perspective. As observed in the introduction, some of the previous studies addressing sentence boundaries did not provide any quantitative data, others looked at sentence splitting without considering sentence joining, and others offered just a partial view of changes motivated by a restricted set of conjunctions. Moreover, most previous studies only discussed sentence-boundary changes from the perspective of a single translation direction. In contrast, the findings provided in 4.2 provided quantitative evidence that, for literary translation, the more widely researched phenomenon of sentence splitting was not more frequent than sentence joining. Furthermore, the performance of the translators working into English did not differ significantly from the performance of the translators working into Portuguese in this respect. Most importantly, the present findings encourage one to question the assumption that there is a tendency for simplification in translation due to the breaking up of source-text sentences, as suggested by unidirectional studies like Vanderauwera (1985), Fabricius-Hansen (1999), Ramm (2004) and Bisiada (2013). If, irrespective of translation direction, splitting sentences is not more common than joining sentences together, their combined effect cannot be one of overall simplification.

The final part of the study examined in more detail a balanced sample of over 1000 parallel text segments involving sentence splitting and joining. Overall, the ways Portuguese and English translators went about splitting and joining sentences was remarkably similar. The findings presented in 4.3.1 and 4.3.2 revealed that, in both translation directions, the majority of changes were due to shifts from soft punctuation marks to hard ones (resulting in sentence splitting) and shifts from hard punctuation marks to soft ones (resulting in sentence joining). These may be regarded as minor changes if compared with changes involving subordination, for example, which affect information density in a much bigger way. However, it matters that changes in punctuation are very frequent, and it is important to understand what actually lies behind them. Although the changes from soft to hard punctuation and from hard to soft punctuation seemed at first glance contradictory, as illustrated in the examples given for these changes in 4.3.1 (category B) and in 4.3.2 (category G), both seem to represent an attempt to normalize the resulting syntax. Indeed, while reading through the data for the purpose of classifying the different types of sentence splitting and joining, it was possible to observe that many changes from soft to hard punctuation marks involved upgrading asyndetic clauses to

independent sentences like in example [7] or replacing a less conventional colon with a full-stop like in [8].

[7] ST:<s>«Spare me the narrow misses, Bill, what have you got?»
TT:<s>«Não me fale do que perdi, Bill. <s>O que é que ainda tem?»
BT:<s>«Don't tell me what I missed, Bill. <s>What do you still have?»

[8] ST:<s>É terrível: não o esqueço um minuto.
LT:<s>It's terrible: I can't forget him for one minute.
TT:<s>It is terrible. <s>I cannot forget him for a moment.

Conversely, many changes from hard to soft punctuation marks seemed to involve joining sentence fragments together like in [9] and [10].

[9] ST:<s>Que dúvida! <s>Todas as suspeitas recaíam sobre a bela filha do dono da casa!
LT:<s>No doubt! <s>Every suspicion fell upon the beautiful daughter of the owner of the house!
TT:<s>Beyond a doubt, all the evidence implicated the beautiful daughter of the owner of the house!

[10] ST:<s>I never did dream much. <s>Which simply means, I understand, that I don't remember my dreams.
TT:<s>Nunca fui muito de sonhar, o que significa simplesmente, tanto quanto sei, que não me lembro do que sonho.
BT:<s> I never dreamed much, which simply means that, as far as I know, I don't remember what I dream.

These findings point in the same direction as Malmkjaer's (1997) previously mentioned observation regarding the normalization of punctuation, and lends further support to Toury's (1995) law of growing standardization in translation, where translators tend to opt for more standard or conservative options. It must be noted, however, that standardization could be more typical of the translation of literature than of other written genres. While literary authors often have poetic license to depart from the rules of a language, not all genres are known to employ unconventional grammar or lexis. For genres that follow conventions more strictly, there may simply be very little left to standardize.

The more detailed analysis of the 1133 parallel text segments focusing on sentence splitting and joining also compared contrasting sentence splitting and joining strategies. In both translation directions, splitting sentences by converting soft punctuation marks into hard ones was significantly more common than joining sentences by converting hard punctuation marks into soft ones. This cannot be interpreted as evidence of simplification, however, for breaking up longer sentences merely by changing punctuation does not necessarily simplify the text. As discussed above, these punctuation changes seem to support standardization more than simplification.

What was also interesting about the comparison of contrasting sentence splitting and joining strategies was the marked tendency for increased coordination in the English translations and for increased subordination in the Portuguese translations, which lend evidence to the possible effect of discoursal differences between languages discussed by Kaplan (1972). Despite this difference between Portuguese and English, both increased coordination and increased subordination result in more explicit links between clauses. When joining independent sentences by coordination, translators insert conjunctions, connecting sentences that are not explicitly linked to each other in source texts. Likewise,

when transforming independent sentences into subordinate clauses, hypotactic connections which are not present in source texts are introduced by translators. In both cases, therefore, translations become more explicit, lending further support to Blum-Kulka's (1986) explicitation hypothesis. Moreover, the fact that in the present study the tendency to transform simple sentences into compound and complex sentences was more pronounced than the reverse tendency to break down compound and complex sentences into simple ones contradicts the idea that translations tend to be syntactically simpler than source texts.

Putting it all together, this study showed that asyndetons seem to be at the root of what, irrespective of language direction, often prompts literary translators to alter the sentence boundaries of source texts, either by joining asyndetic sentences by coordination or subordination (and making the relation between them explicit) or by splitting asyndetic clauses into separate sentences (and normalizing the text in the process).

6. Conclusion

Most translation analyses involving parallel corpora available to date are centred on sub-sentential elements, with very little being known about translation shifts that transcend the level of the sentence. This study made use of a parallel, bidirectional corpus annotated for sentence-boundary shifts to analyse the translation of over 90 thousand source-text sentences, and then examined in closer detail over one thousand segments involving sentence joining and sentence splitting. English literary translators were found to stick a little less closely to source-text sentence boundaries than their Portuguese counterparts, but overall there was an over 90% rate of sentence preservation in both translation directions. Another slight difference noted was Portuguese translators increased subordination, while English translators increased coordination. Apart from that, the performance of both groups was remarkably similar in all other aspects of sentence joining and splitting analysed, suggesting that the influence of common translation norms prevailed over language-specific differences between English and Portuguese. In particular, the study disclosed evidence of a common tendency for standardization and explicitation in both translation directions, and showed that the presence of asyndetons in source texts was what often prompted translators to split or join sentences. It is hoped that the present study will stimulate further research with regard to sentence boundaries in translation, focussing on other language pairs and different genres.

References

Baker, M.1993. 'Corpus linguistics and translation studies: implications and applications' in M. Baker, G. Francis G. and E. Tognini Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, 233-250.

Baker, M. 1996. 'Corpus-based Translation Studies: the challenges that lie ahead' in H. Sommers (ed.) *Terminology, LSP and Translation Studies in Language engineering in honour of J. C. Sager.* Amsterdam and Philadelphia: John Benjamins, 175-187.

Barlow, M. 2002. 'ParaConc: concordance software for multilingual parallel corpora in *Proceedings of LREC. Third International Conference on Language Resources and Evaluation*, Las Palmas, 29-31 May 2002, 20-24. Online http://www.mt-archive.info/LREC-2002-Barlow.pdf [14/04/2016]

Beeby, A., Rodríguez, P. & Sánchez-Gijón, P. (eds.) 2009. *Corpus use and learning to translate (CULT): An Introduction*. Amsterdam and Philadelphia: John Benjamins.

Blum-Kulka, S. 1986. 'Shifts of cohesion and coherence in translation' in J. House and S. Blum-Kulka (eds.) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*. Tübingen: Gunter Narr, 17-35.

Bisiada, M. 2013. *From hypotaxis to parataxis: An investigation of English–German syntactic convergence in translation*. University of Manchester PhD thesis. Online http://ethos.bl.uk/OrderDetails.do?did=1&uin=uk.bl.ethos.603111 [14/04/2016].

Chan, S-W. 2014. 'The development of Translation Technology' in Chan, S-W (ed.) *The Routledge Encyclopedia of Translation Technology*. London: Routledge.

Christ, O. Schulze, B., Hofmann, A. and Koenig, E. 1999. *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual*, Institute for Natural Language Processing, University of Stuttgart, March 8, 1999 (CQP V2.2). Online http://cwb.sourceforge.net/documentation.php [14/04/2016].

Danielsson, P. and Ridings, D. 1997. 'Practical presentation of a 'vanilla' aligner'. Presentation at the *TELRI Workshop on Alignment and Exploitation of Texts*, Ljubljana, 1-2 February 1997.

Delaere, I., De Sutter, G. and Plevoets, K. 2012. 'Is translated language more standardized than non-translated language?' *Target*, 24/2:204-224.

Fabricius-Hansen, C. 1999. 'Information packaging and translation: aspects of translational sentence splitting (German-English/Norwegian' in M. Doherty (ed.) *Sprachspezifissche Aspekte der Informationsverteilung.* Berlin: Akademie Verlag, 175-214.

Frankenberg-Garcia, A. 2008. 'Suggesting rather special facts: a corpus-based study of distinctive lexical distributions in translated texts'. *Corpora*, 3/2, 195-211.

Frankenberg-Garcia, A. 2009. 'Are translations longer than source texts? A corpus-based study of explicitation' in Beeby, A., Rodríguez, P. & Sánchez-Gijón, P. (eds.), 47-58.

Frankenberg-Garcia, A. 2014. 'Understanding Portuguese Translations with the Help of Corpora' in T. Sardinha and T. Ferreira (eds.) *Working with Portuguese Corpora*. London: Bloomsbury, 161-176.

Frankenberg-Garcia, A. 2015. 'Training translators to use corpora hands-on: challenges and reactions by a group of 13 students at a UK university'. *Corpora*, 10/2, 351-380.

Frankenberg-Garcia, A. 2016. 'A corpus study of loans in translated and non-translated texts' in G. Corpas Pastor and M. Seghiri (eds.) *Corpus-based Approaches to Translation and Interpreting: from theory to applications*. Frankfurt: Peter Lang, 19-42.

Frankenberg-Garcia, A. and Santos, D. 2003. 'Introducing COMPARA, the Portuguese-English parallel translation corpus' in F. Zanettin, S. Bernardini & D. Stewart (eds.), 71-87.

Frankenberg-Garcia, A., Santos, D. and Silva, R. 2006. *COMPARA: sentence alignment revision and markup*. Online http://www.linguateca.pt/COMPARA/SentenceAlignment.pdf [14/04/2016]

Gale, W. and Church, K. 1993. 'A program for aligning sentences in bilingual corpora.' *Computational Linguistics*, 19/1, 75-102.

Granell, X. 2015. *Multilingual Information Management: Information, Technology and Translators*. Elsevier.

Hansen-Schirra, S., Neumann, S. and Vela, M. 2006. 'Multi-dimensional annotation and alignment in an English-German translation corpus' in *NLPXML '06 Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing*, 35-42.

Hofland, K. and S. Johansson. 1998. 'The Translation Corpus Aligner: a program for automatic alignment of parallel texts' in S. Johansson and S. Oksefjell (eds.) *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi, 87-100.

Johansson, S. 2007. *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies*. Amsterdam and Philadelphia. John Benjamins.

Johansson, S. and Hofland, K. 2000. 'The English-Norwegian Parallel Corpus: current work and new directions' in P. Botley, T. McEnery and A. Wilson (eds.), 134-147.

Kaplan, R. 1972. *The Anatomy of Rhetoric: prolegomena to a functional theory of rhetoric.* Philadelphia: The Centre for Curriculum Development Incorporation.

Lörscher, W. 2005. 'The Translation Process: methods and problems of its investigation'. *Meta Translators' Journal,* 50(2) 597-608.

Macken, L., Lefever, E. and Hoste, V. 2008. 'Linguistically-Based Sub-Sentential Alignment for Terminology Extraction from a Bilingual Automotive Corpus' in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 529-536.

Malmkjær, K. 1997. 'Punctuation in Hans Christian Andersen's stories and their translations into English' in F. Poyatos (ed.) *Nonverbal Communication and Translation: New Perspectives and Challenges in Literature, Interpretation and the Media*. Amsterdam and Philadelphia: John Benjamins, 151-162.

Mauranen, A. and Kujamäki, P. (eds.) 2004. *Translation Universals, Do They Exist?* Amsterdam and Philadelphia: John Benjamins.

O'Brien, S. 2013. 'The Borrowers: researching the cognitive aspects of translation'. *Target*, 25/1, 5-17.

Pérez-Blanco, M. 2009. 'Translating stance adverbials from English into Spanish: a corpus-based study'. *International Journal of Translation*, 21, 41-55.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Ramm, W. 2004. 'Sentence-boundary adjustment in Norwegian-German and German-Norwegian translations: first results of a corpus-based study' in K. Aijmer and H. Hasselgard (eds.) *Translation and Corpora*. Gothenburg: Acta Universitatis Gothoburgensis, 129-147.

Serbina, T. 2014. 'Sentence splitting in the translation pair English-German', Paper presented at the *4th Using Corpora in Contrastive and Translation Studies Conference*, Lancaster University, UK, 24-26 July 2014.

Toury, G.1987. 'The nature and role of norms in literary translation' in J. Holmes, J. Lambert and R. van den Broeck (eds.) *Literature and Translation*, Leuven: Acco, 83-100.

Toury, G. 1995. *Descriptive Translation Studies – and Beyond.* Amsterdam and Philadelphia: John Benjamins.

Vanderauwera, R. 1985. *Dutch Novels Translated into English: the transformation of a 'minority' literature.* Amsterdam: Rodopi.

Véronis, J. (ed.) 2000. *Parallel Text Processing: alignment and use of translation corpora.* Dordrecht: Kluwer Academic Publishers.

Xiao, R. and Dai, G. 2014. ‘Lexical and grammatical properties of translational Chinese: translation universal hypotheses reevaluated from the Chinese perspective’. *Corpus Linguistics and Linguistics Theory*, 10, 1: 11-55.
Zanettin, F., Bernardini, S. and Stewart, D. (eds.) 2003. *Corpora in Translation Education.* Manchester: St. Jerome Publishing.