Can a corpus-driven lexical analysis of human and machine translation unveil discourse features that set them apart?

Ana Frankenberg-Garcia
University of Surrey

There is still much to learn about the ways in which human and machine translation differ with regard to the contexts that regulate the production and interpretation of discourse. The present study explores whether a corpus-driven lexical analysis of human and machine translation can unveil discourse features that set the two apart. A balanced corpus of source texts aligned with authentic, professional translations and neural machine translations was compiled for the study. Lexical discrepancies in the two translation corpora were then extracted via a corpus-driven keyword analysis, and examined qualitatively through parallel concordances of source texts aligned with human and machine translation. The study shows that keyword analysis not only reiterates known problems of discourse in machine translation such as lexical inconsistency and pronoun resolution, but can also provide valuable insights regarding contextual aspects of translated discourse deserving further research.

Keywords: machine translation, MT, professional translation, discourse, parallel corpora, keyword analysis

1. Introduction

Despite the remarkable advances in machine translation (MT) over the past years, a known limitation of MT is that it still operates predominantly at the level of sentences viewed independently of one another. As a result, sentences that appear to be correct on their own may not fit well into a paragraph or a document because of problems such as incorrect pronoun selection or lexical inconsistency. For example, it is not uncommon to see a feminine noun in one sentence being referred to by a masculine pronoun in an adjacent sentence, or for a word to be translated in different ways throughout a document.

https://doi.org/10.1075/target.20065.fra | Published online: 8 September 2021
Target 34:2 (2022), pp. 278–308. ISSN 0924-1884 | E-ISSN 1569-9986

In response to this challenge, there has been a growing body of research devoted to addressing supra-sentential linguistic dependencies within texts (e.g., Carpuat and Simard 2012; Guillou 2013; Hardmeier 2014; Webber, Popescu-Belis, and Tiedemann 2017; Popescu-Belis et al. 2019). At the same time, Läubli, Sennrich, and Volk (2018, 4791) point out that to assess the quality of MT, there is a growing need to "shift towards document-level evaluation." This is because improvements in supra-sentential links cannot be detected in systems that only evaluate the quality of individual sentences.

One aspect of MT discourse research and evaluation that has received less attention is the question of how texts and their translations are shaped by the contexts in which they occur. This study explores whether a corpus-driven lexical analysis can provide insights into how discourse produced by human translation (HT) and MT differ, drawing attention to not only known issues regarding links between sentences, but also to more elusive problems affecting the contexts that regulate the production and interpretation of texts.

2. Background

A text is only coherent if its readers can activate the knowledge needed to make it coherent. As De Beaugrande and Dressler (1981, 12) explain, "a text does not make sense by itself, but rather by the interaction of text-presented knowledge with people's stored knowledge of the world." Catford (1965) refers to the strictly linguistic environment of words in texts as *co-text*, and to the broader, world-knowledge-dependent way in which words are presented by the producers and interpreted by the recipients of texts as *context*. In this sense, pronouns, lexical repetition, connectives, and other cohesive devices typically addressed in MT discourse research are primarily co-textual. Context, on the other hand, hinges on knowledge-dependent factors affecting how language users produce and interpret words in texts (van Dijk 1977).

At this juncture, it is important to note that the term 'context' has been typically used in MT papers to denote what is being referred to here, from the viewpoint of Translation Studies, as 'co-text'. To be completely clear, in this paper 'context' is not about adjacent words or sentences in a document (this is 'co-text'), but rather about external factors affecting how words are interpreted in texts. For example, the abbreviation 'MT' is a co-textual element of cohesion in this text because it is a term that is consistently used throughout the paper, linking what is said in one sentence with what is stated in other sentences further along. However, 'MT' can only be used coherently from an external, contextual perspective because of my world-knowledge assumption that the use of an abbreviation in

brackets after its full form is first mentioned means that readers will henceforth recognise that in this text 'MT' is being used to refer to 'machine translation'.

In translation, it is important to acknowledge that source- and target-language recipients may not always share the same contextual knowledge that is necessary to make sense of co-textual cues in texts. The use of brackets mentioned above is a standard convention in most written languages, so it can be assumed that it would not be a problem in translation. However, some world-knowledge assumptions are less universal. For example, in a recent news item about a Lisbon football club published in a Portuguese newspaper, the club was referred to as *Sporting* (the name of the club), *os Leões* 'the Lions', *a equipa leonina* 'the lion club', *os verdes e brancos* 'the green and white', and *os lisboetas* 'the Lisboners'. The reporter's expectation was that the readers of the newspaper could activate the contextual knowledge needed to understand that all five terms refer to the same entity, and thus make sense of the co-textual links between them. However, a direct translation of this news piece into another language would lack coherence if the intended target-language readers did not know that Sporting football club's symbol is a lion, that their colours are green and white, and that the club is from Lisbon.

The above example illustrates why mirroring the co-textual cohesive devices of source texts may not always work in translation. In Translation Studies, there has been considerable interest in the ways in which translators mediate discourse, adapting it to target readerships. As House (2006, 356) explains, translation involves recontextualisation, which means "taking a text out of its original frame and context and placing it within a new set of relationships and culturally conditioned expectations." When professional translators choose their words, they are trained to consider not just linguistic equivalence and document-level consistency, but also contextual factors such as the purpose of the translation and the target readership. For example, in a translation from English into Portuguese, should '70 miles' be translated literally into *70 milhas*? Or should the value be converted into kilometres for readers from Brazil and Portugal, where the metric system is used? If the latter, should the exact conversion value of '112.65 km' be given, or would the approximation 'around 110 km' be more appropriate? Or should the original distance in miles be preserved and the metric conversion given in brackets? These are all possible scenarios, but it is not reasonable to decide which strategy to use without contextual information about who and what the translation is for.

When it comes to generic MT, however, the training data used in its development does not normally take translation context into account.[1] According to Koehn and Schroeder (2007, 224), such data "is typically collected opportunistically from wherever it is available." This leaves little room for factoring in how external context has motivated different translation decisions. Moreover, in widely used open-access training data sources like Europarl (Koehn 2005), for example, parallel text alignment is based on linguistic equivalence; however, without the addition of tags indicating translation direction, it is not possible to discriminate which is the source and which is the target language. However, HT is not symmetrical (Klaudy 2009, 2017), so translating A into B is not necessarily the reverse of translating B into A. Translation asymmetry impacts not just linguistic choices, but also discoursal choices, including decisions that depend on context. For instance, Frankenberg-Garcia (2016) and Klaudy (2017) note how foreign words are handled differently in different translation directions, depending on assumptions made about which foreign words target readers can recognise. For example, foreign words left in English are not usually a problem for educated Portuguese readers, but foreign words left in Portuguese can be hard for educated English readerships to understand. To go back to the Sporting football club example, a professional translating that Portuguese news item into another language is likely to add glosses to the expressions that target readers might not understand, or replace those expressions with ones that readers will. Conversely, if translating a foreign news item about Sporting for a Portuguese readership, a professional may deliberately delete glosses which Portuguese readers would find redundant. This illustrates why the reversibility of MT training data can be problematic if contextual knowledge assumptions external to the text are not taken into account.

Another limitation of MT training data that has implications at the level of context is that the translations upon which they are based are not always carried out by professionals. The Open Subtitles and TED talks parallel text collections available from OPUS (Tiedemann 2012), for example, are excellent sources of MT training data representing more informal spoken registers, but because they are the product of fan translations, there is less control over their quality. There is no information about the level of source- and target-language proficiency of the volunteers who carry out the translations, let alone about their understanding of the discourse mediation strategies employed by experienced professional translators. More worryingly, there is no control over the extent to which such collections

1. Note, however, that customised MT trained on domain-specific data can focus on specific contexts. For example, a medical MT engine can be trained to render 'theatre' in the sense of 'operating theatre' rather than in the sense of 'movie theatre'.

have been influenced by MT in the first place, which could result in the circularity of using MT output as training data to develop MT.

In contrast, although translation solutions by experienced professional translators may differ, professionals generally understand what needs to be added to or deleted from a translation, or what needs to be otherwise adapted for successful recontextualisation. For example, professional translation strategies can involve giving a deliberate foreign feel to a target text (for instance, by borrowing words from the source language), or consciously mediating discourse (by adding footnotes or other extra information, for example) to enhance the readability of a text among target language audiences (Schleiermacher [1813] 2004).

In summary, MT research has recognised the limitations of translating isolated sentences, and has made advances towards establishing better links between sentences and developing document-level MT evaluation metrics. However, less attention has been paid to source-text contexts and how discourse is recontextualised in translations. Moreover, although there is no doubt about the value of opportunistic bilingual text collections for the development of MT (especially as there are not enough large, good-quality parallel corpora tagged for translation direction available), they are less suited to helping us understand directional shifts in translation, including how professional translators recontextualise source texts for target-text readerships.

To address this challenge, this study explores whether a corpus-driven lexical analysis that compares professional translation in a known language direction with MT can shed further light on discoursal differences between the two, beyond well-known problems such as lexical inconsistency and pronoun resolution.

3. Method

This section describes the corpus used in the study and how the comparison of HT and MT was undertaken.

3.1 Materials

This study draws on source texts and professional HT from COMPARA (2010), an open-access, bidirectional parallel corpus of Portuguese and English literary fiction (Frankenberg-Garcia and Santos 2003). COMPARA consists of authentic, published translations carried out by professionals from Portuguese into English and from English into Portuguese. Although COMPARA is bidirectional, the present analysis looks only at Portuguese to English translations.

As literary translation does not normally make use of MT (Toral and Way 2018), and as the English translations in COMPARA are authentic, published translations dating back to the 1980s and 1990s – a time before the use of MT became widespread – it can be assumed that the translations in the corpus were not influenced by MT.

To ensure the analysis was not skewed by individual author or translator performances, a balanced corpus comprising the work of fifteen different authors translated by fifteen different translators was used in the study (see Table 1).

Table 1. Authors and translators represented in the study*

Text ID	Source-text title	Author	Translator
PBRF2	*A Grande Arte*	Rubem Fonseca	Ellen Watson
PMMC1	*Vozes Anoitecidas*	Mia Couto	David Brookshaw
PBJS1	*O Xangô de Baker Street*	Jô Soares	Cliff Landers
PBPC1	*O Alquimista*	Paulo Coelho	Alan Clarke
PPSC1	*A Confissão de Lúcio*	Mário de Sá-Carneiro	Margaret Jull Costa
PPJS1	*Sinais de Fogo*	Jorge de Sena	John Byrne
PBMA3	*Dom Casmurro*	Machado de Assis	John Gledson
PPLJ1	*A Costa dos Murmúrios*	Lídia Jorge	Natália Costa
PBAD2	*Os Sinos da Agonia*	Autran Dourado	John Parker
PPMC1	*Um Deus Passeando pela Brisa da Tarde*	Mário de Carvalho	Gregory Rabassa
PBAA2	*O Mulato*	Aluísio Azevedo	Graeme McNicoll
PPEQ3	*Alves e Companhia*	Eça de Queirós	John Vetch
PPJSA1	*Ensaio sobre a Cegueira*	José Saramago	Giovanni Pontiero
PBCB2	*Estorvo*	Chico Buarque	Peter Bush
PBMAA1	*Memórias de um Sargento de Milícias*	Manuel Antônio de Almeida	Ronald Sousa

* Full references are available at https://www.linguateca.pt/COMPARA.

COMPARA's online interface allows for the creation of subcorpora based on selected texts, such as those presented in Table 1, but for copyright reasons it does not permit their full download. Moreover, the tool restricts the number of paral-

lel concordances that can be retrieved such that no more than one third of each bitext (i.e., aligned source and target text) is presented each time a query is performed. As the texts in COMPARA are of unequal length, longer texts will yield more concordances. Therefore, to achieve balance, the total number of source-text words generated by the shortest bitext in the selection was used as a benchmark, and the other texts were cut down to its approximate length. In this way, it was possible to arrive at a balanced corpus of fifteen Portuguese–English bitexts of between 4000 and 5000 source-text words each.

The concordances extracted for each query are in sequential order. However, to cut down the output to within the copyright limit, random concordances may be omitted. This would be a serious limitation if the investigation required one to read the texts linearly, from beginning to end. However, as it will be explained in Section 3.2, the present study did not contemplate an analysis of discourse features that depend on reading longer uninterrupted stretches of text.

To obtain the MT corpus used in the study, the Portuguese source-text segments downloaded from COMPARA were machine translated into English using Google Translate.[2] Google Translate uses neural MT technology for the Portuguese–English language pair (Turovsky 2016), though little else is known about how it operates. This study does not aim to foster the development of Google Translate. The choice for using it was simply that it is free, readily available, and is a popular generic MT system used by the public in general. Note that it was not possible to guarantee that Google Translate did not use the HTs from COMPARA as part of its training data in the first place. However, because this data is not directly available online (it can only be retrieved through specific searches within COMPARA), and because COMPARA is negligible in size when compared to the enormous quantities of training data used by Google, it is highly unlikely that it would exert much influence on how Google Translate operates.

Once the MT output was obtained, it was aligned with the HTs from COMPARA using the full-sentence source-text segments as a common denominator. It was thus possible to obtain a balanced and perfectly aligned parallel corpus of source texts (ST corpus), human translations (HT corpus), and machine translations (MT corpus), as shown in Figure 1.

The corpus was compiled in Sketch Engine (Kilgarriff et al. 2014). Both the HT and MT corpora were tagged with the TreeTagger part-of-speech tagset for English developed by Helmut Schmid and modified by the Sketch Engine team (pipeline version 2).

2. Performed through the now discontinued Google Translator Toolkit (2019).

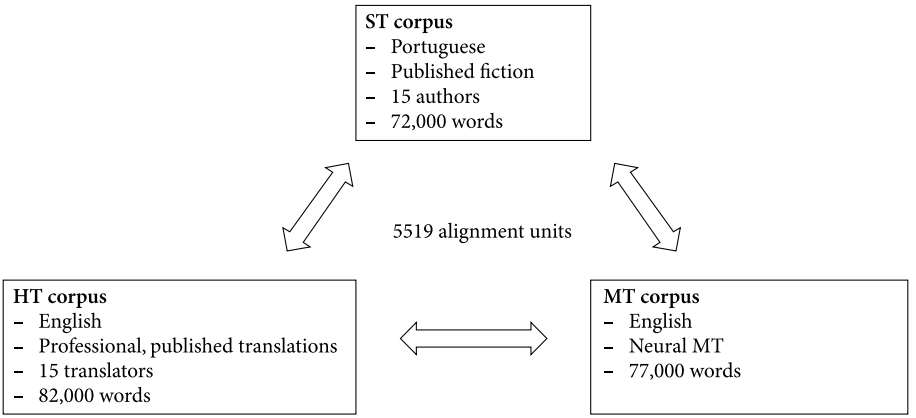

Figure 1. Three-dimensional parallel corpus used in the study

3.2 Procedure

Using parallel corpora to evaluate discourse in MT is a recent technique (Lapshinova-Koltunski and Hardmeier 2017; Guillou et al. 2018). Previous research has been guided by known challenges for MT, like pronoun resolution, in a corpus-based approach. The approach taken in this study is novel in that its starting point is corpus-driven.[3] In other words, instead of running specific corpus queries to examine known issues (e.g., pronouns), in this study the corpus as a whole was used as a point of departure to gain new insights into how HT and MT differ. A standard procedure in corpus-driven approaches is the comparison of lexical distributions in two corpora. For example, Frankenberg-Garcia (2008) uses this method to analyse distinctive lexis in translated and non-translated texts. A similar corpus-driven approach can be applied to uncover distinctive lexis in HT and MT, which can then be further inspected from a corpus-based perspective to determine whether lexical contrasts impact discourse.

The first step in the study involved comparing the HT and MT corpora through keyword analysis. Keyword analysis is a well-known procedure used in corpus linguistics to identify linguistic elements that are exceptionally frequent in a focus corpus when compared to a reference corpus (Kilgarriff 2009). The following formulas were used to extract words that were distinctively more frequent in the MT corpus in relation to the HT corpus, and, conversely, to extract words whose frequency stood out in the HT corpus when compared with the MT corpus:

3. See Tognini-Bonelli (2001) for a discussion of corpus-based and corpus-driven approaches.

$$\frac{MT\ fpm+N}{HT\ fpm+N} \quad \text{and} \quad \frac{HT\ fpm+N}{MT\ fpm+N}$$

MT fpm is the normalised frequency (per million) of an element in the MT corpus, *HT fpm* is the equivalent frequency of the same element in the HT corpus, and *N* is the smoothing parameter used to avoid dividing by zero. As discussed in Kilgarriff (2009), while the standard smoothing parameter is $N=1$, this value can be adjusted so as to prioritise higher- or lower-frequency keywords. Given the relatively small size of this study's corpora, the smoothing parameter used was $N=1000$, which prioritises words in the higher-frequency range. This reduces the chances of capturing idiosyncratic discrepancies concentrated in just one text.

Keyword extraction can be performed on different corpus attributes, such as word forms, tags, lemmas, and so on. In the present analysis, the Sketch Engine *lempos-lc* attribute was used to extract case-insensitive lemmas tagged according to part-of-speech category.[4] This allowed for the conflation of, for example, *His* and *his*, and *car* and *cars*, but at the same time for the separation of *house* (noun) from *house* (verb) in the extraction. After extracting the top 200 HT and MT keywords, a sample was then inspected in further detail. Through close reading of the three-dimensional parallel concordances for keywords in the ST, HT, and MT corpora, a qualitative analysis was undertaken to investigate whether the distinctive lexical distributions observed could impact discourse.

4. Results

Table 2 summarises the distribution of the top 200 keywords in each translation corpus sorted by word class and ordered by keyness score rank. In the last category, 'NOT_TRANSLATED' represents entire source-text sentences which were either intentionally or mistakenly left out of the translation, highlighting the fact that this is a decision or error only human translators can make. Overall, it is possible to see marked differences in the use of modals, prepositions, and pronouns in the two corpora, with substantially more HT than MT keywords pertaining to those closed-class grammatical categories. The HT and MT keywords pertaining to open-class lexical words – nouns (including proper names and foreign words), verbs, adjectives, and adverbs – to reach the study's top-200 threshold were more

4. The tagger failed to classify 7.75% of the lemmas (marked with an *x* in the output), and misclassified 4.25% (e.g., *because* was classified as a preposition). These were manually revised where part-of-speech was straightforward to identify, otherwise they were marked as ambiguous. Indefinite pronouns, proper nouns, and foreign words that had been broadly classified as nouns by the tagger were manually differentiated.

balanced in number. To uncover possible discourse implications behind the lexicogrammatical keyword discrepancies in Table 2 in more depth, ST–HT–MT parallel concordances for a sample of grammatical and lexical keywords will be examined next.

Table 2. Top 200 HT and MT keywords sorted by word class and ordered by keyness score rank

Keyword class	Distinctive in HT	Distinctive in MT
Adjective	own, such, unable, male, special, odd, flat, wise, final, very, splendid, fellow, bright, long, indian	great, little, black, last, much, beautiful, good, silent, ready, thin, full, first, low, high, open, crazy, quiet, sick, front, worth, holy, gray, old, rich, handsome, natural, wet, rare
Adverb	just, over, back, quite, as, together, really, once, now, away, ever, else, ill, merely, around, well, on, certainly, probably, rather, immediately, though, longer, up, out, enough, simply, clearly, very, silently, in, afterwards, too	very, soon, always, also, already, not, anyway, asleep, however, there, maybe, sometimes, little, barely
Foreign	senhor, plaza, senhora	d, mainata, nhonhô, nhonho
Modal	might, should, could, would, can, ought, shall	–
Noun	fellow, wife, part, way, massa, round, feeling, area, bit, kind, place, theatre, mind, affair, colour, cattle, slave, negro, sort, bedroom, reply, music, inspector, evening, thought, use, jacket, destiny, town, raven, stuff, spy-hole, shape, note, maid, country, omen, fine, desk, horse, phone, side, home, line, people, staircase, mummy, moustache, longing, lobby	mosque, house, guy, mr, color, hour, earth, beast, hall, eye, sign, woman, other, face, personal, mustache, legend, crow, background, porch, college, band, ox, animal, photograph, pastor, fight, couch, scent, neighbourhood, devil, yard, ceremony, afternoon, son, floor, street, year, head, mouth, table, step, sheep, square, newspaper, land, suit, stop, partner, jailer, beginning, sage, motorcycle, name, city, wall, foot, doubt, lady, information, stair, care, will, clock, favour, contrary, song, wonder, revenge, gate
Preposition	out, up, off, into, along, about, over, towards, on, through, after, around, onto, for, within, despite, up, near	without, of, in
Pronoun	their, its, his, them, she, her, our, someone, something, myself, everyone, herself, anyone, one, nobody	everything, me, they, it

Table 2. *(continued)*

Keyword class	Distinctive in HT	Distinctive in MT
Proper name	Helen, Mesquita, Gervásio, José, Proserpinus, Alves, Trifenus, Pádua, Lúcio	Helena, Gervasio, Jose, Proserpino, Godfrey, Padua, Azariah, Trifeno, de, Lucius
Verb	hold, get, bring, use, become, go, carry, keep, manage, put, decide, let, realise, stand, need, ring, allow, suggest, inform, weep, place, wear, grow, have, happen, round, build, observe, imagine, begin, catch, wonder, remain, find, stick, reply, summon	give, want, do, enter, know, continue, cry, jump, close, hurt, live, understand, scream, lean, shake, stay, believe, collect, blind, join, serve, count, wish, notice, return, remember, lose, love, confess, wrap, smell, form, resume, fire, delay, conclude, look, come, call, answer, throw, cover, leave, meet, pull, receive, repeat, realize, save, wake, shine, fulfill
Other/ Ambiguous	NOT_TRANSLATED, some, as, any, while, every, whether, which, why, an, those, on, though, like, one, in, whenever	this, these, because, the, so, that, two, another, yeah, if, oh, whose

4.1 Grammatical keywords

This section examines in further detail keyword differences in modals, prepositions, and pronouns. Closed-class words like these occur frequently, generating hundreds of concordance lines each, and it is beyond the scope of this study to manually inspect them all. This part of the investigation focuses on a systematic qualitative analysis of one HT and one MT grammatical keyword per part-of-speech category.

4.1.1 *Modals*

The keyword analysis identified seven distinctive modal verbs in the HT corpus and none in the MT corpus (Table 2). The most distinctive modal in the HT corpus was 'might', with 49 occurrences against just 11 in MT (keyness score 1.34). The modal was present in 80% of the HT texts, so it cannot be dismissed as simply being a matter of stylistic preference. Parallel concordances for 'might' in the HT corpus without the same modal in the MT alignment returned 45 hits. A qualitative analysis of these concordances resulted in 23 HT concordances with no modality in MT (as in Example (1)), and 22 HT concordances with a different modal or a comparable modality marker in MT (as in Example (2)).

(1)

Conc. ID	ST	HT	MT
PPEQ3 302	veio-lhe o terror que o sogro não estivesse em casa	he began to fear that his father-in-law **might** not be at home	terror came to him that his father-in-law was not at home
PMMC1 552	ainda pisava na mina	she **might** tread on a mine	he was still walking in the mine

(2)

Conc. ID	ST	HT	MT
PBAD2 950	**Podem** me envolver	They **might** involve me	You **can** get involved
PBMA3 254	**Talvez** valha a pena dá-la	It **might** be worthwhile giving it here	**Maybe** it's worth giving

Although the source language does not have modal verbs, modality can be expressed through other linguistic resources in Portuguese. The concordances in Example (1) indicate that human translators outperform MT with regard to inferring mood from the context when it is ambiguous or not explicit in the source language. The concordances in Example (2), on the other hand, show that MT can handle mood when it is expressed by means of explicit modality markers in the source language. As shown, possibility can be expressed in Portuguese through the verb *poder* and the adverb *talvez*. This could also explain why the direct translation of the adverb 'maybe' is distinctive in the MT corpus (see Table 2).

4.1.2 *Prepositions*

The keyword analysis highlighted the prevalence of 18 prepositions in the HT corpus against 3 in MT (Table 2). The most distinctive prepositions in the two types of translation – 'out' in the HT and 'without' in the MT – are explored in further detail.

The preposition 'out' returned 272 hits in the HT corpus compared to 126 in the MT corpus (keyness score 1.59). Close reading of the 215 parallel concordances for 'out' in the HT corpus without 'out' in the MT alignment returned

- 175 HT concordances where 'out' is part of a phrasal verb like 'find out', while the MT output is a one-word literal translation from the ST, as in Example (3);
- 23 HT concordances with other senses of 'out', while the MT is literal, as in Example (4);
- 10 HT concordances where 'out' is part of phrases beginning with 'out of' in the sense of 'because of', while the MT is mistranslated, as in Example (5);

– 7 HT concordances where 'out' is part of phrases beginning with 'out of' in the sense of 'without', while the MT uses a one-word literal translation of the ST, as in Example (6).

(3)

Conc. ID	ST	HT	MT
PPEQ3 275	uma pancada surda que o **devastava**	a silent blow that **knocked** him **out**	a deaf thump that **devastated** him
PBAD2 796	O pai mandou que **apagasse** a candeia	His father told him to **put out** the lamp	His father commanded him to **extinguish** the lamp

(4)

Conc. ID	ST	HT	MT
PBAD2 314	Nenhum jeito possível	No way **out**	No way possible
PBAA2 590	ela que vá dando os seus passeios a pé	she should be **out** taking walks	she will go giving her walks on foot

(5)

Conc. ID	ST	HT	MT
PPLJ1 132	a tinham trazido ali por instinto de sobrevivência	had brought her there **out of survival** instinct	had brought her there by instinct for survival
PBMA3 123	Se soubesse, não teria falado, mas falei pela veneração, pela estima, pelo afeto	If I'd known, I wouldn't have spoken, but I did so **out of veneration, out of esteem, out of affection**	If I had known, I would not have spoken, but I spoke of veneration, of esteem, of affection

(6)

Conc. ID	ST	HT	MT
PBAA2 553	Tinha o cabelo à escovinha; os sapatos grandemente **desproporcionados**	His hair was close cropped and his shoes terribly **out of proportion**	She had her hair brushed; the shoes were greatly **disproportionate**
PBJS1 968	E xingava, **descontrolado**	And he cursed, **out of control**	And he cursed, **uncontrolled**

As can be seen from Examples (3) to (6), despite 'out' being a grammatical word, the main reason for its prevalence in the HT corpus is lexical. Apart from the HT being less literal than the MT, it can be seen that the HT use of 'out' in phrasal verbs and other expressions tends to confer a less formal and more idiomatic tone on the translations, indicating a better appreciation of situations where informal language is more appropriate.

The most marked preposition in the MT, 'without', returns 112 hits in the MT corpus compared to 84 in the HT corpus (keyness score 1.19), and there are 50 parallel concordances for 'without' in the MT aligned concordances lacking the same preposition in the HT. Close reading of those concordances revealed that where MT uses 'without', the HT equivalent consists of:

- 25 concordances with negative adverbs such as 'not', as in Example (7);
- 11 concordances with negative prefixes and suffixes such as 'un-' and '-less', as in Example (8);
- 9 concordances with antonymous expressions, as in Example (9);
- 5 concordances with other words or phrases expressing negation, as in Example (10).

(7)

Conc. ID	ST	HT	MT
PBCB2 22	O menino…avista-me **sem** me ver	The kid…looks but does**n't** see me	The boy…sees me **without** seeing me
PPMC1 270	**sem** qualquer escrúpulo	with **no** scruples whatever	**without** any scruple

(8)

Conc. ID	ST	HT	MT
PPLJ1 669	**sem** conseguirem culpar nada	**un**able to blame anything	**without** being able to blame anything
PBMAA1 112	um filho **sem** mãe	A mother**less** child	a son **without** a mother

(9)

Conc. ID	ST	HT	MT
PPLJ1 845	como os homens que vivem **sem** ter tempo	like one of those men who are always rushing through life	like the men who live **without** time
PPEQ3 16	murmurou o guarda-livros, **sem** cessar de escrever	murmured the bookkeeper, as he **went on** writing	he bookkeeper murmured **without** interruption

(10)

Conc. ID	ST	HT	MT
PBRF2 691	**sem** a maioria dos dentes	**missing** most of his teeth	**without** most of his teeth
PPEQ3 31	**sem** a cor viva duma flor	he **lacked** the bright colour of a flower	**without** the living color of a flower

The analysis shows that the translators use a more varied repertoire of expressions of negation. Instead of translating the source-text preposition *sem* literally into 'without', translators resort to oblique translation strategies to produce more idiomatic target language renditions.

4.1.3 *Pronouns*

The keyword analysis identified 15 distinctive pronouns in the HT corpus and 4 in the MT corpus. There were also marked differences in the types of pronouns. Indefinite pronouns ('someone', 'something', 'everyone', 'anyone' and 'nobody') are more salient in the HT than in the MT, where only 'everything' is more frequent. Four gender-marked personal pronouns – 'his', 'she', 'her' and 'herself' – are key in the HT, in contrast to the MT personal pronouns, all of which are gender-neutral. Another interesting finding is the distinctive use of possessives in HT and MT, with 5 key possessives in the HT corpus – 'their', 'its', 'his', 'her', and 'our' – and none in the MT corpus. The most distinctive HT and MT personal pronouns – 'their' and 'me', respectively – were inspected in further detail.

The possessive 'their' has 191 hits in the HT corpus compared to only 112 in the MT corpus (keyness score 1.33). A search for 'their' in the HT corpus without the same form in the MT alignment returned 114 concordances. A qualitative analysis of the pronoun discrepancies indicated that there are:

– 64 concordances where a pronoun not present in the ST is inserted in the HT but not in the MT, as in Example (11);

– 34 concordances where 'their' in the HT results from a less literal translation of the ST, as in Example (12);
– 16 concordances where the pronoun was mistranslated in the MT, as in Example (13).

(11)

Conc. ID	ST	HT	MT
PBRF2 804	Muda de nome, de casa, pinta o cabelo, vai para a Bahia	Move away, change **their** names, dye **their** hair, go to Bahia	Change of name, of house, paints the hair, goes to Bahia
PPJS1 654	através das recordações de pais e tios	through the tales of **their** parents and aunts and uncles.	through the memories of parents and uncles

(12)

Conc. ID	ST	HT	MT
PPJSA1 679	O Governo e a Nação esperam que cada um cumpra o seu dever	The Government and Nation expect every man and woman to do **their** duty	The Government and the Nation expect each one to fulfill **his** duty
PBAA2 148	a dar-lhes a comida	to fix **their** meals	to give **them** food

(13)

Conc. ID	ST	HT	MT
PPEQ3 324	com os lábios unidos aos dele	with **their** lips together	with **his** lips joined to his
PBAD2 477	Iam silenciosos, rosário na mão	They walked in silence, their rosaries in **their** hands	They were silent, the rosary in **his** hand

Looking now at the most salient personal pronoun in the MT, there are 395 hits for 'me' in the MT corpus and 379 in the HT corpus (keyness score 1.09), and 78 parallel concordances for 'me' in the MT without the same pronoun in the HT alignment. One concordance was not translated in the HT corpus. The remaining 77 concordances comprise:

– 31 concordances where the equivalent pronoun 'I' is used in the HT, as in Example (14);
– 26 concordances without a corresponding pronoun in the HT, as in Example (15);

– 15 concordances where the corresponding pronoun in the HT is a possessive, as in Example (16);
– 5 concordances where the HT pronoun refers to another entity (evidencing mistranslation in the MT), as in Example (17).

(14)

Conc. ID	ST	HT	MT
PBJS1 167	há algo aqui que causa-me estranheza	there's one thing that **I** find strange	there is something here that causes **me** strangeness
PBMA3 257	O desuso é que me faz mal	**I**'m out of practice	The disuse is what makes **me** bad

(15)

Conc. ID	ST	HT	MT
PBAA2 163	Não me pareces a mesma	You're not yourself at all	You do not look the same to **me**
PPMC1 24	Mara…deixa-me numa pequena corrida	Mara…runs off.	Mara…leaves **me** in a little run

(16)

Conc. ID	ST	HT	MT
PPJS1 47	porque as experiências não me pertencem	because these experiments are not just **mine**	because the experiences do not belong to **me**
PPSC1 247	por mim, confesso, tive medo	I, for **my** part, felt afraid	for **me**, I confess, I was afraid

(17)

Conc. ID	ST	HT	MT
PBMAA1 294	custou muito a vir	it was hard for him to come back	it was very difficult for **me** to come here
PPJSA1 754	roubaste-me a vista dos olhos	you stole my eyesight	you stole **me** from the eyes

It can be seen that for both pronouns, the reasons for the discrepancies observed have less to do with known problems of MT pronoun resolution (in Example (13) and (17)), and more about professional translators using oblique strategies to render the translation more idiomatic (in Example (12), and (14) to (16)), and to resolve ambiguity (in Example (11)).

4.2 Lexical keywords

This section takes a closer look at lexical keyword differences in the HT and MT corpora through close reading of a selection of concordances with key lexical words. Open-class words like these are less frequent and scattered across specific texts. For example, the distinctive HT proper noun 'Helen' occurs in only one source, unlike a preposition such as 'out', which occurs in all HT and MT texts in the corpus. The analysis of individual lexical keywords is thus not particularly informative, as they could be simply the result of idiosyncratic choices. What is more interesting here is to explore whether there are groups of lexical keywords that behave similarly, which can unveil trends regarding differences between HT and MT. However, in this article there is not enough space to cover every possible pattern emerging from lexical keywords, and thus this part of the analysis focuses on findings pertaining to spellings, proper names, and foreign words.

4.2.1 *Spelling*

An immediately visible difference to emerge in the keyword analysis of the open-class words in Table 2 is spelling differences. The following spelling preferences can be observed in the HT and MT corpora, respectively: 'colour' and 'color', 'moustache' and 'mustache', 'realise' and 'realize', Gervásio' and 'Gervasio', 'José' and 'Jose', and 'Pádua' and 'Padua'. Although these are only surface-form differences that do not affect grammaticality or meanings, they have document-level contextual implications. The choice between British and American spellings reflected in the first three keyword contrasts is not random in the HT, but rather indicative of contextual knowledge of specific target readerships or style guides. The preservation of foreign characters, like the accents used in the three proper names, could in turn be interpreted as a contextual decision to deliberately confer a more exotic feel to the translation, which may happen when it is known from the context that a story plot is set in a foreign country.

4.2.2 *Proper names*

The keyword analysis in Table 2 also presents substantial differences in the use of proper names in the HT and MT. A quantitative summary of ST–HT–MT concordances for the proper names highlighted in Table 2 is shown in Table 3. First, it evidences the well-known problem of lexical consistency in MT. For example, 'Helena' is sometimes machine translated as 'Helena' and sometimes as 'Helen' in the same novel. Similarly, the translation of 'Mesquita', 'Godofredo', 'Gervásio', 'José', and 'Trifeno' – the majority of the proper names analysed – is not consistent in the MT.

Table 3. Key proper names (and their frequencies)

Text ID	ST	HT	MT
PPLJ1	Helena (40)	Helen (41)	Helena (30) Helen (10)
PPJS1	Mesquita (29)	Mesquita (30)	Mosque (25) Mesquita (4)
PPSC1	Gervásio (16)	Gervásio (20)	Gervasio (15) Gervasius (1)
PBMAA1	José (11)	José (11)	José (4) Jose (7)
PBAA2	José (17)	José (20)	José (8) Jose (4) Joseph (5)
PBMA3	José (10)	José (10)	José (9) Jose (1)
PBJS1	José (2)	José (2)	José (2)
PBRF2	–	José (1)	–
PPMC1	Proserpino (10)	Proserpinus (10)	Proserpino (10)
PPEQ3	Godofredo (17)	Godofredo (3) Alves (10) he (4)	Godfrey (9) Godofredo (7) Godfred (1)
PPEQ3	Alves (13)	Alves (13)	Alves (13)
PBMA3	Pádua (8)	Pádua (8)	Padua (8)
PPMC1	Trifeno (8)	Trifenus (8)	Trifeno (7) Trypho (1)
PPMC1	Azarias (8)	Azarias (8)	Azariah (8)
PPSC1	Lúcio (6)	Lúcio (6)	Lucius (6)
PPMC1	Lúcio (7)	Lucius (7)	Lucius (7)
Total	**202**	**210**	**202**

Further, the translation of the proper name 'Mesquita' exemplifies another known problem of MT, namely that of word-sense disambiguation. As shown in Example (18), the MT engine does not distinguish between the surname 'Mesquita' and the common noun *mesquita* 'mosque', translating the name of the character as if it were a person nicknamed 'The Mosque'.

(18)

Conc. ID	ST	HT	MT
PPJS1 545	O **Mesquita** escandalizou-se	**Mesquita** was quite shocked	The **Mosque** scandalized itself

In contrast, the HT of proper names is not only consistent, but also deliberate. As discussed in Section 4.2.1, foreign accents are consciously preserved in the names of characters where plots unfold in non-English speaking settings. Additionally, it is evident that discourse context plays a decisive role in determining whether proper names are translated. 'Helena' in PPLJ1 is nicknamed after the Greek mythology character Helen of Troy, which makes the English translation 'Helen' more appropriate than preserving the Portuguese form 'Helena'. Similarly, 'Proserpino', 'Trifeno', and 'Lúcio' in PPMC1 are characters in a novel set in ancient Rome, which explains why the translator opted for the Latin translations 'Prosperinus', 'Trifenus', and 'Lucius'. Note that in PPSC1, where 'Lúcio' refers to a Portuguese man, the Portuguese rendition of the name is preserved.

The only apparent inconsistency in HT – 'Godofredo' is translated as both 'Godofredo' and 'Alves' in PPEQ3 – occurs because 'Godofredo' (first name) and 'Alves' (surname) refer to the same character. The translator's choice to use the surname 'Alves' is consistent with how the character is referred to in the title of the novel (see Table 1).

With regard to the MT keyword *de* that was tagged as a proper name (see Table 2), the difference with the HT arises because of the inconsistent translation of nobility titles such as 'Visconde **de** Vilar', 'Marquis **de** Salles' but 'Baroness **of** Avare' in the MT, versus 'Viscount **of** Vilar', 'Marquis **of** Salles', and 'Baroness **of** Avare' in the HT.

A less explored trait uncovered by the analysis is the discrepancy in the number of proper names in the HT and MT. As shown in Table 3, the translators add 'Helena' (+1), 'Mesquita' (+1), 'Gervásio' (+4), 'José' (+3 in PBAA1 and +1 in PBRF2) where there are no matching names in the corresponding ST. Close reading of the parallel concordances that contain additional names reveal that the translators added names to remove possible ambiguity of referents in the translation, as exemplified in Example (19).

(19)

Conc. ID	ST	HT	MT
PPLJ1 132	Falando desse modo, tão baixo	As **Helen** spoke, in so soft a voice	Speaking thus, so low
PBAA2 856	não era isso! respondia o outro	it isn't that **José** answered	was not it! answered the other

There are also 4 instances where an ST name ('Godofredo') is replaced with a pronoun in the HT. In contrast, proper names are never added or replaced with pronouns in the MT.

4.2.3 *Foreign words*

Some of the key lexical differences in the HT and MT corpora listed in Table 2 involve non-English words. Even before consulting concordances for those words, it is clear that the ones that are distinctive in HT – *senhor, plaza*, and *senhora* – are easier for English readers to recognise than the distinctive MT foreign words *d, mainata, nhonhô*, and *nhonho*. This is not by chance. As can be seen in Example (20), the Portuguese *senhora* in the HT is very similar to *señora* in Spanish and *signora* in Italian, making it more likely to be understood by English readers. The same applies to *Senhor* in the HT. However, instead of transposing the abbreviated form *Sr.* used in the ST (which target readers might not recognise), the translator spelled out *Senhor* in full, which is again similar to the Spanish *Señor* and the Italian *Signor*. With regard to *plaza*, there is an extra layer of complexity in the HT. By translating the Portuguese *praça* into the Spanish loanword *plaza*, the translator introduced a foreign element to the translation that was not present in the ST. What could at surface be interpreted as an excessive liberty taken by the translator can in fact be justified by the contextual awareness that the plot is set in Spain, and the target English readers would be familiar with the meaning of *plaza* in Spanish-speaking countries (e.g., as in 'Plaza Mayor').

(20)

Conc. ID	ST	HT	MT
PBMAA1 117	Oh! **senhora**! atalhou Leonardo-Pataca	Oh, **senhora**, interrupted Leonardo-Pataca	Oh! **Mrs**! interrupted Leonardo-Pataca
PMMC1 303	Tem a certeza, **Sr.** Paraza?	Are you sure, **Senhor** Paraza?	Are you sure, **Mr.** Paraza?
PBPC1 262	Ficou mais algum tempo olhando a **praça**	He looked at the people in the **plaza** for a while	He spent some time looking at the **square**

(21)

Conc. ID	ST	HT	MT
PBAD2 113	**Nhonhô** quer alguma coisa?	Do you want anything, **massa**?	Does **Nhonhô** want anything?
PPLJ1 365	Quer a **mainata** já ali, com a bandeja, os copos	She wants the **maid** right away, with the tray, the glasses	He wants the **mainata** already there, with the tray, the glasses
PPEQ3 16	O Sr. Machado estava ontem em **D.** Maria	Senhor Machado was at the **Dona** Maria theatre yesterday	Mr. Machado was in **D.** Maria yesterday
PBJS1 786	**d.** Pedro explicou	**Dom** Pedro explained	**d.** Peter explained

In contrast, as shown in Example (21), the use of foreign words in the MT evidences a lack of contextual knowledge. The source-text word *Nhonhô* left untranslated in the MT is a dated form of address used by slaves to their masters. It is unlikely that target English readers would have the background knowledge needed to understand it. In the MT concordance shown in Example (21), *Nhonhô* could even be mistakenly understood to be someone's name. The solution found in the HT was to use the correspondingly dated English form of address 'massa' ('master') instead. Similarly, *mainata*, a Mozambiquean Portuguese word for 'maid' unlikely to be understood by target English readers, is kept as *mainata* in the MT output, but helpfully translated as 'maid' by the translator. The last concordance in Example (21) shows the source-text abbreviated form of address *D.* (pronounced *dona*) kept as *D.* in the MT. In the HT, the translator filled in the contextual knowledge gap by expanding the abbreviation to the full form *Dona*, which is again similar in Spanish and Italian, and thus more likely to be understood by English readers. The addition of 'theatre' to that same concordance without a corresponding equivalent in the ST shows further evidence of how the translator deliberately clarified that *D. Maria* is a theatre, given the contextual awareness that target readers would probably be unfamiliar with the Lisbon cultural scene.

Examples (20) and (21) demonstrate that professional human translators tend to take their readership and the wider context of the translation into account when choosing whether to employ foreign words in a translation. If foreign words are used, they tend to be used deliberately, to confer a foreign flavour to the translation, without compromising reader comprehension. In contrast, the few times foreign words are left untranslated in the MT occur in the rendition of more

obscure Portuguese source-text words in contexts likely to compromise target-reader understanding.

5. Discussion and conclusion

MT research has recently acknowledged the need to address more than just sentences in isolation, focusing on perfecting supra-sentential links and ensuring document-level consistency. This study was motivated by the need to explore discourse produced by HT and MT beyond the text, particularly with respect to how translation texts are shaped by the contexts in which they occur. Using keyword analysis – a methodology used in corpus-driven linguistics to identify contrastive lexical distributions in two different sets of textual data – this study singled out words that are exceptionally frequent in a corpus of professional translation in a known language direction (the HT corpus), and words that are exceptionally frequent in a parallel corpus of generic neural machine translation (the MT corpus).

The fact that the keyword analysis highlights divergent distributions of pronouns, modals, and proper names indicates that the method is sensitive to known challenges in MT discourse research. Pronoun resolution is the most prominent topic of recent work on discourse and MT referred to in Section 1 and related work (e.g., Bawden 2016; Guillou 2016; Luong and Popescu-Belis 2016). The greater use of gender-marked personal pronouns in the HT investigated in this study is in line with the well-known fact that it is problematic for MT to disambiguate gender when it is not expressed in the source language. The keyword analysis also highlights the distinctive use of possessives in HT, in the same way as Luong et al. (2017) discuss problems with possessives in Spanish–English MT, a similar language pair to the one investigated in this study. Modals, too, are known to be problematic in natural language processing (Morante and Sporleder 2012), and have been acknowledged as one of the challenges of MT research (Nakov 2016). Not surprisingly, the keyword analysis shows that modals are used to a much greater extent in the HT corpus. Lexical consistency (also referred to as term consistency and lexical cohesion in the MT literature) is yet another widely discussed issue in MT discourse research, as MT systems operating at sentence level may output translations that are lexically inconsistent at document level (Carpuat and Simard 2012; Guillou 2013). Lexical consistency can be particularly problematic in neural MT where no phrase tables are used to maintain fixed translations (e.g., Dougal and Lonsdale 2020). The same problem was detected in the keyword analysis, which highlights how proper names are translated inconsistently in MT.

The results of the keyword analysis do not just confirm known problems of MT, however. They also draw attention to further differences between MT and professional translation that deserve closer scrutiny. For example, indefinite pronouns are markedly more salient in the HT corpus, but do not seem to have received much attention yet in MT research. While it was not possible to investigate in more detail all the keywords in the study, a qualitative analysis aiming to gain further insight into discoursal differences between HT and MT was undertaken through close reading of the ST–HT–MT concordances of a sample of the HT and MT keywords. For grammatical keywords, the more fine-grained analysis focused on contrasting selected pronouns, modals, and prepositions. For lexical keywords, the focus was on examining spellings, proper names, and foreign words.

Although not all keyword differences necessarily impact discourse, the overall findings of the study indicate that the HT–MT discrepancies observed are less often about MT error and more about professional translators using an array of oblique translation strategies to enhance target-reader experience. Going beyond strictly linguistic equivalence, HT outperformed MT with regard to what House (2006) refers to as recontextualisation: professional translations are less ambiguous, more idiomatic, more appropriate to the situational context of the source text, and deliberately adapted to target-text readerships.

Reduction in ambiguity – or explicitation (Blum-Kulka 1986; Frankenberg-Garcia 2009) – is attested in the parallel concordances where modals, pronouns, and proper names are inserted in the HT without a corresponding prompt in the ST. In such cases, the clue as to which modal, pronoun, or name to add often lies not in the text itself (co-text) but rather in its interpretation (context). For instance, *ainda pisava* in Example (1) is ambiguous in Portuguese, as devoid of context it could be rendered either as 'was still stepping' or as 'might step'. Contextual knowledge of what happens if one steps on the object of the verb – in this case, a landmine – enables human translators to decide which of the two translations is appropriate. Although the study did not focus on MT problems of word-sense disambiguation, they are evident in Example (11), where the translator inferred from the context that *tios* is equivalent to 'aunts and uncles' rather than just 'uncles'; in Example (16), where the translator inferred that *experiência* should be translated as 'experiment' and not 'experience', and in Example (18), where 'Mesquita' is a proper name and not a 'mosque'.

Gains in idiomaticity are observed in parallel concordances where the MT outputs a more literal translation whereas in the HT the translators prefer oblique or indirect translation strategies so as to not upset target-language grammar or style. As shown in many of the examples in the previous section, this is achieved partly through transposition (a change in word class), partly through modulation

(a change in perspective), and partly through reformulation (a complete rewriting) (Vinay and Darbelnet [1958] 2004). For example, using the modal 'might' instead of the adverb 'maybe' in Example (2) is an example of transposition, translating *um filho sem mãe* as 'a motherless child' instead of 'a son without a mother' in Example (8) is an example of modulation, and translating *dar-lhes a comida* as 'fix their meals' instead of 'give them food' in Example (12) is an example of reformulation. Interestingly, the literal translations in all three MT equivalents are not errors, but employ words that appear to be distinctive or overused in MT – 'maybe', 'without', and 'give' (see Table 2).

Reader experience is also enhanced in the HT in places where translators do a better job of conveying register (i.e., combinations of linguistic features that reflect the situation in which language is used [Halliday 1978]). It is clear that at times the degree of formality conveyed in the MT does not fit in with the context of the narrative. For instance, in Example (3) a father would probably not tell his son to 'extinguish the lamp', but rather to 'put out the lamp'. Phrasal verbs like 'put out' are more typical of an informal register than non-phrasal synonyms like 'extinguish'. At the same time, informal words like 'yeah' and 'guy' appear to be overused in the MT (see Table 2), which would suggest an incorrect gauging of the level of formality in different situations of language use. In other places, the translators deliberately use foreign spellings and words to convey the foreign setting of the narrative, like using the Spanish loanword *plaza* to refer to a square in a novel set in Spain in Example (20), or the Latin spelling of the name 'Proserpinus' in a novel set in ancient Rome (Table 3), whereas the MT cannot discern situations where it might be appropriate to borrow words from other languages.

A fourth and final way in which reader experience is heightened in the HT is through translator awareness of possible communication breakdowns among the target readership. This can be seen in the ways translators deliberate which borrowings from other languages are safe to use, and how certain meanings need to be added to fill gaps in target-readership knowledge, in what Pym (2015) refers to as risk-averse translation strategies. For instance, in Example (21) the abbreviated form of address *d.* is machine translated as *d.*, but expanded to *Dona* or *Dom* in the HT. This not only helps English readers understand what the cryptic abbreviation *d.* means (note that the expanded Portuguese forms are similar in Spanish and Italian), but also enhances the foreign register of the narrative, since these words are not normally used in English-speaking contexts. Additionally, the expanded forms disambiguate between the feminine *dona* and the masculine *dom*, thanks to the contextual knowledge that the names that follow are respectively typically female and male. Awareness of what target readers might find difficult to decode is also captured by chance in the HT concordance with *Dona*, where it can be seen that the translator added the word 'theatre' to spell out

that *Dona Maria* is a theatre to readers not familiar with the Lisbon cultural scene. Another recontextualisation shift captured in the example concordances is the HT translation 'every man and woman to do their duty' instead of the more literal MT rendition 'each one to fulfill his duty' in Example (12), which shows how the translator deliberately avoided gendered language. And clearly, even the spelling differences detected, like the choice between British and American spellings, provide evidence of target-readership awareness.

Although there is no room for further details in this article, there are hundreds of concordances in the analysis providing evidence that many of the differences between HT and MT arise because, unlike MT, translators can adapt word choice according to different communicative demands and circumstances of language use. In contrast, opportunistic MT training data that does not distinguish between source and target language cannot distinguish source- and target-language readerships. Another issue is that when MT quality is evaluated devoid of text-external context, the evaluation cannot discriminate between solutions that work well in certain situations but not in others, such as when to use formal and informal target-language equivalents. Moreover, it is important to recognise that while experienced translators are attuned to contextual aspects of discourse, such as register and knowledge gaps among target-language readers, and are trained to employ oblique translation strategies when required, non-professionals tend to approach the task in terms of linguistic equivalence only (Tirkkonen-Condit 1990). MT training data from non-professional translations may therefore be less suitable for capturing strategies used by professionals to mediate discourse.

The corpus-driven keyword analysis undertaken in this study thus not only highlights known problems in MT, but also identifies further challenges for MT discourse research to address. Going beyond document-level consistency, there is a clear call for more studies on how MT discourse can tackle register and the variable situational contexts in which source texts are produced. Although customised MT can go a long way towards addressing some of these issues, the availability of controlled, quality training data is limited. Therefore, one question for the future is whether generic MT of the type used in this study can be trained to infer source-text context and adapt translation output accordingly. Another question is whether MT can be trained to recognise document-level register variation, such as an informal dialogue or quotation within a more formal narrative. Apart from acknowledging source-text context, this study calls for more research into addressing recontextualisation strategies typical of professional translations which take target-reader world knowledge into account.

In addition to providing insights for further MT discourse research, it is hoped the general findings of this study and the specific concordance examples given can be useful to translator education and post-editing training.

Finally, it is important to recognise that the scope of the study is limited, not only because it is exploratory and there is no room to analyse all the MT and HT keywords highlighted in detail, but also because it used only one MT engine and one language pair. Notwithstanding these limitations, this study suggests that corpus-driven keyword analysis can be a promising tool in MT discourse research, as it can not only point to known problems such as pronoun resolution and co-reference, but also unveil new insights about contextual aspects of translated discourse deserving further investigation.

Funding

This research was partly funded by the University of Surrey's Centre for Translation Expanding Excellence in England (E3) Fund, Research England, UKRI.

Acknowledgements

I would like to thank the anonymous reviewers and the editor of this journal for their very helpful comments on a previous version of this manuscript.

References

Bawden, Rachel. 2016. "Cross-lingual Pronoun Prediction with Linguistically Informed Features." In *Proceedings of the First Conference on Machine Translation, Berlin, Germany, 11–12 August*, 564–570. Stroudsburg: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2348

Blum-Kulka, Shoshana. 1986. "Shifts of Cohesion and Coherence in Translation." In *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, edited by Juliane House and Shoshana Blum-Kulka, 17–35. Tübingen: Gunter Narr.

Carpuat, Marine, and Michel Simard. 2012. "The Trouble with SMT Consistency." In *Proceedings of the Seventh Workshop on Statistical Machine Translation, Montréal, Canada, 7–8 June*, edited by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia, 442–449. Stroudsburg: Association for Computational Linguistics. https://doi.org/10.5555/2393015.2393077

Catford, John C. 1965. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. Oxford: Oxford University Press.

COMPARA. 2010. (Version 13.1.17.) Accessed April 12, 2019. http://www.linguateca.pt/COMPARA/index.php

De Beaugrande, Robert, and Wolfgang Dressler. 1981. *Introduction to Text Linguistics*. London: Longman. https://doi.org/10.4324/9781315835839

Dougal, Duane K., and Deryle Lonsdale. 2020. "Improving NMT Quality Using Terminology Injection." In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation, Marseille, France, 11–16 May*, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 4820–4827. Paris: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.593.pdf

Frankenberg-Garcia, Ana. 2008. "'Suggesting Rather Special Facts': A Corpus-Based Study of Distinctive Lexical Distributions in Translated Texts." *Corpora* (3) 2: 195–211. https://doi.org/10.3366/E1749503208000154

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

Frankenberg-Garcia, Ana. 2016. "A Corpus Study of Loans in Translated and Non-Translated Texts." In *Corpus-Based Approaches to Translation and Interpreting: From Theory to Applications*, edited by Gloria Corpas Pastor and Miriam Seghiri, 19–42. Frankfurt: Peter Lang.

Frankenberg-Garcia, Ana, and Diana Santos. 2003. "Introducing COMPARA: The Portuguese–English Parallel Corpus." In *Corpora in Translator Education*, edited by Federico Zanettin, Silvia Bernardini, and Dominic Stewart, 71–87. Manchester: St. Jerome.

Google Translator Toolkit (2019). Accessed December 1, 2019. https://translate.google.com/toolkit

Guillou, Liane. 2013. "Analysing Lexical Consistency in Translation." In *Proceedings of the Workshop on Discourse in Machine Translation, Soa, Bulgaria, 9 August*, edited by Bonnie Webber, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann, 10–18. Stroudsburg: Association for Computational Linguistics. https://www.aclweb.org/anthology/W13-3302.pdf

Guillou, Liane. 2016. *Incorporating Pronoun Function into Statistical Machine Translation*. PhD diss. University of Edinburgh.

Guillou, Liane, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. "A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018." In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Brussels, Belgium, 31 October – 1 November*, edited by Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, 570–577. Stroudsburg: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6435

Halliday, M.A.K. 1978. *Language as a Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.

Hardmeier, Christian. 2014. *Discourse in Statistical Machine Translation*. PhD diss. Uppsala University.

House, Juliane. 2006. "Text and Context in Translation." *Journal of Pragmatics* 38 (3): 338–358. https://doi.org/10.1016/j.pragma.2005.06.021

Kilgarriff, Adam. 2009. "Simple Maths for Keywords." In *Proceedings of Corpus Linguistics Conference*, Liverpool, UK. http://ucrel.lancs.ac.uk/publications/cl2009/

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

Klaudy, Kinga. 2009. "The Asymmetry Hypothesis in Translation Research." In *Translators and Their Readers: In Homage to Eugene A. Nida*, edited by Rodica Dimitriu and Miriam Shlesinger, 283–303. Brussels: Les Editions du Hazard.

Klaudy, Kinga. 2017. "Linguistic and Cultural Asymmetry in Translation from and into Minor Languages." *Cadernos de Literatura em Tradução*, 17, 22–37. https://doi.org/10.11606/issn.2359-5388.v0i17p22-37

Koehn, Philipp. 2005. "Europarl: A Parallel Corpus for Statistical Machine Translation." In *Proceedings of the Tenth Machine Translation Summit, Phuket, Thailand, 12–16 September*, 79–86. Tokyo: Asia-Pacific Association for Machine Translation. https://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf

Koehn, Philipp, and Josh Schroeder. 2007. "Experiments in Domain Adaptation for Statistical Machine Translation." In *Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June*, 224–227. Stroudsburg: Association for Computational Linguistics. https://doi.org/10.3115/1626355.1626388

Lapshinova-Koltunski, Ekaterina, and Christian Hardmeier. 2017. "Discovery of Discourse-Related Language Contrasts through Alignment Discrepancies in English–German Translation." In *Proceedings of the Third Workshop on Discourse and Machine Translation, Copenhagen, Denmark, 8 September*, edited by Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann, 73–81. https://doi.org/10.18653/v1/W17-4810

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. "Has Machine Translation Achieved Human Parity? A Case for Document-Level Evaluation." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October – 4 November*, edited by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, 4791–4796. Stroudsburg: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1512

Luong, Ngoc-Quang, and Andrei Popescu-Belis. 2016. "A Contextual Language Model to Improve Machine Translation of Pronouns by Re-ranking Translation Hypotheses." In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation, Riga, Latvia*, special issue of *Baltic Journal of Modern Computing* 4 (2): 292–304.

Luong, Ngoc-Quang, Andrei Popescu-Belis, Annette Rios Gonzales, and Don Tuggener. 2017. "Machine Translation of Spanish Personal and Possessive Pronouns Using Anaphora Probabilities." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol 2, Short Papers, Valencia, Spain, 3–7 April*, edited by Mirella Lapata, Phil Blunsom, and Alexander Koller, 631–636. Stroudsburg: Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2100

Morante, Roser, and Caroline Sporleder. 2012. "Modality and Negation: An Introduction to the Special Issue." *Computational Linguistics*, 38 (2): 223–260. https://doi.org/10.1162/COLI_a_00095

Nakov, Preslav. 2016. "Negation and Modality in Machine Translation." In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics, Osaka, Japan, 12 December*, edited by Eduardo Blanco, Roser Morante, and Roser Saurí, 41. Stroudsburg: Association for Computational Linguistics. https://www.aclweb.org/anthology/W16-5005.pdf

Popescu-Belis, Andrei, Sharid Loáiciga, Christian Hardmeier, and Deyi Xiong, eds. 2019. *Proceedings of the Fourth Workshop on Discourse in Machine Translation, Hong Kong, China, 3 November*. Stroudsburg: Association for Computational Linguistics. https://www.aclweb.org/anthology/volumes/D19-65/

Pym, Anthony. 2015. "Translating as Risk Management." *Journal of Pragmatics* 85: 67–80. https://doi.org/10.1016/j.pragma.2015.06.010

Schleiermacher, Friedrich. (1813) 2004. "On the Different Methods of Translating." In *The Translation Studies Reader*, 2nd ed., edited by Lawrence Venuti, 43–63. London: Routledge.

Tiedemann, Jörg. 2012. "Parallel Data, Tools and Interfaces in OPUS." In *Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2214–2218. Stroudsburg: Association for Computational Linguistics. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Tirkkonen-Condit, Sonja. 1990. "Professional vs. Non-Professional Translation: A Think-Aloud Protocol Study." In *Learning, Keeping and Using Language: Selected Papers from the Eighth World Congress of Applied Linguistics, Sydney, 16–21 August 1987*, edited by M.A.K. Halliday, John Gibbons, and Howard Nicholas, 381–394. Amsterdam: John Benjamins. https://doi.org/10.1075/z.lkul2.28tir

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6

Toral, Antonio, and Andy Way. 2018. "What Level of Quality Can Neural Machine Translation Attain on Literary Text?" In *Translation Quality Assessment: From Principles to Practice*, vol. 1, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 263–287. Cham: Springer. https://doi.org/10.1007/978-3-319-91241-7_12

Turovsky, Barak. 2016. "Found in Translation: More Accurate, Fluent Sentences in Google Translate." *Google* (blog), November 15, 2016. https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

Van Dijk, Teun A. 1977. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. Harlow: Longman.

Vinay, Jean-Paul, and Jean Darbelnet. (1958) 2004. "A Methodology for Translation." In *The Translation Studies Reader*, 2nd ed., edited by Lawrence Venuti, 128–137. London: Routledge.

Webber, Bonnie, Andrei Popescu-Belis, and Jörg Tiedemann, eds. 2017. *Proceedings of the Third Workshop on Discourse in Machine Translation, Copenhagen, Denmark, 8 September*. https://www.aclweb.org/anthology/W17-4800

Address for correspondence

Ana Frankenberg-Garcia
Centre for Translation Studies
University of Surrey
Stag Hill Campus
GU2 7HX GUILDFORD
United Kingdom
a.frankenberg-garcia@surrey.ac.uk
https://orcid.org/0000-0001-9623-7990

Publication history

Date received: 25 April 2020
Date accepted: 7 August 2021
Published online: 8 September 2021