ISSN 2499-4553

# IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 6, Number 1
june 2020

aAccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)


Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn PDF 9791280136404

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_6_1

# IJCoL

**Volume 6, Number 1**
**june 2020**

## CONTENTS

# Towards Automatic Subtitling: Assessing the Quality of Old and New Resources

Alina Karakanta*
Fondazione Bruno Kessler / University of Trento

Matteo Negri**
Fondazione Bruno Kessler

Marco Turchi†
Fondazione Bruno Kessler

*Growing needs in localising multimedia content for global audiences have resulted in Neural Machine Translation (NMT) gradually becoming an established practice in the field of subtitling in order to reduce costs and turn-around times. Contrary to text translation, subtitling is subject to spatial and temporal constraints, which greatly increase the post-processing effort required to restore the NMT output to a proper subtitle format. In our previous work (Karakanta, Negri, and Turchi 2019), we identified several missing elements in the corpora available for training NMT systems specifically tailored for subtitling. In this work, we compare the previously studied corpora with MuST-Cinema, a corpus enabling end-to-end speech to subtitles translation, in terms of the conformity to the constraints of: 1) length and reading speed; and 2) proper line breaks. We show that MuST-Cinema conforms to these constraints and discuss the recent progress the corpus has facilitated in end-to-end speech to subtitles translation.*

## 1. Introduction

Screens have become an essential component of modern life. Screens are omnipresent and they come in different shapes and sizes; from classical television screens, to tablets, smartphones and smartwatches. As a result, we constantly interact with audiovisual material; whether for entertainment, work or education, audiovisual material now comes not only in the form of films and series, but as home-made Youtube videos, online courses, live-streamed events, virtual conferences and meetings. This unprecedented amount of content is generated by content creators from all across the globe, speaking different languages, and reaches an equally diverse audience in terms of linguistic coverage and accessibility needs. Subtitling is probably the fastest and more cost-efficient way for facilitating the access to information from any part of the world to any part of the world.

This omnipresence of screens has led to a growing need for Machine Translation (MT) of subtitles in various applications. Subtitling, as a form of translation, is a complex process consisting of several stages and the actual task of translation is only one step in a long pipeline of sub-tasks (transcription, timing, adaptation). Understandably, manual approaches to subtitling involve different professional roles and are laborious and

* Fondazione Bruno Kessler - Via Sommarive 18, Povo, Trento, Italy. E-mail: akarakanta@fbk.eu
** E-mail: negri@fbk.eu
† E-mail: turchi@fbk.eu

costly. What is more, subtitling has to conform to spatial and temporal constraints. For example, a subtitle cannot be longer than a defined number of characters (length) and it should remain on screen for sufficient time in order to be read by the viewers (reading speed). While length and reading speed can be modelled as a post-processing step in an MT workflow using simple language-independent rules, subtitle segmentation, i.e. where and if to insert a line break between subtitles, depends on semantic and syntactic properties. Subtitle segmentation is particularly important, since it has been shown that a proper segmentation by phrase or sentence significantly reduces reading time and is vital for easy comprehension of the content (Perego 2008; Rajendran et al. 2013). Hence, fully or at least partially automated solutions for subtitle-oriented NMT should aim at reducing, on the one hand, the effort of generating a transcription of the source language text, and on other hand the effort of restoring the output to a proper subtitle format. These solutions would greatly contribute in reducing post-processing effort and speeding-up turn-around times. Automated approaches though, especially NMT, are data-hungry. Performance greatly depends on the availability of large amounts of high-quality data (up to tens of millions of parallel sentences), specifically tailored for the task. In the case of subtitle-oriented NMT, this implies having access to large subtitle training corpora.

There are large amounts of available parallel data extracted from subtitles (Lison and Tiedemann 2016; Pryzant et al. 2018; Di Gangi et al. 2019; Karakanta, Negri, and Turchi 2020b). In our previous work (Karakanta, Negri, and Turchi 2019), we explored whether these large, publicly available parallel data compiled from subtitles conform to the temporal and spatial constraints necessary for achieving quality subtitles. These corpora are usually obtained by collecting files in a subtitle specific format (*.srt*) in several languages and then parsing and aligning them at sentence level. As shown in Figure 1, this process compromises the subtitle format by converting the subtitle blocks into full sentences. With this "merging", information about subtitle segmentation (line breaks), as well as all the information related to the timing of the subtitle (timestamps), is lost. We concluded that, while most of the subtitling corpora conform to some extent to the subtitling constraints, the loss of line breaks and information about the duration of the utterance is a limitation towards the development of effective end-to-end NMT solutions geared to subtitling. We raised several open issues for creating corpora for subtitling-oriented NMT: i) **subtitling constraints:** a subtitling corpus, in order to be representative of the task, should respect the subtitling constraints; ii) **duration of utterance:** since the translation of a subtitle depends on the duration of the utterance, time information is highly relevant; iii) **integrity of documents:** a subtitle often occupies several lines, therefore the order of subtitles should be preserved whenever possible; iv) **line break information:** while parallel sentence alignments are indispensable, they should not compromise line break and subtitle block information.

This paper extends our previous work by adding to the comparison of the existing subtitling corpora a new corpus, MuST-Cinema (Karakanta, Negri, and Turchi 2020b). This speech-to-subtitles corpus has been built to enable training end-to-end speech translation systems which to directly translate speech into properly formed subtitles, and therefore addresses the open issues described above. To evaluate compliance to length and reading speed constraints we employ character counts while for proper line breaks we use the Chink-Chunk algorithm (Liberman and Church 1992). We show that MuST-Cinema, except for its other advantages (audio input, utterance duration, line breaks) has a high conformity to the subtitling constraints. Finally, we discuss the recent progress in end-to-end NMT solutions for subtitling driven by MuST-Cinema. The paper is structured as follows: In Section 2 we present the existing corpora built

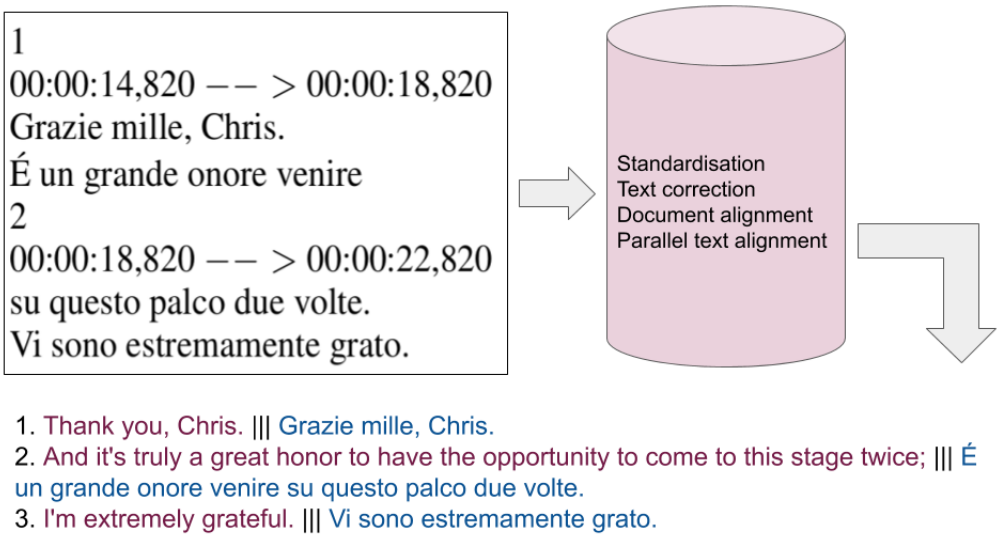


**Figure 1**
Subtitle blocks (top, 1-2) as they appear in an .srt file and the processed output for obtaining aligned sentences (bottom).

from subtitles and previous work on subtitle segmentation in a monolingual or cross-lingual setting. Section 3 explains the technical aspects of subtitling in relation to the spatial and temporal constraints and presents the criteria filters used to isolate the material that conforms to these constraints. In Section 4 we describe the experiment. The results and their analysis are presented in Section 5 and finally, Section 6 summarises and concludes our work.

## 2. Related work

### 2.1 Subtitling corpora

Several projects have attempted to collect parallel subtitling corpora. The most well-known one is the OpenSubtitles[1] corpus (Lison and Tiedemann 2016), extracted from 3.7 million subtitles across 60 languages. Since subtitle blocks do not always correspond to sentences (see Figure 1), the blocks are merged and then segmented into sentences using heuristics based on time codes and punctuation. Then, the extracted sentences are aligned to create parallel corpora with the time-overlap algorithm (Tiedemann 2008) and bilingual dictionaries. Although the 2018 version of OpenSubtitles has high-quality sentence alignments, it does not resemble the realistic subtitling scenario described above, since time and line break information are lost in the merging process. The same methodology was used for compiling MontenegrinSubs (Božović et al. 2018), an English – Montenegrin parallel corpus of subtitles, which contains only 68k sentences.

1 http://www.opensubtitles.org/

The Japanese-English Subtitle Corpus JESC (Pryzant et al. 2018) is a large parallel subtitling corpus consisting of 2.8 million sentences. It was created by crawling the internet for film and TV subtitles and aligning their captions with improved document and caption alignment algorithms. This corpus is aligned at caption level, therefore its format is closer to our scenario. On the other hand, non-matching alignments are discarded, which might hurt the integrity of the subtitling documents. As we will show, this is particularly important for learning proper line breaks between subtitle blocks.

A corpus preserving both subtitle segmentation and order of lines is SubCo (Martínez and Vela 2016), which comprises machine and human translated subtitles for English–German. However, it only consists of 2 source texts (∼150 captions each) with multiple student and machine translations. Therefore, it is not sufficient for training competitive NMT systems, although it could be useful for evaluation because of the multiple reference translations.

The EuroparlTV Multimedia Parallel Corpus (EMPAC) (Serrat Roozen and Martínez Martínez 2020) is an English–Spanish Corpus compiled from videos from the European Parliament's Multimedia Centre. The corpus is a substantial resource of institutional subtitling, since it has been built from .srt files corresponding to 280 hours of video, a total of 2.5 million tokens. Thanks to a complete pipeline, the corpus features linguistic annotation of the subtitles, alignment at document and subtitle level and identification of wrong subtitle segmentation based on a rule-based approach for English and Spanish. EMPAC has been used for a detailed analysis of accessibility in terms of subtitling constraints at the European Parliament (Serrat Roozen 2019). The findings showed a tendency towards longer and faster subtitles as the years went by.

Also deviating from the domain of films and TV series, corpora for Spoken Language Translation (SLT) have been created based on TED talks. WIT$^3$, the Web Inventory of Transcribed and Translated Talks (Cettolo, Girardi, and Federico 2012), is a multilingual collection of transcriptions and translations of TED talks. The talks are aligned at sentence level without audio information. Based on WIT$^3$, the IWSLT campaigns (Ansari et al. 2020) are annually releasing parallel data and their corresponding audio for the task of SLT, which are extracted based on time codes but again with merging operations to create segments. MuST-C (Di Gangi et al. 2019; Cattoni et al. 2021) is to date the largest multilingual corpus for end-to-end speech translation. It contains (*audio*, *source language transcription*, *target language translation*) triplets, aligned at segment level. Its creation process is the opposite from IWSLT; the authors first align the written parts and then match the audio. MuST-C has been widely used for the deployment of end-to-end ST systems. However, the translations are merged to create sentences, therefore they are far from the suitable subtitle format.

MuST-Cinema (Karakanta, Negri, and Turchi 2020b) has been created to address the challenges we described in our previous work (Karakanta, Negri, and Turchi 2019). It is practically an extension of the MuST-C corpus, where the parallel sentences have been annotated with symbols corresponding to subtitle breaks. `<eol>` corresponds to line breaks (breaks inside the subtitle block), while `<eob>` corresponds to block breaks (the whole text appearing in a single video frame). In MuST-Cinema, the sentences in Figure 1, annotated with breaks become:

Grazie mille, Chris. <eol>
É un grande onore venire <eob> su questo palco due volte. <eol>
Vi sono estremamente grato. <eob>

Moreover, the fact that MuST-Cinema is a speech corpus (audio as source language input) has opened up new possibilities for translating directly from the audio into subtitles, without the need for complex cascades composed of several modules (ASR for transcription, NMT, classifier for segmentation into subtitles).

### 2.2 Readable subtitles

Automatically generating readable subtitles has recently received growing attention. As far as subtitle segmentation is concerned, automatic techniques have so far mainly focused on monolingual scenarios. Álvarez, Arzelus, and Etchegoyhen (2014) trained Support Vector Machine and Logistic Regression classifiers on correctly/incorrectly segmented subtitles to predict line breaks. Extending this work, Álvarez et al. (2017) used a Conditional Random Field (CRF) classifier for the same task, but this time differentiating between line breaks (next subtitle line) and subtitle breaks (next subtitle block). Song et al. (2019) used a neural network-based approach to subtitle segmentation, closer to the problem of boundary detection. They employed a Long-Short Term Memory Network (LSTM) to predict the position of the period in order to improve the readability of automatically generated Youtube captions. Recently, Karakanta, Negri, and Turchi (2020b) trained a sequence-to-sequence model which receives a full sentence and generates the same sentence inserting symbols which correspond to subtitle breaks. Focusing on the length constraint, Liu, Niehues, and Spanakis (2020) proposed adapting an Automatic Speech Recognition (ASR) system to incorporate transcription and text compression, for generating more readable subtitles.

A recent line of works has paved the way for NMT systems which generate readable subtitles in a multilingual scenario. Matusov, Wilken, and Georgakopoulou (2019) customised an NMT system for the task of subtitling. They introduced a segmentation module based on human segmentation decisions, trained on OpenSubtitles, and with penalties well established in the subtitling industry. Karakanta, Negri, and Turchi (2020a) were the first to propose an end-to-end solution for Speech Translation into subtitles. Their findings indicated the importance of prosody, and more specifically pauses, to achieving subtitle segmentation in line with the speech rhythm. They further confirmed the different roles of line breaks (new line inside a subtitle block) and subtitle block breaks (the next subtitle appears on a new screen); while block breaks depend on speech rhythm, line breaks follow syntactic patterns. All these works show that subtitle segmentation and the generation of readable subtitles is a complex and dynamic process which depends on several and varied factors.

## 3. Criteria for assessing subtitle quality

### 3.1 Background

The quality of the translated subtitles does not only depend on fluency and adequacy, but also on their format, i.e. the way they appear on screen. Both these aspects are crucial for the evaluation of subtitles. We assess whether the available subtitle corpora conform to the constraints of length, reading speed (for the corpora where time information is available) and proper line breaks on the basis of the criteria for subtitle segmentation

mentioned in the literature of audiovisual translation (Cintas and Remael 2007; Carroll and Ivarsson 1998) and the TED talk subtitling guidelines[2]:

1. **Characters per line**. The space available for a subtitle is limited. Subtitles need to be concise. The length of a subtitle depends on different factors, such as size of screen, font, age of the audience and country. Typical lengths consider max. 42 chars for Latin alphabets, 14 for Japanese and Korean, 16 for Chinese (including spaces).
2. **Lines per subtitle**. Subtitles should not take up too much space on screen. The space allowed for a subtitle is about 20% of screen space. Therefore, a subtitle block should not exceed 2 lines.
3. **Reading speed**. The on-air time of a subtitle should be sufficient for the audience to read and process its content. Moreover, a subtitle should be synchronised to the speech, in the sense that it should match as much as possible the start and the end of an utterance. The duration of the utterance (measured either in seconds or in feet/frames) is directly equivalent to the space a subtitle should occupy. Even for very short subtitles, a minimum duration (1 second) is set in order to avoid thee effect of the subtitle "flashing" in from of the eyes of the viewers. On the other hand, a subtitle should not remain visible on the screen for too long, as this causes the audience to re-read it. There is no clear consensus in the research world on the ideal reading speed (Cintas and Remael 2007; Swarkowska 2013; Romero-Fresco 2015). Depending on the audience needs, age and screen, reading speed rates range between 12-21 characters per second.
4. **Preserve 'linguistic wholes'**. This criterion is related to subtitle segmentation. Subtitle segmentation does not rely only on the allowed length, but should respect linguistic norms. To facilitate readability, subtitle splits should not "break" semantic and syntactic units. In an ideal case, every subtitle line (or at least subtitle block) should be self-contained and/or represent a coherent linguistic chunk (*i.e.* a sentence or a phrase). For example, a noun should not be separated from its article. Lastly, the segmentation of subtitles should respect the coherence with the audio and video modalities, therefore it should correspond to the natural pauses in speech and the shot changes in the video.
5. **Equal length of lines**. Another criterion for splitting subtitles relates to aesthetics. There is no consensus about whether the top line should be longer or shorter, however, it has been shown that subtitle lines of equal length are easier to read, because the viewer's eyes return to the same point on the screen when reading the second line.

While subtitle length and reading speed are hard constraints that can be controlled directly by the software used by subtitlers, subtitle segmentation is left to the decision of the subtitler. A sentence could be segmented into subtitles in multiple, equally plausible ways. Subtitlers often have to compromise the aesthetics in favour of the linguistic wholes or vice versa, and resort to omissions and substitutions in order to conform to

2 https://www.ted.com/participate/translate/guidelines

the constraints as much as possible. This shows that several interplaying factors define segmentation decisions. Therefore, accurately modelling these segmentation decisions based on the large available corpora is of great importance for a high-quality subtitle-oriented NMT system.

### 3.2 Quality criteria filters

In order to assess the conformity of the existing subtitle corpora to the constraints mentioned above, we implement the following filters, inspired by the literature in audiovisual translation described in the section 3.1.

*Characters per line (CPL).* As mentioned in the introduction of this work, the information about line breaks inside subtitle blocks is discarded in the process of creating parallel data. Therefore, we can only assume that a subtitle fulfils the criteria 1 and 2 above by calculating the maximum possible length ($max\_length$) for a subtitle block; 2 * 42 = 84 characters for Latin scripts and 2 * 14 = 28 for Japanese. If $CPL > max\_length$ then the subtitle doesn't conform to the length constraints.

*Characters per second (CPS).* This metric relates to reading speed. For the corpora where time codes and duration are available, we calculate CPS as the total number of characters in the subtitle, divided by the total duration of the utterance as follows:

$$CPS = \frac{\#chars}{duration} \tag{1}$$

*Chink-Chunk.* Chink-Chunk (Liberman and Church 1992) is a low-level parsing algorithm which can be used as a rule-based method to insert line breaks between subtitles. It is a simple but efficient way to detect syntactic boundaries. It receives Part-of-Speech information to distinguish between content vs. function words. However, tensed verbs can behave like auxiliaries, while objective pronouns can behave like nouns, acting as content words. Therefore, the two defined categories are the following: *chink*→function words + tensed verbs and *chunk*→content words + objective pronouns. The parsing algorithm greedily matches {chink* chunk*}. We apply this simple algorithm to detect subtitles which preserve linguistic wholes, by accepting subtitles ending either at punctuation marks (logical completion) or when the subtitle break occurs between a chunk followed by a chink[3]. Here, we use this algorithm to compute statistics about the type of subtitle block breaks in the data (punctuation break, chunk-chink break or other). The algorithm is described in Figure 1.

Since no original information about subtitle line breaks (inside a subtitle block) is preserved in any of the corpora, the criterion of equal length of lines cannot be explored in this study.

3 In practice, the decision of how to split a subtitle is a much more complex procedure depending on several factors, such as number of characters per line, speaker changes, shot and scene changes. In industry, language- and country-specific rules are applied to determine the segmentation into subtitles. In this work, we selected the Chink-Chunk algorithm as a simple, language-independent solution to evaluate segmentation.

**Algorithm 1:** Algorithm for assessing subtitling segmentation

```
Result: Number of correctly/incorrectly segmented subtitles
if POS_last in ['PUNCT', 'SYM', 'X'] then
    punc +=1;
else
    if POS_last in chunk and POS_next in chink then
        chunk-chink +=1;
    else
        other_split +=1;
    end
end
correct = punc + chunk-chink
return correct, other_split
```

## 4. Experiments

For our experiments we consider the corpora which are large enough to train NMT systems; OpenSubtitles, JESC, MuST-C and MuST-Cinema. We focus on three language pairs, Japanese, Italian and German, paired with English, as languages coming from different families and having a large portion of sentences in all corpora. Table 1 shows the statistics of the corpora. MuST-C and MuST-Cinema contain the same number of sentences[4]. For this analysis we take into account only the training sets, which are the same for the two corpora; the only difference between the corpora is the annotation of the subtitle breaks.

Computing the percentage of subtitles conforming to the characters per line criterion is performed as follows: for OpenSubtitles and MuST-C, the number of characters corresponds to the total number of characters in each sentence (line in the parallel corpus). For MuST-Cinema instead, the sentences are split at the subtitle block symbols `<eob>` and the number of characters is computed for each subtitle separately.

Calculating the conformity to the criterion of reading speed (characters per second) is possible only when access to timestamps is possible. For OpenSubtitles, timestamps have been removed from the parallel corpora, however they are preserved in the monolingual corpora in XML format. In order to compute CPS for OpenSubtitles, we use the file IDs to identify the files which were used to compile the parallel data and calculate the percentage of conforming subtitles based on the monolingual data. Therefore, CPS is computed here at the subtitle level. The same holds for MuST-Cinema, where we can refer to the original subtitles to obtain CPS for individual subtitles. For MuST-C, CPS is computed at the sentence level, i.e. the characters of the full sentence are divided by the duration of the utterance. JESC does not contain any metadata, therefore computing CPS is not possible. The current versions of MuST-C and MuST-Cinema do not include Japanese as a source language, however we will still perform the analysis on JESC and OpenSubtitles.

4 For more information on the structure, dimension, number of speakers, organisation, etc. please refer to the related publications

**Table 1**
Size of the parallel corpora in sentences. MuST-C and MuST-Cinema contain the same number of sentences.

| Corpus | EN-IT | EN-DE | EN-JA |
|---|---|---|---|
| OpenSubtitles | 35.216M | 22.512M | 2.083M |
| JESC | - | - | 2.801M |
| MuST-C(inema) | 258K | 234K | - |

For the Chink-Chunk algorithm we preprocess the data using the Stanza toolkit (Qi et al. 2020). We first tokenise and perform Multi-Word Token (MWT) expansion to split the words into syntactic units. Then, we tag the data to obtain universal POS tags[5].

We apply each of the quality criteria filters discussed in Section 3.2 to the corpora both on the source and the target side independently. Then, we take the intersection of the outputs of all the filters to obtain the total number of lines/sentences which conform to all the criteria.

## 5. Results and analysis

**Table 2**
Percentage of data preserved after applying each of the quality criteria filters on the subtitling corpora independently. Percentages are given on source and target side (s/t), except for the *Total* where source and target are combined. For OpenSubtitles, time information is only present in its monolingual corpora (m).

| EN- | Corpus | Format | Time | CPL (s/t) % | CPS (s/t) % | CC (s/t) % | Total% |
|---|---|---|---|---|---|---|---|
| IT | MuST-C | segment | ✓ | 49 / 48 | 78 / 72 | 99 / 99 | 45 |
| | OpenSubtitles | segment | ✓(m) | 95 / 94 | 84 / 80 | 99 / 99 | 78 |
| | M-Cinema | subtitle | ✓ | 95 / 91 | 91 / 86 | 89 / 88 | 83 |
| DE | MuST-C | segment | ✓ | 51 / 47 | 77 / 66 | 99 / 99 | 42 |
| | OpenSubtitles | segment | ✓(m) | 95 / 95 | 84 / 83 | 99 / 99 | 80 |
| | M-Cinema | subtitle | ✓ | 94 / 91 | 90 / 80 | 90 / 87 | 78 |
| JA | OpenSubtitles | segment | ✓(m) | 96 / 93 | 87 / 90 | 99 / 98 | 86 |
| | JESC | subtitle | - | 97 / 94 | - | 88 / 87 | 85 |

Table 2 shows the percentage of preserved lines/sentences after applying each criterion.

*Length.* The analysis based on the characters per line filter shows that OpenSubtitles, JESC and MuST-Cinema conform to the quality criterion of length in at least 91% of the cases. Despite the merging operations to obtain sentence alignments, OpenSubtitles still preserves a short length of lines, possibly because of the nature of the text of film and TV series subtitles. A manual inspection shows that the text contains mainly short

5 The performance of the Stanza models for the task of UPOS can be found here: https://stanfordnlp.github.io/stanza/performance.html

dialogues and the long sentences are parts of descriptions or monologues, which are rather rare. On the other hand, the merging operations in MuST-C create long sentences that do not resemble the subtitling format. This could be attributed to the format of TED talks. TED talks mostly contain text in the modality 'written to be spoken'. The talks are prepared, usually delivered by one speaker with few (if any) dialogue turns. Despite having the same data as basis, MuST-Cinema shows a high conformity to the criterion of length, since the annotation of the subtitle and line breaks allow us to consider subtitles instead of full sentences. A lower conformity for the target side (91% vs 94%/95%) is due to a small number of TED talks which contain a large percentage of non-conforming subtitles. In Karakanta (2020) we observed that these talks come from the earlier years of the program and the non-conformity is due to the fact that the translations of these early talks are verbatim translations of the transcription, without any omission or substitution strategies to generate proper subtitles. The conformity of TED talks to the constraint of length has grown over the years, possibly because of the introduction of a subtitling tool and a stricter quality assurance process involving several linguists. In fact, all talks after 2012 have almost 100% conformity to the constraint of length.

*Reading speed.* Conformity to the criterion of reading speed is achieved to a lesser degree, as shown by the characters per second filter. OpenSubtitles shows a conformity between 80%-84% for Italian and German, while for Japanese, where the allowed number of characters per line is lower, the conformity reaches 90%. It should be taken into account that the threshold of 21 characters per second which we chose for Latin scripts in this analysis is already considered quite high for some audiences, and, as a result, the criterion of reading speed seems to be less important or more difficult to respect by subtitlers. This could be attributed to the fact that the subtitle files in OpenSubtitles come from different sources, therefore it is possible that they abide by different genre and client specifications. The mix of professional, amateur (fansubs) and machine-translated subtitles might also contribute to this trend. Unfortunately, time information is not present in JESC, therefore a full comparison is not possible for Japanese. On the other hand, MuST-Cinema, being in subtitling format, conforms better to reading speed, with 80%-91% of the subtitles having a reading speed equal or less than 21 characters per second.

*Linguistic wholes.* The Chink-Chunk algorithm shows interesting properties of the subtitle breaks for all the corpora. MuST-C and OpenSubtitles conform to the criterion of preserving linguistic wholes in 99% of the sentences, which does not occur in the corpora in subtitle format. JESC shows an acceptable segmentation for 88-87% of the subtitles included and MuST-Cinema 87-90%. Since JESC is compiled by removing captions based on unmatched time codes, the integrity of the documents is possibly broken. Subtitles are removed arbitrarily, so consecutive subtitles are often not kept in the order they appear in the .srt files. Removing subtitles arbitrarily makes this type of corpora unsuitable for document-level MT techniques, which have often been preferred in the MT of subtitles, since increasing the context improves performance by enhancing coreference resolution and disambiguation. This advocates for the importance of preserving the order of subtitles when creating subtitling corpora. In MuST-Cinema, no subtitles were removed inside the sentences (only individual non-aligned sentences). Therefore, what can this non-conformity be attributed to?

In order to provide an explanation, we further analysed the categories of breaks based on the Chink-Chunk algorithm: 1) breaks after punctuation marks, 2) breaks between a content and a function word, and 3) incorrect breaks. The percentages are

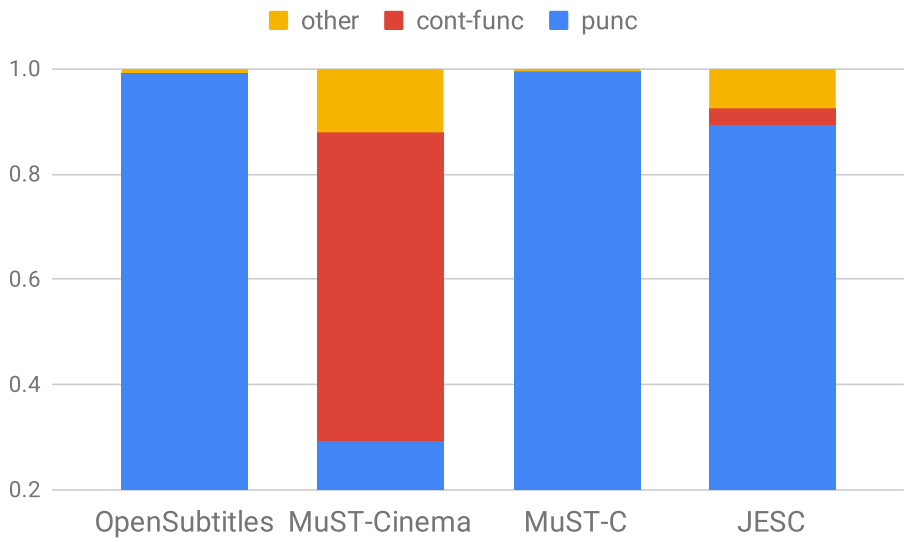


**Figure 2**
Percentages of the positions where breaks occur in the total breaks in each corpus. *punc*: the break occurs after a punctuation mark. *cont-func*: the break occurs between a chunk and a chink (content and function word). *other*: the break occurs in another position.

shown in Figure 2. A close inspection of the breaks shows that OpenSubtitles and MuST-C end in a punctuation mark in 99.9% of the cases. This result confirms that the two corpora are created by merging the subtitles into proper sentences. Even though they preserve logical completion, these corpora do not contain sufficient examples of line breaks inside a sentence. We observe a similar tendency for JESC, where 90% of the sentences end in a punctuation mark, but the content-function category represents only a 3% of the cases. On the other hand, for MuST-Cinema, about half of the subtitle breaks occur between a chunk and a chink, and therefore inside the sentences. This shows a much larger variation in the positions of breaks. In a realistic subtitling scenario, an NMT system at inference time will often receive unfinished sentences, either from an audio stream or a subtitling template. Therefore, line break information is valuable for training NMT systems that learn to translate and segment. This has been shown in Karakanta, Negri, and Turchi (2020a), where training on the MuST-Cinema data, annotated with the special symbols corresponding to subtitle and line breaks, achieved high conformity to the subtitling constraints, without any modification in the NMT architecture. Similarly, the segmentation module presented in (Matusov, Wilken, and Georgakopoulou 2019), which is trained on the monolingual OpenSubtitles data (which contain breaks as metadata), led to significant reductions in post-editing effort.

The total retained material (last column in Table 2) shows that OpenSubtitles and MuST-Cinema are the most suitable corpora for producing quality subtitles in the investigated languages. For Italian, the retained material for MuST-Cinema is 83% of the total material and 78% for OpenSubtitles. The opposite tendency is shown for German, with 80% for OpenSubtitles and 78% for MuST-Cinema. Two factors should be taken into account in this comparison, when considering training subtitling-oriented NMT systems with these corpora. Firstly, the size of the corpora, with OpenSubtitles being

more than 10 times larger than MuST-Cinema, and secondly, the subcategorisation of the genre of the kind of subtitles contained. Müller and Volk (2013) have shown that the subtitles in corpora based on TED and OpenSubtitles are different and MT systems trained on each subgenre cannot be used interchangeably. Moreover, one serious bottleneck of OpenSubtitles in developing Speech Translation solutions for subtitling is the lack of audio input, since most films and series are protected by copyright and therefore access to the audio is restricted. For JESC, 85% of the sentences passed the filters. However, a fair comparison is not possible, given that no information about reading speed is present and thus the data was filtered with only 2 out of the 3 filters. We showed that corpora in subtitling format (JESC and mainly MuST-Cinema) contain useful information about line breaks not ending in punctuation marks, which are mostly absent in the parallel corpora of OpenSubtitles. However, some of this information is still present as metadata in the monolingual OpenSubtitles. A process similar to the creation process of MuST-Cinema (annotating the sentences with symbols indicating the breaks based on the original .srt files) could be applied to OpenSubtitles to annotate the parallel corpora with symbols representing subtitle breaks. This would increase the size of the corpora annotated with subtitle breaks, which might prove useful for an NMT output already segmented in subtitles, reducing significantly the post-processing effort.

## 6. Conclusions

In this paper, we explored whether the existing parallel subtitling resources conform to the subtitling constraints of length, reading speed and proper segmentation. More specifically, we added a new subtitling corpus, MuST-Cinema to the comparison of conformity to the constraints. We found that all subtitling corpora (OpenSubtitles, JESC and MuST-Cinema) generally conform to length, reading speed and proper line breaks, despite the merging operations for aligning parallel sentences. However, the same is not the case with MuST-C, where merging has resulted in a corpus not representing the subtitle format.

MuST-Cinema has addressed the open issues described in (Karakanta, Negri, and Turchi 2019). There are no missing subtitles inside the sentences, which means that the order of subtitles is kept intact. Line and block break information is present in the form of special symbols. Its largest advantage compared to the other subtitling corpora is the audio input, which has allowed for end-to-end Speech translation solutions for subtitling (Karakanta, Negri, and Turchi 2020a), eliminating the need for a human transcription of the audio and for a separate module to segment the sentences into subtitles. Here, we have shown that MuST-Cinema has a high conformity to the subtitling constraints, reaching (and in some cases exceeding) the conformity level of OpenSubtitles, which was previously found to be the corpus most representative of the subtitling format. Still, OpenSubtitles has the advantage of a much larger corpus size and its availability for more than 60 languages. All in all, both resources have advantages and disadvantages. MuST-Cinema and OpenSubtitles could be used complementarily to support multiple domains, such as films, series, interviews, talks, documentaries, and single as well as multiple speaker events. Therefore, an intelligent combination of old and new subtitling resources based on these advantages could be advantageous for the creation of NMT solutions for subtitling.

## References


Álvarez, Aitor, Haritz Arzelus, and Thierry Etchegoyhen. 2014. Towards customized automatic segmentation of subtitles. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 229–238, Cham. Springer International Publishing.

Álvarez, Aitor, Carlos-D. Martínez-Hinarejos, Haritz Arzelus, Marina Balenciaga, and Arantza del Pozo. 2017. Improving the automatic segmentation of subtitles through conditional random field. In *Speech Communication*, volume 88, pages 83–95. Elsevier BV.

Ansari, Ebrahim, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online, July. Association for Computational Linguistics.

Božović, Petar, Tomaž Erjavec, Jörg Tiedemann, Nikola Ljubešić, and Vojko Gorjanc. 2018. Opus-Montenegrinsubs 1.0: First electronic corpus of the Montenegrin language. In *Conference on Language Technologies & Digital Humanities*, Ljubljana, September.

Carroll, Mary and Jan Ivarsson. 1998. *Code of Good Subtitling Practice*. Simrishamn: TransEdit.

Cattoni, Roldano, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language Journal*, 66.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. Wit$^3$: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16$^{th}$ Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Cintas, Jorge Diaz and Aline Remael. 2007. *Audiovisual Translation: Subtitling*. Translation practices explained. Routledge.

Di Gangi, Mattia Antonino, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Minneapolis, MN, USA, June.

Karakanta, Alina. 2020. Subtitling in Transition: the case of TED Talks. In *Book of Abstracts: Translation in Transition (TT5)*, Kent State University, October, 2020.

Karakanta, Alina, Matteo Negri, and Marco Turchi. 2019. Are Subtitling Corpora really Subtitle-like? In *Sixth Italian Conference on Computational Linguistics, CLiC-It*, Bari, November, 2019.

Karakanta, Alina, Matteo Negri, and Marco Turchi. 2020a. Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 209–219, Online, July. Association for Computational Linguistics.

Karakanta, Alina, Matteo Negri, and Marco Turchi. 2020b. MuST-cinema: a speech-to-subtitles corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3727–3734, Marseille, France, May. European Language Resources Association.

Liberman, Mark and Kenneth Church. 1992. Text analysis and word pronunciation in text-to-speech synthesis. *Advances in Speech Signal Processing*, pages 791–831.

Lison, Pierre and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from Movie and TV subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*, Portorož, Slovenia, may.

Liu, Danni, Jan Niehues, and Gerasimos Spanakis. 2020. Adapting end-to-end speech recognition for readable subtitles. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 247–256, Online, July. Association for Computational Linguistics.

Martínez, José Manuel Martínez and Mihaela Vela. 2016. SubCo: A learner translation corpus of human and machine subtitles. In *Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia, may.

Matusov, Evgeny, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy, August. Association for Computational Linguistics.

Müller, Mathias and Martin Volk. 2013. Statistical machine translation of subtitles: From opensubtitles to ted. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language*

*Processing and Knowledge in the Web*, pages 132–138, Berlin, Heidelberg. Springer Berlin Heidelberg.

Perego, Elisa. 2008. Subtitles and line-breaks: Towards improved readability. *Between Text and Image: Updating research in screen translation*, 78(1):211–223.

Pryzant, Reid, Yongjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English Subtitle Corpus. In *Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan, May.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online, July.

Rajendran, Dhevi J., Andrew T. Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. 2013. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, 21(1):5–21.

Romero-Fresco, P. (ed.), editor. 2015. *The Reception of Subtitles for the Deaf and Hard of Hearing in Europe*. Peter Lang.

Serrat Roozen, Iris. 2019. *Análisis descriptivo de la accesibilidad a los contenidos audiovisuales de webs del Parlamento Europeo*. Ph.D. thesis, Universitat Jaume I, Castelló de la Plana, 7.

Serrat Roozen, Iris and José Manuel Martínez Martínez. 2020. EMPAC: an English–Spanish corpus of institutional subtitles. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4044–4053, Marseille, France, May. European Language Resources Association.

Song, Hye-Jeong, Hong-Ki Kim, Jong-Dae Kim, Chan-Young Park, and Yu-Seop Kim. 2019. Inter-sentence segmentation of YouTube subtitles using long-short term memory (LSTM). 9:1504.

Swarkowska, A. 2013. Towards interlingual subtitling for the deaf and hard of hearing. *Perspectives: Studies in Translatology*, 21:68–81.

Tiedemann, Jörg. 2008. Synchronizing translated movie subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, Marrakesh, Morocco.