DELA Corpus - A Document-Level Corpus Annotated with Context-Related Issues

Sheila Castilho
ADAPT Centre - School of Computing
Dublin City University
sheila.castilho@adaptcentre.ie

João L. Cavalheiro Camargo
Western Paraná State University
joao.camargo@unioeste.br

Miguel Menezes
University of Lisbon
lmenezes@campus.ul.pt

Andy Way
ADAPT Centre - School of Computing
Dublin City University
andy.way@adaptcentre.ie

Abstract

Recently, the Machine Translation (MT) community has become more interested in document-level evaluation especially in light of reactions to claims of "human parity", since examining the quality at the level of the document rather than at the sentence level allows for the assessment of suprasentential context, providing a more reliable evaluation. This paper presents a document-level corpus annotated in English with context-aware issues that arise when translating from English into Brazilian Portuguese, namely ellipsis, gender, lexical ambiguity, number, reference, and terminology, with six different domains. The corpus can be used as a challenge test set for evaluation and as a training/testing corpus for MT as well as for deep linguistic analysis of context issues. To the best of our knowledge, this is the first corpus of its kind.

1 Introduction

Machine translation (MT) is now widely used in a variety of fields, mainly due to advancements in neural models (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). As a result of these recent advances, scientists have been increasingly attempting to include discourse into neural machine translation (NMT) systems (Wang, 2019; Lopes et al., 2020). Thus, researchers started to consider a more suitable evaluation for these document-level systems as the standard MT automatic evaluation metrics have been shown to underestimate the quality of NMT systems (Shterionov et al., 2018) and the appropriateness of these metrics for document-level systems has been challenged (Smith, 2017) since they are not sensitive to their improvements (Voita et al., 2019).

Accordingly, document-level human evaluation of MT has attracted the community's attention since it allows for a more thorough examination of the output quality with context. While a few works have taken into account document-level human evaluation (Läubli et al., 2018; Toral et al., 2018; Barrault et al., 2019; Castilho, 2020, 2021), one common practice for document-level evaluation is the usage of test suites with context-aware markers (Bawden et al., 2018; Guillou et al., 2018; Müller et al., 2018; Voita et al., 2019; Cai and Xiong, 2020). However, the concept of document-level evaluation, in terms of how much text must be shown, remains uncertain (Castilho et al., 2020). While most research on document-level MT evaluation works with contrastive pairs, very few works have tried to use full documents for human evaluation (Läubli et al., 2018; Castilho, 2020, 2021) and challenge test sets (Rysová et al., 2019; Vojtěchová et al., 2019). Methodologies for assessing MT at the document-level have been looked into (Barrault et al., 2019, 2020) as well as the types of issues that come with different methodologies (Castilho, 2020, 2021).

We present a document-level corpus annotated with context-aware issues when translating from English (EN) into Brazilian-Portuguese (PT-BR). In total, 60 documents from six different domains (literary, subtitles, news, reviews, medical, and europarl) were annotated with context-aware issues, namely gender, number, ellipsis, reference, lexical ambiguity, and terminology. The corpus can be used as a challenge test set for the evaluation and as a training/testing corpus for MT and quality estimation, as well as for deep linguistic analysis of context issues. Moreover, we believe that the annotation can be also used for close-related languages such as Spanish.

2 Related Work

Document-level MT evaluation has attracted interest in the field as it allows for the evaluation of

suprasentential content, which in turn, provides more meaningful insights on the MT output. However, the definition of what constitutes a document-level MT evaluation is still unclear (Castilho et al., 2020).

Context plays an important role as it is widely used in translation and interpreting literature (Baker, 2006), although it lacks a precise definition for practical purposes, including in everyday work of a professional translator (Melby and Foster, 2010). For Melby and Foster (2010, p 3), context in translation could be studied "either for the purpose of analysing existing translations or for the purpose of improving the production of new translations". For the authors, context can be categorised into *non-text* (non-linguistic variables) and *text* (linguistic aspects), where the latter is divided into four aspects of context: relating to the source text: *co-text* (the version of the document itself) and *chron-text* (past and future versions); and relating to other text: *rel-text* (monolingual related texts) and *bi-tex* (bilingual related texts). In this work, we adopt Melby and Foster's view of context that is important to the analysis of translations, and focus (i) on the co-text, i.e. the boundaries within the document translated, and (ii) in the non-text, where the name of the authors, speakers, and products have an effect on the translation.

In a survey with native speakers, Castilho et al. (2020) tested the context span for the translation of 300 sentences in three different domains, namely reviews, subtitles, and literature. The results showed that over 33% of the sentences tested were found to require more context than the sentence itself to be translated or evaluated, and from those, 23% required more than two previous sentences to be properly evaluated. The authors found that ambiguity, terminology, and gender agreement were the most common issues to hinder translation. Moreover, differences in issues and context span were found between domains. Their recommendations include to show whole documents when possible, include information on text type, topic, product, hotel and movie names in case of reviews, and include visual context whenever possible (non-text). This shows that document-level evaluation enables the assessment of textual cohesion and coherence types of errors which are impossible at times to recognise at sentence level.

Regarding overall MT evaluation, a few attempts have been made to perform human evaluation with document-level set-ups. Läubli et al. (2018) compared sentence-level evaluation versus document-level evaluation with pairwise rankings of fluency and adequacy to evaluate the quality of MT against human translation (HT) with professional translators. Their results show that document-level raters clearly preferred HT over MT, especially in terms of fluency. The authors argue that document-level evaluation enables the identification of certain types of errors, such as ambiguous words, or errors related to textual cohesion and coherence.

The Conference for Machine Translation (WMT), which has been running since 2006 and only evaluated sentences, attempted document-level human evaluation for the news domain for the first time in 2019 (Barrault et al., 2019). Their direct assessment (DA) (Graham et al., 2016) required crowdworkers to assign a score (0-100) to each sentence. They asked raters to evaluate (i) whole texts, (ii) single consecutive segments in their original order, and (iii) single random phrases. In the following year, WMT20 changed the approach and expanded the context span to include full papers, requiring raters to evaluate specific segments while seeing the complete document, as well as to assess the content's translation (Barrault et al., 2020).

Castilho (2020, 2021) tested for the differences in inter-annotator agreement (IAA) between single sentence and document-level set-ups. In Castilho (2020), the author asked translators to evaluate the MT output of freely available online systems in terms of fluency, adequacy (Likert scale), ranking and error annotation in two different set-ups: (i) translators give one score per single isolated sentence, and (ii) translators give one score per document. The results showed that IAA scores for the document-level set-up reached negative levels, and the level of satisfaction of translators with that methodology was also very low. Nonetheless, it avoided cases of misevaluation that happen in isolated single sentences. Following on from that work, Castilho (2021) modifies the document-level set-up and re-runs the experiment with more translators, where she compares the IAA in evaluation of (i) random single sentences, (ii) evaluation of individual sentences where translators have access to the full source and MT output, and (iii) evaluation of full documents. Results showed that a methodology where translators assess individual sentences within the context of a document yields a good level

of IAA compared to the random single-sentence methodology, while a methodology where translators give one score per document shows a very low level of IAA. The author demonstrates that the methodology of assigning one score per sentence in context avoids misevaluation cases which are extremely common in the random sentences-based evaluation set-ups. Moreover, the author posits that the higher IAA agreement in the random single sentence set-up is because "raters tend to accept the translation when adequacy is ambiguous but the translation is correct, especially if it is fluent" (Castilho, 2021, p 42), and asserts that the single random sentence evaluation method should be avoided as the misevaluation issue is especially problematic when assessing the quality of NMT systems as they have an improved fluency level.

One current way of evaluating document-level issues is the use of test suites designed to better evaluate translation of the addressed discourse-level phenomena. Commonly, these test suites are contrastive, that is, each sample sentence in the test has both correct and wrong translations for a given phenomena (Bawden et al., 2018; Guillou et al., 2018; Müller et al., 2018; Voita et al., 2019; Cai and Xiong, 2020). The higher the accuracy of the model in rating correct translations over incorrect ones, the better the quality is deemed to be. Test suites with document-level boundaries are still scarce, e.g. Vojtěchová et al. (2019) present a test suite designed to evaluate coherence when testing MT models trained for the news domain on audit reports, and Rysová et al. (2019) designed a document-level test suite to assess three document-level discourse phenomena, namely information structure, discourse connectives, and alternative lexicalisation of connectives.

Given the above, the need to move toward document-level methodologies in MT is indisputable. Moreover, with the lack of resources for the topic of document-level MT, the document-level corpus annotated with context-aware issues presented here can be used as a challenge test set for evaluation and as a training/testing corpus for MT as well as for deep linguistic analysis of context issues.

3 Corpus Compilation

The corpus was collected from a variety of freely available sources. Following a pre-determined list of context issues found in Castilho et al. (2020) that hindered the translation of single sentences and sentence pairs, the annotators searched for challenging English texts for the MT systems when translating into PT-BR. In total, 60 full documents (57217 tokens) were collected from six different domains: literary, subtitles, news, reviews, medical, and legislation (europarl). Table 1 shows the statistics of the corpus.

Domains	#Docs	#Sent.	Av. Sent. Lgth
Subtitles	9	1074	18.69
Literary	4	756	9.76
News	15	634	17.17
Reviews	28	608	13.42
Medical	3	339	13.02
Legislation	1	272	23.70
TOTAL	60	3683	15.57

Table 1: Full corpus statistics, where *average sentence length* is calculated as words per sentence.

Each domain has their plain text and .xls versions of the documents segmented into sentences with sentence id and document boundary tags, and all documents contain the source (url or corpus) where the documents were retrieved from. What follows is a detailed description of each domain is provided.[1]

3.1 Subtitles

To compile the corpus for the subtitle domain, nine full TED Talks were selected from the Opus Corpus (Tiedemann, 2012) from a variety of different topics and speakers, where: doc1: education, doc2: climate change, doc3: astronomy, doc4: computers, doc5: creativity, doc6: science, doc7: technology, doc8: anthropology, and doc9: psychology. We chose these talks specifically in order to obtain a blend of different topics and speakers' genders.

	#Sent.	#Tokens	Av. Sent. Lgth
doc1	105	1671	15.91
doc2	98	1309	13.35
doc3	40	650	16.66
doc4	71	1213	17.08
doc5	176	3654	20.88
doc6	130	2485	19.26
doc7	77	1384	18.45
doc8	167	4213	25.53
doc9	210	3346	16.00
TOTAL	1074	19925	18.55

Table 2: Corpus statistics for each document in the subtitle domain.

[1] Although some of the documents were already segmented by sentence (i.e. Opus and WMT), the full corpus was manually checked for sentence segmentation.

3.2 Literary

	#Sent.	#Tokens	Av. Sent. Lgth
doc1	205	2921	14.24
doc2	122	2002	16.40
doc3	76	689	9.06
doc4	353	1767	5.00
TOTAL	756	7379	9.76

Table 3: Corpus statistics for each document in the Literary domain.

To compile the corpus for the literature domain, four documents[2] were selected:
doc1: one chapter from a fan-fiction story.[3]
doc2: one excerpt from "The Road to Oz" book.[4]
doc3: a short story generated with the PlotGenerator website.[5]
doc4: a short play generated with the PlotGenerator website.

Note that a blend of contemporary and classic excerpts, combining descriptive and fast moving styles, were gathered. Note too that the synthetic stories (doc3 and doc4) were generated as they allowed the researchers to add a good number of possible issues, including lexical ambiguities cases that can only be solved with a larger context than two consecutive sentences which is rather difficult to find in "natural" texts. Nonetheless, English native speakers then revised both stories for fluency and readability. Table 3 shows the statistics for each document in the literary domain.[6]

3.3 News

The news domain was compiled with 15 documents gathered from different sources. Table 4 shows the statistics of the corpus.[7]

Five documents were gathered from the WMT series (four documents from WMT19[8] and one from WMT20[9]), and their size varied from 13 to 32 sentences. Ten documents were gathered from several news websites,[10] and they varied from 23-35 sentences.

	#Sent.	#Tokens	Av. Sent. Lgth
docs 1-5	112	2293	20.47
docs 6-15	521	8578	16.46
TOTAL	633	10871	17.17

Table 4: Corpus statistics for each document in the news domain.

3.4 Reviews

The reviews domain was compiled with 28 documents gathered from reviews available on Amazon[11] and TripAdvisor[12] websites. Table 5 shows the statistics of the corpus.[13]

Reviews gathered from Amazon consist of users' reviews about a variety of products and movies, totalling 25 reviews, and vary from 6 to 84 sentences. The reviews were sought by searching products that could generate lexical ambiguities, such as "plant", "ship", etc. Reviews gathered from TripAdvisor consist of 3 reviews about places, and vary from 23-35 sentences.

	#Sent.	#Tokens	Av. Sent. Lgth
docs 1-25	520	6901	13.27
docs 26-28	88	1261	14.32
TOTAL	608	8162	13.42

Table 5: Corpus statistics for each document in the review domain. Documents 1-25 are product reviews gathered on the Amazon website, and documents 26-28 are location reviews gathered on the TripAdvisor website.

3.5 Medical

The medical domain corpus was compiled with three full documents, where two of them were collected from Autopsy reports available on the Medical Transcriptions website,[14] and one document was collected from the leaflets available on the

[2] Excerpts of two copyrighted books are in the process of being granted permission, and if so, they will be added to the corpus.

[3] "Harmonic Resonances" (based on the Carrie film) fan fiction (archiveofourown.org/works/26524723/chapters/64650841), last accessed 01 June 2021.

[4] Chapter 3 "Queer Village" (www.gutenberg.org/files/485/485-h/485-h.htm#chap03), last accessed 21 June 2021.

[5] https://www.plot-generator.org.uk/ last accessed 21 June 2021.

[6] Note that the Av. Sent. Lgth for the literature domain is skewed because of doc4, which – due to its play format where the names of each character is given in a single line before they speak – contains a great number of very short sentences. The Av. Sent. Lgth for literature when doc4 is left out is 13.9.

[7] Note that for the news domain, we grouped documents due to space constraints.

[8] http://www.statmt.org/wmt19/

[9] http://www.statmt.org/wmt20/

[10] mercurynews.com, zdnet.com, usmagazine.com, machinedesign.com, nytimes.com, thejournal.ie, thesun.ie, theconversation.com, goodhousekeeping.com, allthatsinteresting.com, last accessed 01 June 2021.

[11] amazon.com

[12] tripadvisor.com

[13] Note that for the review domain, we grouped documents due to space constraints.

[14] mtsamples.com

Royal College of Obstetricians and Gynaecologists (RCOG).[15]

	#Sent.	#Tokens	Av. Sent. Lgth
docs 1-2	243	2912	11.98
doc 3	96	1503	15.65
TOTAL	339	4415	13.02

Table 6: Corpus statistics for documents in the medical domain. Documents 1-2 were compiled from autopsy reports, while document 3 was compiled from medical leaflets

3.6 Legislation

For the legislation domain, we chose an excerpt of Europarl (Koehn, 2005)[16] taken from the Opus Corpus (Tiedemann, 2012).

	#Sent.	#Tokens	Av. Sent. Lgth
doc 1	272	6465	23.7

Table 7: Corpus statistics for documents in the legislation domain extracted from the Europarl corpus.

4 Methodology for Annotation

Following literature on document-level test suites (see Section 2), together with issues found when trying to define how much context span is needed to translate and evaluate MT (Castilho et al., 2020), we compiled a list of context-aware issues that are challenging for MT when translating from EN into PT-BR to be annotated:

1- Gender	2- Number
3- Ellipsis	4- Reference
5- Lexical Ambiguity	6- Terminology

Three annotators tagged those issues that might occur in a translation from EN into PT-BR when no context information is given. For example, in the following single sentence given to a translator to translate:

"And thanks for the case."

The translator will not be able to translate this sentence with absolute certainty because:
i) it is not possible to know the gender of the person who is saying 'thanks' as Portuguese differentiates between masculine and feminine genders.
ii) it is not possible to know what the word "case" is as this word has a few different meanings that would fit this sentence, i.e it could be some type of protective box (a case for my phone, a case for my glasses), a woman's bag, a pencil case, a folder, a suitcase, or a police case to be investigated, each one with a different translation in Portuguese. Consequently, the translation of "for" will have a different gender depending on the meaning of the word "case".

When evaluating the translation of the source sentence given by 3 different MT systems (Google Translate[17] (GG), Microsoft Bing[18] (MS) and DeepL[19] (DPL) the translator has to evaluate all three systems' outputs as correct:
GG: "E obrigado pelo caso." (masculine, police case)
MS: "E obrigado pelo estojo." (masculine, pencil case)
DPL: "E obrigado pela caixa." (masculine, box)

That is because without a wider context, it is impossible to know the correct translation or the sentence, which should be:

HT: *"E obrigada pela capa."* (feminine, phone case)

Therefore, the issues tagged in the corpus are issues that might arise in the translation of sentences when the full context is not given. Annotators used different MT systems to help check for issues that would go unnoticed when only looking at the source text.

Moreover, a few modifications to the source text were performed in order to add those issues and make the translation more challenging for MT, such as modifying the gender, substituting the name of a product for 'it', splitting a sentence into two, etc. These modifications are explained in the spreadsheet file for each line modified, so researchers can decide if they can use or not documents that had the source modified.

4.1 Annotation of Context-Related Issues

As previously mentioned, six context-related issues were tagged in the corpus when they could not be solved within the sentence they appeared. A detailed guideline was developed as the annotators gathered the corpus and discussed how the annotation would be better performed. Figure 1 shows the decision tree that guides the annotation of the context-related issues.

[15] Copyright permission was granted by both websites.
[16] http://www.statmt.org/europarl/
[17] https://translate.google.com/
[18] https://www.bing.com/translator
[19] https://www.deepl.com/en/translator

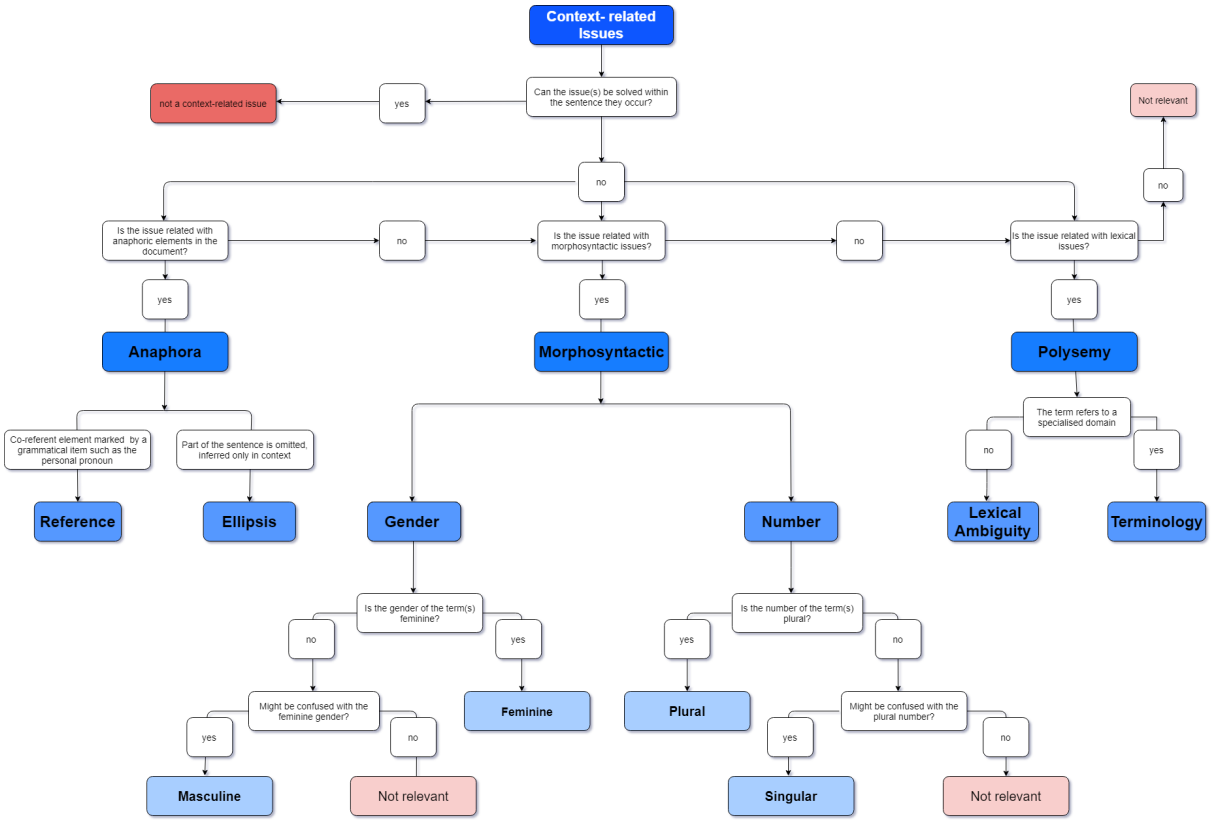

Figure 1: Decision tree used to guide the annotation of context-related issues.

4.1.1 Reference

Reference is associated with the notion that "anaphoric references and noun phrase organizers may serve as cohesive ties linking separate sentences into unified paragraphs [aiding] the reader's memory structure" (Fishman, 1978, p 159). Differently from ellipsis which is generally dependent on the previous clause, reference can reach back a long way in the text and extend over a long passage (Halliday and Matthiessen, 2013), thus being of significance for the present work.

In the annotation guide, we annotated the reference whenever we faced a disruption or ambiguity in the referential chain, e.g., we only annotated dependent referential units. Moreover, and similar to all annotated categories, the disagreement had to be expressed at the document level, e.g. the issue could not be solved only by looking at the sentence. In example **A)**, we annotate the second individual *it* as being a referential issue because there is not enough lexical material in the sentence to properly establish the referent, thus affecting translation correctness and final text readability.

A) *It is understandable though since* ***it*** *was shipped from China.*
reference –> it = the ship
it = o navio.

In example **B)**, we annotated *they* as being a referential unit issue, due to the fact that there is not enough lexical material in the sentence to determine its referent. Moreover, we also tagged this issue as a gender problem since there is no information in the source sentence that allows one to determine that the referential unit should be translated into PT-BR as a plural feminine pronoun.

B) *They actually hurt*
reference –> they = the choices
gender –> they = feminine
They actually hurt = Elas / As escolhas realmente machucam.

4.1.2 Ellipsis

Ellipsis is a form of anaphoric cohesion where there is an omission from a clause, and, so, the reader must "presuppose something by means of what is left out" (Halliday and Matthiessen, 2013, p 635). Ellipsis differs from reference as the relationship it entails is lexicogrammatical rather than semantic

(Halliday and Matthiessen, 2013).

In the annotation guide, we annotate ellipsis exclusively when the omission of information affects the translation of that specific single sentence which needs a broader context to be understood. For example, in **C)**, ellipsis is tagged because the omission with the explicit indication of "do" causes lexical ambiguity that cannot be solved within the sentence.[20] Therefore, the tags and the solution for the issues are both ellipsis and lexical ambiguity, with the translation of the ellipsis also containing an explanation of the lexical ambiguity caused by it.

C) *In my laughter, I bellied out a "YES, I do!!"*
ellipsis –> do = think
lexical ambiguity –> do = make (incorrect) *vs* think (correct)
Sim, eu faço! (Yes, I make, incorrect) *vs*
Sim, eu **acho!** (Yes, I "think", correct)

In example **D)**, ellipsis is tagged because the omission causes gender issues that cannot be solved within the sentence. Therefore, the tags and the solution for the issues are both ellipsis and gender, with the translation of the ellipsis also containing an explanation of the lexical ambiguity caused by the ellipsis:

D) *Several more are planned for the rest of the year, including The Angry Beavers Seasons 1 & 2, Danny Phantom Season 1, Aaahh!! Real Monsters Season 1, Catdog Season 1 Part 1 and The Wild Thornberrys Season 2 Part 1.*

ellipsis –> several more = releases
gender –> several more are planned = feminine
Several more are planned... = Vári**as** outr**as** estão planejad**as**..."

Sentence **E)** is an example of ellipsis with the auxiliary *do* that has not been tagged in the corpus because the omission is solved within the sentence:

E) *Also, not once did I **feel a blast of hot** air like I **do** when taking things out of the oven.*
do = feel a blast of hot air

[20]It is only with the previous sentence "*He came back in the house and said 'So you think this is funny?!' up the stairway at me and I LOST IT*" that we can solve "I do" as being "I think so".

4.1.3 Gender

As Portuguese is a language in which grammatical gender (feminine and masculine) plays a very significant role, the definition of gender used is from a grammatical standpoint, where word classes like adjectives, articles or pronouns are bound to respect and reflect a word's gender (Grosjean et al., 1994).

In the annotation guide, we annotated gender whenever facing a gender issue e.g., gender disagreement, unsolvable within the sentence itself and requiring broader context information.[21] For example, in example **F)**, gender (feminine) is tagged because the issue is not possible to be solved within the sentence. Since the default of the PT-BR is to have everything in the masculine, translations (both HT and MT) follow the same pattern. Therefore, we tag the word that needs to be in a different gender and the solution for its gender marker, with the translation containing an explanation:

F) *I'm surprised to see you back so early.*
gender –> surprised = feminine
surprised = surpresa

In example **G)**, we note that not only the pronoun "they" needs to be tagged with the feminine gender tag, but also the expression "staying at", as it is translated with an adjective in Portuguese:

G) *She waited for a few minutes longer, but nothing happened, no one followed her, so she made her way back to the motel they were staying at.*
gender –> they = feminine
gender –> staying at = feminine
they were staying at = elas estavam hospedadas

Gender was also tagged even when the most used translation for the the given term was a neutral one, because the adjective could still be translated with one of its synonyms. For instance, in example **H)**, the adjective "glad" has its most common translation as "feliz" which is used for both masculine and feminine gender. If a

[21]Note that gender was most exclusively tagged as *feminine* when a problem with the agreement was obvious. As the MT systems typically tend to translate the gender into the masculine form (when no specific gender markers are given in the single sentence) for PT-BR, the masculine gender was only tagged when there was an ambiguity in the sentence.

translator chooses to translate the text as "I'm glad = Estou feliz", no gender marking is needed. However, synonyms of that translation would need to be translated into feminine (satisfeita, grata, encantada, animada), and so, gender is tagged for that case:

H) *I'm so glad that it comes with the extender, so I have more levels to use to continue to get smaller.*
gender –> glad = feminine
reference –> it = waist cincher

4.1.4 Number

Number agreement is one of the basic coherence devices within a text, and it is "part of the code, used to signal that linguistic constituents carrying the same number are linked regardless of whether they appear together or apart in an utterance"(Bock et al., 1999, p 330), in the entirety of the text, and thus is significant for the present work.

In the annotation guide, we annotated number whenever we faced a number disagreement within the referential chain, e.g. (i) noun or pronoun, (ii) verb and noun/pronoun, (iii) adjective, caused by lack of enough contextual information within the sentence.[22] In example **I)**, the number category was applied to the word *you* because it is not possible to identify within this single sentence whether we are facing a pronoun in the plural or singular.

I) *I was praying for you.*
number –> you = plural
you –> vocês

Example **J)** depicts a mistranslated number agreement chain into PT-BR which originated from the absence of contextual evidence in the sentence that allowed us to determine whether *you* should be translated in the plural rather than in the singular. Furthermore, as a consequence of this initial mistranslation, the adjective *agreeable* was affected, being translated in the singular rather than the plural.

J) *You should be more agreeable.*
number –> you = plural
number –> agreeable = plural
number –> should be = plural
gender –> agreeable = feminine
You should be more agreeable. –> Você**s** dever**iam** ser mais simpátic**as/** agradáv**eis**.

4.1.5 Lexical Ambiguity

Lexical ambiguity refers to the fact that "a single word form can refer to more than one different concept" (Rodd, 2018, p 2). Lexical ambiguity can be divided into two categories: (i) one takes into account a word's morphological aspect (verbs, adjectives) referred to as syntactic ambiguity, e.g. the word "play" can be either the act of taking part in a sport or the conducting of a sporting match; and (ii) the second focuses on the fact that a word can assume different meanings according to context, e.g. the word "ball" as in *They danced till dawn at the ball* versus *This dog can be entertained all day with a ball* (Small et al., 2013, p 4), which is referred as semantic ambiguity.

In the annotation guide, we annotated lexical ambiguity, the more generic term, whenever we faced one of the two cases above ((i) and (ii)) and whenever they appeared to be detrimental to the translation and understandable only within the broader context, rather than at sentence level. In example **K)**, lexical ambiguity is tagged because the clause *I lost it*, without context, can be interpreted either as someone losing something or someone losing control:

K) *He came back in the house and said "So you think this is funny?!" up the stairway at me and I LOST IT.*

lexical ambiguity –> lose something *vs* to lose control
I lost it –> Eu o/a perdi *vs* Eu perdi o controle

In example **L)**, lexical ambiguity is tagged because the word *Period* is polysemic, meaning simultaneously menstruation, a portion of time, and a punctuation mark, and by the fact that there is not enough lexical information at a sentence level to disambiguate the complete meaning.

L) *Period.*
lexical ambiguity –> period = era/menstruation *vs* full stop
Período *vs* Ponto final

[22] Note that number was most exclusively tagged as *plural* for the pronoun "you" (and its referential chain (verb/adjectives)) when a problem with the agreement was obvious. As the MT systems typically tend to translate "you" in the singular (when no specific plural markers are given in the single sentence) for PT-BR, the pronoun was only tagged for singular when there was an ambiguity in the sentence.

4.1.6 Terminology

Terminology, according to Sager and Nkwenti-Azeh (1990) (as cited in (Kast-Aigner, 2018)), can have three different definitions: (i) the theory behind the relationship between a concept and a term; (ii) terminology curatorship activities, e.g. collection, description, processing and presentation of terms; and (iii) the vocabulary of a specific field. In the present work, we perceived terminology as (iii) i.e. the lexical expression of a domain-specific area.

In the annotation guide, we annotated terminology whenever we faced a wrongly domain-specific word translation caused by contextual poor sentences. In the following example **M)**, the category terminology was applied to the word 'farm' because its meaning shifts from "a piece of land used for crops and cattle raising", its more generalised conceptualisation, into a more domain-specific concept, "an area of land with a group of energy-producing windmills or wind turbines".

M) *The center will also conduct testing (power curve, mechanical loads, noise, and power quality) at its own experimental wind farm*
terminology –> generalised lexic (farm) *vs* domain-specific lexic (park)
wind farm –> parque eólico

4.2 Format

The annotation was performed for each sentence, which are tagged one per line, in the order they appear in the sentence, followed by their explanation/solution, along with modifications performed in the source (if any) and translations of some cases and notes. Sentences with no context-related issues are followed by two Xs for the issue and the solution. For Reference and Ellipsis, the term that contains the issue is stated along with an equals sign (=) and the explanation of what it refers to. For Gender and Number, the issue is tagged along with an equals sign (=) and the solution (feminine/masculine or singular/plural) is given. For Lexical Ambiguity and Terminology, the term (or terms) is stated along with an equals sign (=) and a contrasting solution is given, the wrong meaning(s) compared to (vs) the correct one. Table 8 illustrates how the annotation format is performed for each issue.

The corpus will be made freely available in two formats. One is a spreadsheet (.xls) containing the tagged corpus in all domains and full information. This .xls format will allow for filtering specific issues or sentences and enable users/researchers to see the rationale of the annotation. The corpus will be also available in plain text (.txt) format, containing the segment id, sentence, issue and explanation all in one line.[23] This format will allow for an automatic use of the corpus, for training or as a test suite. Figure 2 shows the .xls and .txt formats.

4.3 Agreement

As previously mentioned, three annotators compiled and annotated the corpus. Their backgrounds include linguistics, translation and computational linguistics. Throughout the process of compilation and annotation, the annotators worked closely together to discuss the corpus compilation and also what issues should be tagged. Disagreements were discussed and resolved, and then the annotation process would resume. This process helped to refine the list of issues as well as to develop and finalise the guidelines. The corpus annotation carried out by the three first annotators was corrected at the final stage in order to ensure that it follows the established version of the guidelines.

In order to reveal some possible weaknesses of the annotation guidelines and the decision tree, another expert annotator was involved at the final stage. The fourth annotator worked with 9% of the documents from the original collection, where at least one document of each domain was selected randomly. The annotation was done according to the guidelines and the decision tree used by the first three annotators (see Figure 1). During the annotation process, the annotator was given the guidelines, decision tree and was explained what the goal of the annotation was, but was not allowed to communicate with the other annotators. We then calculated inter-annotator agreement using Cohen's Kappa (Cohen, 1960) treating the first annotation (performed by the three annotators) as the gold standard.

Results show that the overall Kappa score was 0.61 meaning that, by using the guidelines and the decision tree on a first try, we could reach a substantial agreement (Landis and Koch, 1977). We note that the majority of disagreement cases are related to agreeing whether or not a sentence contains an issue to be annotated, while our gold

[23] Modifications and translation are not provided in this format.

Issue	Explanation (solution)	Translation & notes
Reference	it = support group	o grupo de suporte
Ellipsis	I do = I think	Eu acho
Gender	it = feminine	Ela
Number	surrender = plural	Entreguem-se
Lexical ambiguity	paper = news (wrong) vs research article (right)	O jornal vs O artigo
Terminology	wind farm = farm (wrong) vs park (right)	Fazenda eólica vs Parque eólico

Table 8: Annotation format for every context-related issue.

Example .xls format

ID	text = source	issue	explanation	modification	extra explanations and notes
<seg id="1">	Willa G. Tsarones	X	X	subsititution = William >> Willia changed gender to add gender issues to the translation	
<seg id="2">	Reviewed in the United States on November 24, 2017	X	X		
<seg id="3">	I really liked this waffle maker after receiving it.	X	X		
<seg id="4">	I felt it was designed well and is highly functional and easy to use.	reference gender gender	it = waffle maker it = feminine well designed = feminine		aqui o "designed" vira "desenhado", então deve-se aplicar no feminino para "máquina"
<seg id="5">	I make 2 waffles twice every weekend with this Cusinart WAF-F20 and have been doing so for the last 10 months with great success.	X	X		
<seg id="6">	I really liked it.	reference	it = waffle maker		
<seg id="7">	Then, one weekend after cleanup, I looked at the rear support post and stared for awhile and asked myself "What an I looking at?"	X	X		
<seg id="8">	I was dumbfound.	gender	dumbfound = femine		

Example .txt format

```
<seg id="1">    Willa G. Tsarones   X   X
<seg id="2">    Reviewed in the United States on November 24, 2017  X   X
<seg id="3">    I really liked this waffle maker after receiving it.    X   X
<seg id="4">    I felt it was designed well and is highly functional and easy to use.   reference   it= waffle maker  gender    it=feminine gender  well designed=feminine
<seg id="5">    I make 2 waffles twice every weekend with this Cusinart WAF-F20 and have been doing so for the last 10 months with great success.   X   X
<seg id="6">    I really liked it.  reference   it=waffle maker
<seg id="7">    Then, one weekend after cleanup, I looked at the rear support post and stared for awhile and asked myself "What an I looking at?"   X   X
<seg id="8">    I was dumbfound.    gender  dumbfound=femine
```

Figure 2: Example of one excerpt of the corpus in the .xls and .txt format

standard has 170 issues annotated in this portion of the corpus, the fourth annotator found 106 issues. After this IAA was calculated, we discussed the annotation produced by annotator 4 and revised the corpus.

5 Conclusion

We have presented a document-level corpus annotated with context-aware issues when translating from EN into PT-BR, namely gender, number, ellipsis, reference, lexical ambiguity, and terminology. This first version of the corpus contains 60 documents, with 3680 sentences, in six different domains: subtitles, literary, news, reviews, medical and legislation. To the best of our knowledge, this is the first corpus of its kind.[24]

With the rise in NMT quality and the claims of human parity, the need to move towards a more fine-grained evaluation involving the whole document is beyond question. Moreover, with the lack of resources for the document-level MT area, this document-level corpus can be used as a challenge test set for evaluation and as a training/testing corpus for MT as well as for deep linguistic analysis of context issues. We believe that the annotation can be also used for closely-related languages such as Spanish.

We intend to increase the corpus, adding more documents, domains and more context-aware issues. The full translation into PT-BR is ongoing, and we want to annotate it for other languages, starting with the Romance language family.

[24]The corpus and guidelines will be freely available at https://github.com/SheilaCastilho/DELA-Project

Acknowledgements

We would like to than Helena Moniz and Vera Cabarrão for the fruitful discussions and invaluable help. This project was funded by the Irish Research Council (GOIPD/2020/69). ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Dublin City University, is funded by the Science Foundation Ireland through the SFI Research Centres Programme (Grant 13/RC/2106_P2).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly

Learning to Align and Translate. In *Proceedings of ICLR*, San Diego, CA.

Mona Baker. 2006. Contextualization in translator-and interpreter-mediated events. *Journal of pragmatics*, 38(3):321–337.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (WMT 19)*, pages 1–61, Florence, Italy.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Kathryn Bock, Janet Nicol, and J Cooper Cutting. 1999. The ties that bind: Creating number agreement in speech. *Journal of Memory and Language*, 40(3):330–346.

Xinyi Cai and Deyi Xiong. 2020. A test suite for evaluating discourse phenomena in document-level neural machine translation. In *Proceedings of the Second International Workshop of Discourse Processing*, pages 13–17, Suzhou, China. Association for Computational Linguistics.

Sheila Castilho. 2020. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online. Association for Computational Linguistics.

Sheila Castilho. 2021. Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45, Online. Association for Computational Linguistics.

Sheila Castilho, Maja Popović, and Andy Way. 2020. On Context Span Needed for Machine Translation Evaluation. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Anne Stevens Fishman. 1978. The effect of anaphoric references and noun phrase organizers on paragraph comprehension. *Journal of Reading Behavior*, 10(2):159–170.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

François Grosjean, Jean-Yves Dommergues, Etienne Cornu, Delphine Guillelmon, and Carole Besson. 1994. The gender-marking effect in spoken word recognition. *Perception & Psychophysics*, 56(5):590–598.

Liane Guillou, 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*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

M.A.K. Halliday and C.M.I.M. Matthiessen. 2013. *Halliday's Introduction to Functional Grammar*. LSE International Studies. Taylor & Francis.

Judith Kast-Aigner. 2018. *A Corpus-Based Analysis of the Terminology of the European Union's Development Cooperation Policy: with the African, Caribbean and Pacific Group of States*. Peter Lang International Academic Publishers.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

António V Lopes, M Amin Farajian, Rachel Bawden, Michael Zhang, and André F T Martins. 2020. Document-level Neural MT: A Systematic Comparison. In *22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity?

A Case for Document-level Evaluation. In *Proceedings of EMNLP*, pages 4791–4796, Brussels, Belgium.

Alan Melby and Christopher Foster. 2010. Context in translation: Definition, access and teamwork. *The International Journal for Translation & Interpreting Research*, 2.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

RCOG. Managing premenstrual syndrome (pms) - patient information leaflet. https://www.rcog.org.uk/en/patients/patient-leaflets/managing-premenstrual-syndrome-pms/. London: March 2018, used with the permission of the Royal College of Obstetricians and Gynaecologists. Last acessed August 2021.

Jennifer Rodd. 2018. Lexical ambiguity. *Oxford handbook of psycholinguistics*, pages 120–144.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. 2019. A test suite and manual evaluation of document-level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy. Association for Computational Linguistics.

J.C. Sager and B. Nkwenti-Azeh. 1990. *A Practical Course in Terminology Processing*. J. Benjamins Publishing Company.

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'dowd, and Andy Way. 2018. Human versus Automatic Quality Evaluation of NMT and PBSMT. *Machine Translation*, 32(3):217–235.

Steven L Small, Garrison W Cottrell, and Michael K Tanenhaus. 2013. *Lexical Ambiguity Resolution: Perspective from Psycholinguistics, Neuropsychology and Artificial Intelligence*. Elsevier.

Karin Sim Smith. 2017. On Integrating Discourse in Machine Translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of NIPS*, pages 3104–3112, Montreal, Canada.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of WMT*, pages 113–123, Brussels, Belgium.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 5998–6008, Long Beach, CA.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. 2019. SAO WMT19 test suite: Machine translation of audit reports. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 481–493, Florence, Italy. Association for Computational Linguistics.

Longyue Wang. 2019. *Discourse-aware neural machine translation*. Ph.D. thesis, Ph. D. thesis, Dublin City University, Dublin, Ireland.