


Comparison of Segmentable Units as Indicators of Two Texts Being Parallel

Afonso Xavier Canosa

University of Santiago de Compostela, Galiza, Spain

canosarodrigues@gmail.com

 <https://orcid.org/0000-0002-8767-3640>

Abstract

A bitext produced from a Portuguese historical text and its English translation, Fernão Mendes Pinto's *Pilgrimage*, serves as a case study to describe the creation of a parallel corpus and investigate which linguistic and textual units are the best indicators of alignability. The process of building the corpus goes through preparation of transcriptions, annotation, segmentation and sentence alignment. Once the bitext is ready, the corpus is used to inquire which units appear as more relevant to predict that both texts are parallel. From the largest content units, those of chapters, to sentences, word types, tokens and characters, the latest, despite being the unit with less textual and linguistic significance, were found to be the best indicator of both texts being alignable.

2012 ACM Subject Classification Computing methodologies → Machine translation

Keywords and phrases parallel corpora, text alignment, bitexts

Digital Object Identifier 10.4230/OASIS.SLATE.2018.16

Category Short Paper

1 Introduction

Parallel corpora are increasingly important for the development and evaluation of machine translation and Natural Language Processing applications. Yet, parallel texts can serve more specific research purposes, such as careful examination and comparison of versions and translations in classical humanities research. This is the case of Tartaria, a parallel corpus created from the the section that describes Fernão Mendes Pinto's stay with the Tartars, comprising chapters 117-131 from the Portuguese first edition *Peregrinacam* (PT 1614)¹ and chapters 38-41 from its English version (EN 1653)². As translation was one of the reasons for misreadings of Pinto's report (e.g. Figure 1 the exotic term *bada* is translated as rhinoceros without any apparent motivation in the source), a parallel corpus allows researchers to detect those segments that they may be more interested in and focus on relevant sentences only, allowing for optimization of expensive and time-consuming translation tools. The expected result should output a table with two texts, a source and its translation, in such a disposition that enables an easy comparison of both (Figure 1). The process of creating this parallel corpus required a balance between machine and human-performance. One of the first tasks to solve was to find out if the English version is a direct translation of the Portuguese text, or, on the other hand, if the target only follows a narrative and offers an independent free version of the source. To answer this question without direct inspection of both texts, textual units

¹ <http://purl.pt/82>

² <http://purl.pt/16425>



© Afonso Xavier Canosa;

licensed under Creative Commons License CC-BY

7th Symposium on Languages, Applications and Technologies (SLATE 2018).

Editors: Pedro Rangel Henriques, José Paulo Leal, António Leitão, and Xavier Gómez Guinovart

Article No. 16; pp. 16:1–16:7



OpenAccess Series in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

200	<p>E auido entre ambos cōselho sobre esta noua, assentaraõ de mãdarem as embarcaçoẽs todas quatro a Huzanguee, & elles ambos com poucos dos seus irẽse por terra a Fanaugrem onde tinhão por nouas que el Rey estaua, o que logo se pões em effeito cõ parecer tambem dẽsta princesa, a qual lhes mãdou dar todas as caualgaduras que ouuerão mister para sy & para os seus, & oito badas para leuarem o seu fato.</p>	<p>When as they had consulted a while upon these news, they resolved to send their four vessels away to Usamguee, and themselves to travel by land to Fanaugrem, where they understood the King was. This deliberation taken they put incontinently into execution, & that by the advice of this Princess, who for that purpose caused them to be furnished with horses for themselves, & their people, as also with eight Rhinocerots for the transportation of their baggage.</p>
-----	---	---

■ **Figure 1** Example of alignment used to research the translation of the term *bada*.

can be evaluated in order to search for the one that shows a closer correspondence between both texts, hence serving as evidence for alignability of source and target in a relation of direct translation.

2 Parallel corpora and units considered in alignments

Bilingual and multilingual corpora are core resources for the training and evaluation of automate machine translation and natural language processing tools [4, 1]. A distinction can be made between corpora that represent direct translations and those that are only comparable, showing a similarity in content, yet not being a literal translation of each other [14, 12]. In a parallel corpus, texts are aligned so that a direct correspondence is made between text sequences from one language to the other. The final alignments show not only sentences with the same content, but also omissions (1:0), additions (0:1), and more complex correspondences when added or omitted text comes together with perfect matches [8]. The corpus can be further enriched through annotation and is usually encoded with standard schemas [9], though output formats vary depending on the final use [8].

It is in the field of automate alignment that the issue of which units represent an indicator of two texts being parallel appears as a relevant question. Two units are considered in common sentence alignment algorithms [8, 10]: sentence length [2] and word matches [5]. The use of both resulted in hybrid solutions [6, 11]. Even more elaborated models claiming to outperform previous hybrid methods [7], use the semantic similarity of sentences as a result of computing TF-IDF values (hence word-based) across each language to later align target and source based on sentence metrics (sentence again).

3 Tartaria parallel corpus

The process of building a parallel corpus of the Portuguese and English chapters related to the Tartars in Fernão Mendes Pinto’s travels was aimed not only at producing an output for comparative studies, but also for research on NLP tasks such as NERC. An experimental approach was considered to find a balance between automation and the need for a high quality product, only achieved with human validation.

3.1 Transcriptions

The first step was text transcription. Although the Portuguese version of Pinto’s travels has been partially transcribed and published online [3, 13], the chapters relevant to Tartaria are not included in public corpora. It was possible to find, though, a transcription of the whole 1614’s text from a publishing house. Even if the provided transcription had some important unreported errors (gaps of whole pages and typos), those chapters relevant for the Tartar corpus could be used without any modification other than small typos, validated against the facsimile of the DLNLP.

```

-<text>
  -<div type="chapter" n="38">

    <head>CHAP. XXXVIII</head>
  -<head>
    -<div2 type="page" n="149">
      A Tartar Commander enters with his Army into the Town of Quincay,
      the taking of it by the means of some of us Portugals.
    </div2>
  </head>
  -<div1>
    -<div2 type="page" n="149">
      WE had been now eight months and an half in this captivity, where

```

■ **Figure 2** Excerpt showing all elements and attributes required to annotate Tartaria corpus.

The English text was in-house transcribed following the edition available at the DLNLP. Despite the use of OCR to start, the whole process of transcription required careful manual revision. Hyphens and reformed words at the end of the line were discarded and a regular font format was used for the whole text, even if the original displays place names, demonyms and anthroponyms in italics. Missing characters were marked with square brackets. During the process of text alignment, once the corpus was already built, some words still needed correction, usually due to confusion of similar characters such as *s* and *f*.

At the end of this stage there were two text files containing raw transcriptions of the first editions.

3.2 Annotation

The corpus was annotated for main structural elements (Figure 2), showing chapters, pages (folios in the Portuguese edition) and divisions to distinguish chapter headings from main content.

Geographical named entities were annotated in each corpus using NERC tools and human-validated to produce a gold standard of annotated place-names.

At the end of this stage the corpus comprised two annotated documents showing chapter, title, main text and either folio or page components.

3.3 Text segmentation

Each text was segmented in chapters and sentences. A direct observation in chapter segmentation is that there is no direct correspondence between the number of chapters in both languages. As a first intuition, this could be explained by the English version bringing more content together, but by important omissions in content as well. The purpose of creating an aligned corpus was also to answer this question and find content gaps if any.

As all chapters have a heading with a pseudo-paragraph, and there is a different number of chapters, it was obvious from the very beginning that some sentences would result in an omission in the translation. A script parsing the annotated corpus stripped tags and split sentences using dots, semicolons, exclamation and interrogation marks as delimiters. Regular expressions handled exceptions for chapter headings in 1653's text, where semicolons have a function similar to that of commas.

■ **Table 1** Processed segments during corpus preparation. *Tag*: annotated with specific tag. *Script*: retrievable using a script to parse annotated text, even if the category is not annotated. *Db*: relational table in a database. *Txt*: file with raw text in txt format. *HTML*: displayed as web page.

Segment type	PT 1614	EN 1653	Total	Retrieval
Chapters	15	5	20	tag, script, db
Pages	36	20	56	tag, script, db
Sentences	222	353	575	script, db, txt
Word types	3401	2806	6027	script
Tokens	18040	19159	37199	script
Aligned sentences	240	230	470	script, db, HTML

3.4 Bitext

Hunalign³ [11], an open source for automate alignment, was first applied to create the parallel structure of the corpus. Test rounds considered the use of an in-house built dictionary with the 100 words of highest frequency and a gazetteer built from geographical named entities from the NERC annotation of the Portuguese corpus. The best result was converted to a table with two columns, one for the source language, another for the translation. Even if the automate alignment brought related sentences close enough to prefer this procedure over a bitext produced manually from scratch, in order to obtain a golden corpus, results had to be manually corrected and validated. Misaligned rows were arranged and grouped to fit the original style of the source where sentences, defined as a stretch of text ending in a given punctuation mark, are more similar to pseudo-paragraph than to sentences as perceived by modern standards.

4 Results

Through the process of annotation and sentence alignment, a web environment enabled the visualization of the results and served as a repository for the experimental data and related products. Table 1 shows the textual segments obtained in the final parallel corpus.

A web-based interface allows the retrieval of text in different dispositions for either the bitext or any of the Portuguese and English versions only. As an example, figure 3 shows how the tagged text from figure 2 is displayed in a more readable format.

The following units were considered for visualization outputs.

Chapters. A series of scripts evaluates the number of chapters and displays them in HTML format.

Pages. The text grouped by pages following the disposition as in the original first editions.

Sentences. List of sentences using punctuation marks as delimiters. This was also the starting point to generate the bitext. The final segment is, however, not the result of delimiters only, but of the match of both versions for alignment. It is not always the case that an aligned row contains only one grammatical sentence in either the source or the translation. A larger unit, pseudo-paragraphs, may apply. Nevertheless, neither term defines the category well. Apart from those alignments where sentence is an accurate description of the matched segments, there are also complex and compound-complex

³ <http://mokk.bme.hu/resources/hunalign/>

Showing 5 chapters from *Tartaria Corpus EN (1653)*.

(1) Chap. 38

Page 149

CHAP. XXXVIII

A Tartar Commander enters with his Army into the Town of Quincay, and that which followed thereupon; with the Nauticors besieging the Castle of Nixiamcoo, and the taking of it by the means of some of us Portugals.

WE had been now eight months and an half in this captivity, wherein we endured much misery, and many incommodities, for that we had nothing to live upon but what we got by begging up and down the Town, when as one Wednesday, the third of July, in the year 1544, a little after midnight there was such a hurly burly amongst the people, that to hear the noise and cries which was made in every part, one would

■ **Figure 3** Excerpt retrieved from the web environment showing chapters in English.

sentences split in two different alignments. These units fall two levels below pseudo-paragraphs in a hierarchy of text segments. As sentence was the initial term to describe this segment type, aligned sentences was kept as the most representative categorical label. The actual aligned unit represents a balance between single sentences and pseudo-paragraphs made of coordinated sentences (part of a bigger segment which is still a sentence).

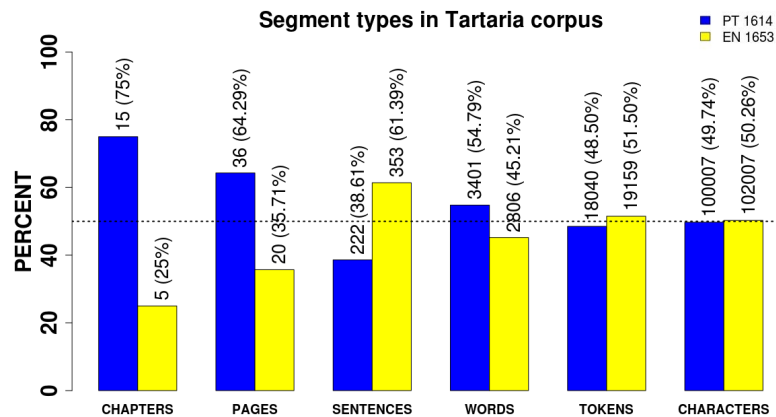
Words. Ranking of word types arranged by its Zipfian distribution and aligned in two rows, PT (1614) and EN (1653). Lexical items with highest frequencies and named entities were selected to create dictionaries for the automatic text alignment tests.

Aligned sentences. An alignment considers an empty sentence when there is no counterpart in either source or translation resulting in an omission (1:0). Relevant complex alignments (m:n) appear in pseudo-paragraphs with embedded transcriptions of the language of the Tartars along with the Portuguese translation (hence allowing for a triple alignment). The whole corpus shows all the alignment types as a list in a table following the narrative sequence.

5 Discussion: which units show evidence of both texts being alignable?

Initial analysis was directed towards answering if both texts were able to generate a parallel corpus. This would mean that for most segments in the source PT 1614 (L1) from chapter to sentence, there is an equivalent segment in the target translation EN 1653 (L2). It could be the case that both texts were not direct translations of each other, but just an account of similar events following a common narrative, though still comparable corpora. This may still allow an alignment at the top of the hierarchy, chapters in our segmentation. On the other hand, if the target is a literal translation, an alignment at the level of sentences is expected. Another possibility is neither chapters nor sentences having the same number of units in L1 and L2, as in a less literal translation that modifies text disposition. In this last case, some units may show no equivalents, though some others would be expected to emerge as indicators of both texts having a translation relation. The procedure was finding out which category allows evaluation in terms of L1 → L2 having ratio 1:1, that is, a proportion of 50% for both L1 and L2 taking the corpus as a whole. Figure 4 shows a comparative graph of the size of the corpus for each language in absolute and relative terms.

Very early in the process, it was noticeable that the highest segmented category, chapters, shows dissimilarity (L1 75%, L2 25%). Each chapter has a heading that adds extra blank lines to a page, so more chapters would slightly affect the number of pages too. This extra



■ **Figure 4** Comparative graph of segment types in Tartaria corpus.

content is not, however, conclusive enough for the different number of pages. Issues such as typography, page size and layout, not considered in the annotation, may explain a different number of pages in both editions.

The number of sentences split by punctuation marks has unbalanced ratios too (L1 38%, L2 61.39 %). It is worth noting that $L1 < L2$, hence there is a contradiction with the higher categories (chapters and pages) where $L1 > L2$. Again, non-considered variables such as editorial preferences, different punctuation standards and more grammar-dependent syntactic disposition of clauses may also explain the different number of sentences in each language regardless of both texts being alignable.

The category of word types still shows a difference between both texts (L1 54.79%, L2 45.21 %). A direct observation of the data shows that morphological features are relevant factors to explain word forms variability. The word form at the top of the Zipfian distribution is the determinate article with ratio 4:1 and values L1 (a, o, os, as) : L2 (the). However, following the same example, the fact that the same word form has more than one different expression in another language, does not necessarily affect the number of tokens. Thus, tokens, the variable appearing as a direct measure of text length, show a more balanced ratio 1 : 1.07 (L1 48.5%, L2 51.5%), an indicator of one text being a translation of the other.

Finally, characters show the closest balance. In fact, round percentages stand for the desired 1:1 (L1 49.74%, L2 50.26%). An explanation for this highest accuracy is that characters do not only represent a similar number of tokens in both texts, but also capture some phonetic and morphological properties of words. In fact, if a word in the source language is polysyllabic or has a derivational or compound structure, its equivalent is most often expected to show a more complex structure in the target language too.

6 Conclusion

Different units were compared to research which one would be the best predictor of two texts being alignable in terms of source and translation. Characters show a parallel ratio, around 1:1, becoming the most accurate feature for predicting the alignability of both texts. From a linguistic point of view, it is intriguing that a unit without semantic value stands as more relevant than morphologically rich and syntactic relevant tokens and word types. The inferred hypothesis to consider for future work is that, when used as a measure of length for

the largest units, characters capture some aspects of the morphologic-syntactic structure. Although they have been extensively used as indicators in text-alignment tasks, to the best of my knowledge, the linguistic implications of such a basic and easily observable phenomenon have not been explained and are still open for further research.

References

- 1 Christian Buck and Philipp Koehn. Findings of the WMT 2016 bilingual document alignment shared task. In *First Conference on Machine Translation – Shared Task Papers*, volume 2, pages 554–563, 2016.
- 2 William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1993.
- 3 Charlotte Galves, Aroldo Leal de Andrade, and Pablo Faria. Tycho brahe parsed corpus of historical Portuguese. <http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip>, 2017.
- 4 Philipp Koehn. EuroParl: A parallel corpus for statistical machine translation. In *Machine Translation Summit*, pages 79–86, 2005.
- 5 I. Dan Melamed. A geometric approach to mapping bitext correspondence. *CoRR*, 1996. URL: <http://arxiv.org/abs/cmp-1g/9609009>.
- 6 Robert C. Moore. Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas*, pages 135–144, 2002.
- 7 Xiaojun Quan, Chunyu Kit, and Yan Song. Non-monotonic sentence alignment via semisupervised learning. In *51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 622–630, 2013.
- 8 André Santos. A survey on parallel corpora alignment. In *Master of Informatics Internal Conference, Universidade do Minho*, pages 117–128, 2011.
- 9 Alberto Simões and Sara Fernandes. XML schemas for parallel corpora. In *XATA 2010: 9ª Conferência Nacional em XML, Aplicações e Tecnologias*, pages 59–69, 2011.
- 10 Hai-Long Trieu, Phuong-Thai Nguyen, and Kim-Anh Nguyen. Improving moore’s sentence alignment method using bilingual word clustering. In *Knowledge and Systems Engineering*, pages 149–160, 2014. doi:10.1007/978-3-319-02741-8_14.
- 11 Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. In *Recent advances in natural language processing IV : selected papers from RANLP 2005*. John Benjamins, 2007.
- 12 Krzysztof Wolk and Krzysztof Marasek. Unsupervised comparable corpora preparation and exploration for bi-lingual translation equivalents. *CoRR*, 2015. URL: <http://arxiv.org/abs/1512.01641>.
- 13 Marcos Zampieri and Martin Becker. Colonia: Corpus of historical Portuguese. In *Non-standard Data Sources in Corpus-based Research*. Shaker Verlag, 2013.
- 14 Federico Zanettin. *Translation-driven corpora: Corpus resources for descriptive and applied translation studies*. Routledge, 2014.