

UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Automatic Subtitling
A new Paradigm

Alina Karakanta

Advisors

Matteo Negri Fondazione Bruno Kessler

Marco Turchi Fondazione Bruno Kessler

Committee

Dr. Evgeny Matusov AppTek

Prof. Maarit Koponen University of Eastern Finland

Prof. Rico Sennrich Universität Zürich

Dr. Marco Guerini Fondazione Bruno Kessler

11 November 2022

Abstract

Audiovisual Translation (AVT) is a field where Machine Translation (MT) has long found limited success mainly due to the multimodal nature of the source and the formal requirements of the target text. Subtitling is the predominant AVT type, quickly and easily providing access to the vast amounts of audiovisual content becoming available daily. Automation in subtitling has so far focused on MT systems which translate source language subtitles, already transcribed and timed by humans. With recent developments in speech translation (ST), the time is ripe for extended automation in subtitling, with end-to-end solutions for obtaining target language subtitles directly from the source speech. In this thesis, we address the key steps for accomplishing the new paradigm of automatic subtitling: data, models and evaluation. First, we address the lack of representative data by compiling MuST-Cinema, a speech-to-subtitles corpus. Segmenter models trained on MuST-Cinema accurately split sentences into subtitles, and enable automatic data augmentation techniques. Having representative data at hand, we move to developing direct ST models for three scenarios: offline subtitling, dual subtitling, live subtitling. Lastly, we propose methods for evaluating subtitle-specific aspects, such as metrics for subtitle segmentation, a product- and process-based exploration of the effect of spotting changes in the subtitle post-editing process, and finally, a comprehensive survey on subtitlers' user experience and views on automatic subtitling. Our findings show the potential of speech technologies for extending automation in subtitling to provide multilingual access to information and communication.

Keywords

Subtitling, Speech Translation, Neural Machine Translation, Evaluation

Contents

1 Introduction **3**

1.1 The audiovisualisation of communication 3

1.2 Accessibility and audiovisual translation 6

1.3 Automatising audiovisual translation 9

1.4 Innovative aspects and contributions 11

1.5 Structure of the thesis 13

1.6 Activities and dissemination 15

2 Audiovisual Translation **21**

2.1 Audiovisual translation and its remit 22

2.1.1 Multimodality 24

2.1.2 Types of AVT 27

2.2 Subtitling 31

2.2.1 Types of subtitling 31

2.2.2 Subtitling constraints 35

2.2.3 Quality 42

2.2.4 Tools and master templates 45

2.2.5 Volunteer subtitling at TED 49

2.3 Conclusion 51

3 Neural Machine Translation **53**

3.1 NMT and sequence-to-sequence learning 54

3.1.1 Attention . . . 57
3.1.2 Transformer . . . 60
3.2 Speech translation . . . 62
3.2.1 Cascaded speech-to-text translation . . . 63
3.2.2 Direct speech-to-text translation . . . 65
3.2.3 Joint generation of transcription and translation in ST 68
3.2.4 Simultaneous speech translation . . . 69
3.3 Automatic subtitle segmentation . . . 71
3.4 Machine translation for subtitling . . . 72
3.5 Evaluation . . . 75
3.5.1 Human evaluation . . . 75
3.5.2 Automatic evaluation . . . 78
3.6 Conclusion . . . 81

4 Speech Subtitling Data 83
4.1 Are Subtitling Corpora really subtitle-like? . . . 84
4.1.1 Existing subtitling data . . . 85
4.1.2 Conformity criteria . . . 86
4.1.3 Results . . . 88
4.2 The MuST-Cinema corpus . . . 90
4.2.1 Corpus compilation . . . 90
4.2.2 Corpus structure . . . 93
4.2.3 Iterative re-annotation of line breaks . . . 96
4.2.4 Evaluation and analysis . . . 97
4.3 Automatic subtitle segmentation . . . 101
4.3.1 MuST-Cinema annotation . . . 102
4.3.2 Heterogeneous data for subtitle segmentation . . . 110
4.3.3 Multimodal and multilingual segmentation . . . 117
4.4 Conclusions . . . 124

5 End-to-end Models for Subtitling **127**
5.1 Is 42 the answer to everything in automatising subtitling? 128
5.2 Offline automatic subtitling 130
5.2.1 Experimental setting 131
5.2.2 Results 135
5.3 Joint generation of captions and subtitles 143
5.3.1 Models 145
5.3.2 Evaluation of subtitling consistency 147
5.3.3 Experimental setting 149
5.3.4 Results 151
5.4 Live subtitling 156
5.4.1 SimulST for live subtitling 158
5.4.2 Evaluation of display modes 161
5.4.3 Experimental setting 165
5.4.4 Results 166
5.5 Conclusions 171

6 Evaluation of Subtitle-specific Aspects **173**
6.1 Is BLEU all you need? 174
6.2 Evaluating subtitle segmentation 177
6.2.1 BLEU_{br} vs. BLEU_{nb} 178
6.2.2 *Sigma*: a new Subtitle Segmentation Score 180
6.2.3 Experimental setting 181
6.2.4 Results 183
6.3 Auto-spotting in the spotlight 189
6.3.1 Post-editing corpus 189
6.3.2 Experimental setting 191
6.3.3 Results 193
6.4 User experience and subtitlers' views 198

6.4.1 Methodology 199
6.4.2 Results 203
6.4.3 Discussion 209
6.4.4 Recommendations for improvement 212
6.5 Conclusions 214

7 Conclusion **217**
7.1 A new paradigm 217
7.2 Open challenges 221
7.3 The future of AVT 224

Bibliography **227**

A Machine Translation for Dubbing **281**
A.1 Does dubbing have two shades? 282
A.2 On/Off-screen dubbing 286
A.2.1 Data and annotation 286
A.2.2 Analysis of textual, prosodic and phonetic features 288
A.3 The two shades of dubbing in NMT 293
A.3.1 Experimental setting 294
A.3.2 Results 295
A.3.3 Analysis 297
A.4 Conclusions 299

B Reproducibility checklist **301**

C List of publications **303**

List of Tables

2.1 Types of AVT, adapted from Pérez-González (2014). . . . 28

2.2 Subtitling constraints acc. to the TED guidelines. 37

4.1 Subtitle blocks (left, 1-2) as they appear in an .srt file and the processed output for obtaining aligned sentences (right). 85

4.2 Percentage of data conforming to each of the quality criteria on the corpora. 88

4.3 Number of sentences and target side words for each language of MuST-Cinema for the training, development and test set for all 7 languages. 94

4.4 Percentage of sentences containing only conforming subtitles based on the 42 and 84 CPL criterion, number of sentences containing <eol>symbols and the percentage of these sentences in the corpus (in parentheses). 95

4.5 Subtitling conformity statistics for the 7 target languages of MuST-Cinema and the source (English). 98

4.6 Examples of sentences output by the segmenter. 109

4.7 Dataset sizes in sentences for English. 112

4.8 Segmentation results on the two test sets. 112

4.9 Segmentation results on seen languages. 121

4.10 Segmentation results on *unseen* languages. 123

5.1 Results for translation quality ($BLEU_{br}$, $BLEU_{nb}$), for conformity to the subtitling constraints (CPL, CPS) and for subtitle segmentation (TER_{br}) for the four systems. 135
5.2 Average pause duration and standard deviation (in seconds) for the category without breaks (None), and for the categories with the two types of break symbols. 139
5.3 Examples of translations by the cascade (CAS) and the end-to-end model (E2E) compared to the source sentence (EN) and the reference (REF). 141
5.4 Results for quality (WER and BLEU), subtitling conformity (CLP for length, CPS for reading speed and *Sigma* for segmentation), and subtitling consistency (Structural and Lexical) for French and German model outputs. 151
5.5 BLEU, Average Lagging (AL) in ms, conformity to the length constraint (CPL) and segmentation (*Sigma*) on three language pairs of MuST-Cinema. 167
5.6 Reading speed (*rs*) mean and standard deviation in characters per second (CPS), percentage of subtitles with a *rs* of max. 21 CPS and display delay (in ms) on three language pairs of MuST-Cinema. 169

6.1 Examples of the noise inserted in each of the three experiments. 182
6.2 Segmentation scores of *Sigma* and the examined metrics for the output of four systems after projecting their boundaries to the reference. 188
6.3 Total number of interaction events and total event duration in minutes for text and spotting edits. 197
6.4 Respondents per language pair. 201

6.5 Main issues related to automatic translation, autospotting and segmentation, and number of statements. 206

A.1 Corpus statistics in utterances and source/target tokens for the three classes. 287
A.2 Mean and standard deviation on the source/target ratio of different units (characters, syllables, vowels vs. consonants and utterance duration in seconds). 289
A.3 Area under ROC curve results for classifying words as on/off (only for English input), or utterances as on/off/mixed (for all inputs). 289
A.4 Concurrent peaks between source and target segments for frequency (f0) and intensity. 290
A.5 Prosodic phrase length in mean number of phrases per segment and mean duration. 291
A.6 Percentage of phonetic class correspondence between source and target for the three classes. 'c' stands for consonant and 'v' for vowel. 292
A.7 BLEU scores for different systems for the entire test set (Total) and the three classes. 296
A.8 Perplexity on a general Spanish language model for human dubbing and two MT outputs. 298

List of Figures

2.1 The combination of signs (verbal, non-verbal) and communication channels (aural, visual) results in semiotic elements. 25
2.2 Subtitles and their categorisation acc. to (Cintas and Remael, 2021). 32
2.3 Example of a subtitle not conforming to the subtitling constraints and the same subtitle as it should be ideally displayed. 41
2.4 Two typical instances of subtitle editors. 46
2.5 Example of a subtitling template opened in a spreadsheet. 48

3.1 The Transformer model architecture. 61
3.2 Typical ST cascade architecture. 64
3.3 An example of the wait-k policy with k=2. 70

4.1 Subtitle file (top) and the full sentence with subtitle breaks (bottom). 92
4.2 Structure of MuST-Cinema. 93
4.3 Percentage of subtitles in the MuST-Cinema training data conforming to CPL$\leq$ 42 after segmenting with different models. 97
4.4 Percentage of subtitles longer than 84 characters per subtitle block for each year. 99
4.5 Percentages of the positions where breaks occur in the total breaks in each corpus. 100

4.6 Class labels for the NCRF model. 104
4.7 Accuracy of different segmentation methods based on Precision, Recall and F1, BLEU$_{br}$and length conformity (CPL). 107
4.8 Accuracy for different data mixtures of MuST-Cinema based on F1, BLEU$_{br}$and length conformity (CPL). 108
4.9 Parallel multimodal segmenter architecture. 118

5.1 Schema of the models used in the experiments. 146
5.2 Source captions (En) and Target subtitles (Fr) for the same subtitle template. 147
5.3 Example sentence displayed in the three different modes. . 159

6.1 Reference-based evaluation for perfect/imperfect texts . . . 178
6.2 Comparing n-grams counted by BLEU$_{br}$ and/or BLEU$_{nb}$. 179
6.3 Behaviour of BLEU$_{br}$ when gradually transforming the reference segmentation. 184
6.4 Linearity of BLEU$_{br}$ wrt BLEU$_{nb}$, for instances where only text was submitted to noise. 185
6.5 Values of BLEU$_{br}$ (a) and *Sigma* (b) after applying segmentation noise at different values of BLEU$_{nb}$. 186
6.7 Areas of interest for selecting the number of mouse clicks performed in each. 192
6.8 Scatterplots for the selected variables. 194
6.9 Correlation between HTER, TTE and dSpot. 196
6.10 Respondents' answers to their working settings. 203
6.11 User experience scores. 204
6.12 User experience scores for the subtitlers in Section 6.3 (TED talks). 205

A.1 Example of a translation done with lip sync and one not. . 284

A.2 An example: “No problem, robot”, of the classification task along with the annotations (s2_1_250). 287

Acknowledgements

I would like to thank...

...the reviewers of this thesis, Prof. Maarit Koponen and Dr. Evgeny Matusov, for patiently going through the 300 pages of this thesis and for their interest in improving it through their comments.

...the rest of the committee members, Prof. Rico Sennrich and Dr. Marco Guerini for accepting to take part in the defence.

...my advisor, Matteo Negri, for reminding me that there is always more space for creativity (even at 23:58 UTC-12).

...my advisor, Marco Turchi, for transmitting his enthusiasm on the topic of automatic subtitling.

...my MT group colleagues and co-authors, Luisa, Mauro, Marco, Sara, for their precious collaboration and the great team work.

...my PhD buddies and friends, Mattia, Surafel, Duygu, Amir, Dennis, Beatrice, Flor, Samuel, Samy, Rasoul, François, Jitao for sharing experiences, problems, worries and successes.

...Prof. François Yvon for what I've learnt from him during my internship in LIMSI in autumn 2021.

...Prof. Anna Matamala for the research stay in March 2022 and the discussions on media accessibility.

...the Murga Trentinerante for the colour explosions.

...the Capoeira São Salomão for keeping me away form my computer in the evenings.

...Fiddle, without which life would not have been so beautiful.

...Fabrizio, for mentoring me and knowing when to let go.

...David, for supporting me and knowing when not to let go.

...my brother Michalis, because he stills looks up at me no matter how much I mess up.

...my parents for knowing I would make the right decision once again.

Chapter 1

Introduction

1.1 The audiovisualisation of communication

The history of humankind has always been driven by technology. From the invention of the wheel to the industrial revolutions which led to our current Industry 4.0, technology is what marked these historical turning points. In our lifetime, we have witnessed how the Internet democratised knowledge and how personal computers were made a commodity for everyone's use. Since Internet's launching in 1995, the number of global Internet users has been growing at a rate of 2-4% per year[1]. As of early 2022, more than 5 billion people use the Internet, corresponding to 66% of the global population.[2] This expansion of the Internet has been facilitated by devices getting smaller and more portable. Smartphones are giving instant access to information at the tips of our fingers and completely lift spatial and temporal barriers by allowing us to reach others at any point any place.

All these technological affordances have changed the way we communicate and share information. A large part of our exchanges takes place in virtual environments, such as social media, messaging applications, video platforms. As a result, our communication has become highly digitalised

[1] https://www.internetworldstats.com/emarketing.htm
[2] https://www.internetworldstats.com/stats.htm

and our interactions are more audiovisual and multimedial than ever. This has led to an increasing audiovisualisation of our communication (Diaz-Cintas, 2013). Screens have a primary role in this process. The ubiquity of screens is seen in all spheres of public and private life. Screens have largely substituted other communication instruments, such as paper and telephones, and are now the main tool to entertain, communicate, inform, educate, do business and shop. The media products appearing on screens semiotically integrate different modalities, combining written text, images and sound.

But it is not only the media itself that is becoming more complex, but also the roles of the agents involved in media creation and consumption. Today's culture has been characterised as a culture of media convergence (Jenkins, 2006), a "*melting pot where old and new media intersect, where grassroots and corporate media collide, where the power of the media producer and the power of the consumer interact in unexpected ways*". The ease by which anyone can become media producer has led to a boom in user generated content, largely promoted by social media, while more than 500 hours of audiovisual content are uploaded every minute on YouTube.[3] The complex mediascape is composed by users, the video game industry, edutainment producers, social media networks, video-on-demand (VOD) platforms and over-the-top (OTT) services, among many other agents.

A special mention should be given to the influence of OTT services and VOD platforms on the global media distribution. OTT is used to describe internet-based services that allow users to stream media content over the internet without the need for cable or satellite subscriptions. Many of the most popular OTT services are VOD or streaming platforms such as Netflix, Amazon Prime, Hulu and Disney Plus, which gained great momentum over the past decade, sidelining the TV and broadcast industry. From an

[3]https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

economical point of view, OTT services are driving the audiovisual industry. The OTT media revenue generated in 2020 amounted to 106 billion U.S. dollars and it is expected to double by 2026, reaching over 210 billion.[4] In a 2016 Nielsen online survey conducted in 61 countries, nearly two-thirds of the global respondents (65%) said they watched some form of VOD programming.[5] Another survey on the state of online video conducted in 10 countries found that in 2020 repondents spent almost 8 hours per week watching some type of online content, an increase of 85% over the 2016 average of 4,28 hours. In 2021, users streamed 452k hours on Netflix and 694k hours on Youtube per minute.[6] Video on demand is becoming a part of daily viewing habits for a large part of the world's population.

In early 2020, the global pandemic upended the world as we knew it. The way we work, engage with others, entertain ourselves was completely transformed, since all activities moved online. This marked a technological reset (Schwab and Malleret, 2021) which accelerated the digital transformation. The amount of data generated was unprecedented. The mobility restrictions did not only have a direct impact on in-home media consumption but contributed to the establishment of remote work, which was instituted beyond jobs and tasks that involve screen work. The need to communicate with co-workers and peers gave rise to video conferencing applications, with the most popular being Zoom, Microsoft Teams, Google Meet, etc. This trend was not limited to the initial period of confinement since, in 2021, 856 minutes of webinars were hold on Zoom per minute.[7] Moreover, online education, conferences and seminars were and are still largely held online or in hybrid mode, which produces an enormous amount of hours of

[4]https://www.statista.com/statistics/260179/over-the-top-revenue-worldwide/

[5]https://www.nielsen.com/wp-content/uploads/sites/3/2019/04/global-video-on-demand-report-mar-2016.pdf

[6]https://www.domo.com/learn/infographic/data-never-sleeps-9

[7]https://www.domo.com/learn/infographic/data-never-sleeps-9

recordings. The hybrid mode seems to be here to stay, as workers prefer the flexibility and easiness provided by some amount of remote working.

All this shows that today's mediascape is a complex patchwork where types of media, agents, producers and consumers intermingle. In this zero-distance world, multilingual communication is more important than ever. The transition to an open virtual space and the technological advances have enabled access to any type of information we are looking for. But is universal access to information and communication a reality or a utopia?

1.2 Accessibility and audiovisual translation

Globalisation, digitisation and the new prospects offered by technologies seem to be democratising the access to knowledge and information through the amounts of content available and the increased flexibility to interact with it. But what happens when we do not understand the language of an audiovisual product? When we cannot hear the audio? Or when we cannot see the image? The audiovisual experience is not complete, and access to information is impeded.

The answer comes from the field of Media Accessibility (MA). Media accessibility is defined as "*access to media and non-media objects, services and environments through media solutions, for any person who cannot or would not be able to, either partially or completely, access them in their original form*" (Greco, 2019). Although MA initially focused on particular groups, especially people with disabilities, e.g. the blind, and the deaf and hard-of-hearing, it later expanded to consider not only sensorial but also linguistic barriers (Orero and Matamala, 2007). Under this perspective, ensuring access across linguistic and sensorial borders to all types of audiovisual products means modifying the audiovisual product to increase access to the aural or visual mode or to transform the verbal mode (Remael

and Reviers, 2019).

The transformation of the verbal mode to overcome linguistic barriers is the focus of Audiovisual translation (AVT). AVT is a branch of translation studies which focuses on the translation of audiovisual (AV) texts. An AV text is defined as "*a communication act involving sounds and images*" (Zabalbeascoa, 2008, p. 1). AVT hosts under its umbrella several modes, the most prominent ones being subtitling, dubbing, voice-over and interpreting. Each of these modes has its own pros and cons, and is favoured in different situations and scenarios. AVT poses challenges for the translator due to the multimodal nature of the source and the form requirements of the target. Regarding multimodality, in audiovisual products the meaning is constructed not only from the source speech, but also from the interaction of images, written text, sounds, gestures, special effects etc. Information flows in through different channels: the visual and the aural, containing verbal and non-verbal elements (Chaume, 2000). In addition, the translation is constrained by synchrony requirements (Mayoral et al., 1988). For example, in dubbing the translation needs to be synchronised with the image and the phonetics; sound signals should correspond to the visible lip movements on screen. Subtitling requires spatial synchrony in terms of the space subtitles occupy and synchrony of time, that is subtitles should correspond to the spoken signal. Lastly, AV content can take many different forms and cover many different genres with their own specificities: a sitcom is different from a news broadcast or a political debate. All this shows that AVT is a complex translation process and requires highly-qualified professionals, who posses not only translation but also technical skills (video and audio processing, file format conversion, etc).

So by modifying the audiovisual products through AVT we can provide multilingual access to information and enable multilingual virtual communication. Let's estimate how much time it would take to achieve this,

for example, for the content uploaded on Youtube. The ATAA (French association for audiovisual translation) in their guide for Subtitling and Dubbing for Film and Television (ATAA, n.d.) provide recommendations for the time that should be allowed for adaptation. According to these recommendations, a documentary or hour-long episode of a TV series (52 min.) would take 1 week to be subtitled or dubbed, while a feature-length film 2-3 weeks. This estimation is for the adaptation step alone, without including "*any preparatory work, like spotting, or post-translation work like simulations and verifications*". Of course these recommendations refer to ideal situations which are sadly far from reality. Audiovisual translators are often asked to deliver the translation in much less time.[8] Even with the most optimistic estimation of a 6:1 rate of production (6 minutes of work for each minute of video) –under continuous production settings–, it would take 3000 work minutes to adapt the content uploaded on Youtube every minute into just one language, let alone in all world's languages.

Given this disproportionate production-translation rate, providing multilingual access to audiovisual content is a Sisyphean task. We cannot expect that multilingual access can be ensured for all audiovisual products relying simply on human resources in a traditional way. The exploding demand for localisation has led some to speak of a 'talent crunch' in AVT (Estopace, 2017). Despite this claimed shortage of highly-skilled talent, the working conditions of professionals in the AVT sector are worsening, with dropping rates and always faster turn-around-times. This paradox is deteriorating the capacity shortages since few AVT professionals nowadays earn a decent wage and are forced to drop the industry (Slator, 2022).

This situation leads us to consider two main impediments in providing multilingual access to audiovisual content. Firstly, the human resources are not and will never be sufficient to translate all audiovisual products

[8] https://www.proz.com/forum/subtitling/227494-how_long_subtitling_may_take.html

which need to be adapted in different languages. Second, the precious human resources are mistreated and misused and the technology is rarely developed with the aim to support them by reducing effort spent on tedious tasks and allowing them to focus on the creative parts of their tasks.

Since human resources cannot keep up with the deluge of AV production, how do we ensure access to the audiovisual products?

1.3 Automatising audiovisual translation

A solution to the problems laid out in the previous section could be given by technology. Fully or partially automatising AVT could increase access to audiovisual products. Automatisation in AVT has mainly focused on approaches applying machine translation (MT) on a transcript of the source language dialogues. MT has demonstrated mixed results in the AVT field (Burchardt et al., 2016) and has received strong reproval and resistance by AVT professionals (Audiovisual Translators Europe, 2021). Main criticisms are its inability to deal with spoken, creative texts, and the difficulty to apply synchrony constraints on the target. However, MT is only one step in the process of automatisation. Significant human effort is still needed for transcribing the speech, condensing it to a certain length, synchronising the transcript to the video, etc., before MT is applied on this prepared text.

Recent developments in speech technologies, such as in automatic speech recognition (ASR) and speech translation (ST), have set the floor for extending automation in AVT. Improved ASR achieves more accurate transcription and recognition of speech, even under difficult noise conditions. ST, besides more robust cascaded architectures, is witnessing the proliferation of direct models, which translate directly from the audio signal into the target language. These technologies are in line with the multimodal nature of AVT and could thus be exploited to reduce the human effort in

the adaptation process.

For subtitling, the most prominent type of AVT, integration of speech technologies and machine translation would imply moving from MT for subtitling to a new paradigm, that of automatic subtitling. In MT for subtitling, the source speech is transcribed, timed and segmented into source language subtitles (or captions) by a subtitler to create a subtitle template. This template is then translated automatically with an MT system into one or multiple languages. In automatic subtitling, these preparatory steps could be substituted by speech recognition, automatic prediction of timestamps and automatic segmentation, after which MT can be applied. Skipping the template creation step would lead to significant time and cost reductions. Compared to a human template though, a performance drop is expected both in translation, due to speech recognition errors, and in the accuracy of the timestamps and segmentation, since both are determined automatically. Another disadvantage of relying on the source timestamps and segmentation (regardless of human or automatic) for generating target subtitles is that they may not be optimal for the target language, e.g. because of different syntax or an increase in the number of characters per subtitle after translating.

Direct automatic subtitling solutions, inspired from recent advances in end-to-end speech translation (ST), could provide target language subtitles directly from the video, without the need for intermediate representations. Direct approaches avoid error propagation and allow for end-to-end training by jointly learning all components, which may lead to better performance compared to learning each component separately, as it happened with neural MT and statistical MT. In addition, speech as input signal contains richer information than text (prosody, pauses, etc.). This information, even though essential for subtitling, is lost in the transcription process. Lastly, a direct approach would not be restricted by the source

language and could better model the timestamps and segmentation based on the needs and norms of the target language. For these reasons, the focus of this thesis is to take advantage of the promising elements of direct approaches for promoting automatic subtitling.

This new paradigm of extended automation in subtitling can help tackle the problem of lack of multilingual access in two ways: First, through **full automation**, by providing subtitles in a fully automatic way, e.g. for gisting purposes. The transition to models generating subtitles from the audio without any human intervention in a fully automatic way can provide access to products that would not have been translated otherwise (e.g. user generated content). Second, through **augmentation**, by providing support to human translators in AVT. This consists in the integration of speech and ST technologies in subtitling tools tailored to their needs, in a way that reduces the effort from repetitive, tedious tasks and allows them to focus on creative aspects. On the other side of the coin, the automatisation of AVT can benefit research in language technologies. Because of the specific challenges related to the translation of AV texts, AVT can serve as a new test bed for pushing the limits of MT and ST, leading to novel methods and architectures.

1.4 Innovative aspects and contributions

This thesis addresses the automatisation of subtitling from a computational perspective, with the aim of obtaining target language subtitles directly from the source audio. This marks a transition to a new paradigm of fully automatic subtitling compared to 'old style subtitling'. All computational stages are covered. Since the development of automatic systems requires high-quality task-specific data, the first contribution is the compilation of suitable subtitling data containing the source speech. Building on this first

prerequisite, the second contribution is the development of end-to-end architectures for three popular subtitling scenarios: offline subtitling ('traditional' subtitling, pre-recorded), dual subtitling (joint generation of source language captions-target language subtitles) and live subtitling (real-time creation of subtitles). The progress in a field cannot be marked without appropriate evaluation metrics and methods. Therefore, the third contribution of the thesis focuses on specific metrics taking into account the formal requirements of subtitling and a comprehensive study of post-editing effort and subtitlers' views on automatic subtitling. The main contributions of this thesis can be categorised under three topics: resources, architectures and evaluation.

Resources: We provided two resources for promoting research in automatic approaches to subtitling.

- The MuST-Cinema corpus (Karakanta et al., 2020b): The only publicly available speech-to-subtitles corpus, containing 250 hours of speech, source language captions and target language subtitles in seven language pairs.[9]

- MuST-Cinema-PE (Karakanta et al., 2022b): The first corpus containing product-, process- and participant-based data collected from post-editing automatic subtitles for two language pairs of the MuST-Cinema test set.

Systems: We experimented with the application of ST systems in the domain of subtitling, understanding their potential for subtitling-related tasks. Specifically, we have investigated:

- Neural models for subtitle segmentation under different data conditions and their feasibility for synthesising new speech subtitling data (Karakanta et al., 2020b,c; Papi et al., 2022)

[9]https://ict.fbk.eu/must-cinema/

- The first end-to-end speech translation system for subtitling and a comparison between machine translation, cascade and end-to-end systems for the generation of subtitles (Karakanta et al., 2020a)
- The implementation and evaluation of architectures for generating dual subtitles (Karakanta et al., 2021a)
- Simultaneous ST systems for live subtitling and an exploration of alternative output visualisation methods (Karakanta et al., 2021b)

Evaluation: We promoted the evaluation of subtitling-specific aspects as well as the methodology for conducting user-based evaluations in automatic subtitling. Our contributions here include:

- A review of methodological issues in the evaluation of machine translation for subtitling and a reproducibility checklist for experimental research in automatic subtitling (Karakanta, 2022)
- EvalSubtitle: A systematic comparison of metrics for evaluating subtitle segmentation and a toolkit to compute them (Karakanta et al., 2022c)
- Sigma: A new metric for evaluating subtitle segmentation for end-to-end subtitling systems (Karakanta et al., 2022c)
- A broad survey of subtitlers' views and experience with post-editing automatic subtitles, leading to a guide with recommendations for the improvement of future automatic subtitling tools (Karakanta et al., 2022a)

1.5 Structure of the thesis

The thesis is structured in seven chapters. Chapter 1 contains the Introduction, which lays out the context, a short description of the problem and the proposed solution, its innovative aspects, structure and other

information on publications and activities throughout the three years of the PhD. Two background chapters follow, which present the necessary information to guide the reader through the thesis: Chapter 2 discusses the challenges of AVT with a specific focus on subtitling and Chapter 3 describes the workings of neural machine translation, speech translation and their evaluation. Moving to the main contributions of the thesis, the chapters follow a tripartite structure, describing the ingredients for approaching subtitling from a computational perspective: data, models and evaluation. Chapter 4 reviews existing subtitling corpora and their limitations for developing automatic subtitling solutions, the compilation of the necessary speech subtitling corpora, and automatic subtitle segmentation, both from an application side and as a method for subtitling data augmentation. Chapter 5 presents direct automatic subtitling models for offline subtitling, for the generation of dual subtitles and for live subtitling. The last contribution chapter, Chapter 6 is dedicated to the evaluation of subtitle-specific aspects. After proposing new metrics for evaluating the segmentation, we analyse the effort in post-editing automatic subtitles and the subtitlers' user experience, proposing roadmaps for promoting reproducibility in research at the crossroads between AVT and MT and for improving future automatic subtitling tools. Chapter 7 concludes the thesis, offering a view into the open challenges for the future of automatic subtitling. Although not the core of this thesis, Appendix A contains one chapter on the exploration of the on/off-screen speech in dubbing and a study on screen-value-aware MT for dubbing.

A note on the use of personal pronouns for the –possibly– perplexed reader: Parts of the thesis where 'we' appears refer to work conducted in collaboration with other researchers, while 'I' refers to individual work (such as the writing of this thesis) or opinions not necessarily endorsed by other collaborators.

1.6 Activities and dissemination

During the PhD journey, I was lucky enough to have collaborated with researchers from various fields and specialisations in Translation Studies and Machine Translation. However, not all the research resulting from these collaborations can form an integral part of this thesis. For this reason, I include an Appendix with a full list of the publications achieved during these three years (Appendix C), divided by topic. A list of selected publications appearing in the thesis is presented here:

- Alina Karakanta. 2022. Experimental research in Automatic Subtitling: At the crossroads between Machine Translation and Audiovisual Translation. *Translation Spaces*. DOI: https://doi.org/10.1075/ts.21021.kar
- Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. Post-editing in automatic subtitling: A subtitlers' perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 259–268, Ghent, Belgium. European Association for Machine Translation.
- Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. Towards a methodology for evaluating automatic subtitling. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 333–334, Ghent, Belgium. European Association for Machine Translation.
- Alina Karakanta, François Buet, Mauro Cettolo, François Yvon. 2022. Evaluating Subtitle Segmentation for End-to-End Generation Systems. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. Marseilles, France, pp. 3069–3078.
- Alina Karakanta, Sara Papi, Matteo Negri, Marco Turchi. 2021. Simultaneous Speech Translation for Live Subtitling: from Delay to Display. In

Proceedings of the 1st Workshop on Automatic Spoken Language Translation in Real-World Settings (ASLTRW), pages 35–48, Virtual. Association for Machine Translation in the Americas.

- Alina Karakanta, Marco Gaido, Matteo Negri, Marco Turchi. 2021. Between Flexibility and Consistency: Joint Generation of Captions and Subtitles. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 215–225, Bangkok, Thailand (online). Association for Computational Linguistics.

- Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, Marco Turchi. 2021. Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. August 2021. Association for Computational Linguistics.

- Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. 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.

- Alina Karakanta, Matteo Negri and Marco Turchi. 2020. Point Break: Surfing Heterogeneous Data for Subtitle Segmentation. In Proceedings of the *Seventh Italian Conference on Computational Linguistics*, CLiC-It. Bologna, Italy.

- Alina Karakanta and David Orrego-Carmona. In press. Subtitling in Transition: the case of TED talks. *Translation in Transition: Bridging Human and Machine Intelligence*, American Translators Association series. John Benjamins.

- Alina Karakanta, Matteo Negri and Marco Turchi. 2020. MuST-Cinema: a Speech-to-Subtitles Corpus. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May 13-15 2020.

- Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. Towards Automatic Subtitling: Assessing the Quality of Old and New Resources. *IJCoL – Italian Journal of Computational Linguistics* vol. 6, n. 1.

- Alina Karakanta, Matteo Negri, Marco Turchi. 2019. Are Subtitling Corpora Really Subtitle-like? In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-It)*, Bari, Italy.

I always considered important that researchers are not isolated in front of their computer but stay in contact with the real-life applications and uses in the field where their research is applied. Even more so in my case, because of the blending of the fields of Translation Studies, Audiovisual Translation and Machine Translation. For this reason, the dialogue between these disciplines was one of the goals of my thesis work. This dialogue was promoted through dissemination of the state of the art to potential users and through interdisciplinary publications and collaborations. Below I list the activities undertaken towards these goals during the PhD.

EAMT Grant I was recipient of an European Association for Machine Translation (EAMT) grant under its programme "2021 Sponsorship of Activities - Students' edition". The project, titled "Towards a methodology for evaluating automatic subtitling", aimed at collecting subtitle post-editing data in a real use case scenario where professional subtitlers edit automatically generated subtitles. Based on the collected data, the overarching goal of the project is to investigate automatic subtitling, the factors contributing to post-editing effort and their relation to the quality of the

output. The grant resulted in a special release of MuST-Cinema, with additional references and productivity data (keystrokes, post-editing time measurements etc.), and a publication at the EAMT 2022 conference.

Collaborations Despite the limited possibilities for onsite interactions during the pandemic, I had the pleasure to secure collaborations with several researchers. The first collaboration with Prof. Timo Baumann from the University of Hamburg in late 2019 aimed at exploring differences between on/off-screen dubbing and the application of synchrony constraints in MT depending on the degree of synchrony required. This collaboration resulted in two publications at COLING 2020 and ICMI 2020. On the same topic, Erick García Chávez from the University of Trento conducted a four-month internship with our group in 2021, laboriously working on discovering prosodic and phonetic patterns distinguishing between on/off-screen dubbing. He presented this work at UCCTS 2021.

The compilation of our speech subtitling corpus based on TED Talk subtitles made possible a collaboration with Prof. David Orrego-Carmona from Warwick University on non-professional subtitling in TED talks. We conducted a corpus-based investigation of particular characteristics of TED subtitles and the effect of subtitling tools on the translated subtitles, resulting in a chapter to appear in a book of the American Translators Association series (John Benjamins).

During the COST LEAD-ME 2021 Summer School "Eye tracking in media accessibility research - methods, technologies and data analyses" in July 2021, a group was created to work towards a final project. The participants were Anita Kwiatkowska, Vilelmini Sosoni, Rocío Varela, Claudia Wiesinger and I and the title was "Fully automatic vs. human subtitling: comprehension, user experience and perception". We used eye-tracking and questionnaires to respond to whether fully automatic subtitles require

more processing effort on the part of viewers than human subtitles and to test the impact of fully automatic subtitles on perception and comprehension. Even though the study was just a pilot and therefore the results would not be suitable for publishing following high-research standards, this work is the first of its kind and will hopefully set the basis for more reception studies on automatically generated subtitles.

In autumn 2021, I completed a two-month research stay in Laboratoire Interdisciplinaire des Sciences du Numérique (LISN) at Université Paris-Saclay with Prof. François Yvon, under a European Commission funded project "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (grant #820437). The aim of the stay was to study metrics for evaluating the automatic generation of subtitles for TV programmes, with a particular focus of evaluating subtitle segmentation. The stay was concluded with a publication at LREC 2022 and the release of EvalSubtitle, a toolkit featuring a collection of metrics for evaluating subtitle segmentation.

Invited talks, presentations and lectures As part of the dissemination, I took part in the following activities, reaching out to various research communities, the users of AVT technologies and students.

- *Automatic subtitling: a hands-on session.* Workshop for the European Masters students at Ghent University (EM-TTI visiting scholars grant). 26 October 2022.
- *Post-editing of speech translation output.* Lecture for the European Masters students and scholars at Ghent University (EM-TTI visiting scholars grant). 24 October 2022.
- *Post-editing in automatic subtitling: Auto-spotting in the spotlight.* Presentation at Translation in Transition 6. Prague. 23 September 2022.

- *Post-editing in automatic subtitling: students' and subtitlers' perspectives.* Invited lecture for the MA subtitling course at the University of Warsaw. 30 May 2022.

- *Automatic subtitling: Post-editing effort and subtitlers' perspectives*: Invited talk at Translating Europe Workshop - AVT workflows and the role of automation. Rome (Online). 24 May 2022.

- *Automatic subtitling: towards a new paradigm*: Invited talk at TransMedia Catalonia International Meeting - Universitat Autónoma de Barcelona. 18 March 2022.

- *Dubbing through the MT-glass and What We Found There*: Invited lecture Setmana de la TAV. TRAMA group, Uni Jaume I. 18 December 2021.

- *Automatic subtitling: Towards a new paradigm*: Invited talk at the STL Seminar. LIMSI, Paris-Saclay. 9 November 2021.

- *Subtitles of tomorrow: More Machine- than Human-translated?*: Presentation at the Palermo Translation Symposium Audiovisual Translation and Computer-Mediated Communication: Fostering Access to Digital Mediascapes. University of Palermo. 7-8 October 2021

- *Subtitling and the Machine: The place of subtitlers in automatic subtitling workflows*: Presentation at IATIS conference. Universitat Pompeu Fabra Barcelona. 14-17 September 2021.

- Alina Karakanta, Anita Kwiatkowska, Vilelmini Sosoni, Rocío Varela, Claudia Wiesinger. 2021. *Fully automatic vs. human subtitling: comprehension, user experience and perception.* Project presentation LEAD-ME Summer Training School Warsaw 2021. 5-9 July 2021.

Chapter 2

Audiovisual Translation

This chapter is an introduction to Audiovisual Translation (AVT) detailing the background knowledge required for understanding AVT as a particular type of translation and the challenges it brings for translators and machines. It presents main research directions of this rich translation field in the past 30 years, focusing on subtitling as the core subject of this thesis. Even though the first conceptions of AVT can be found in some seminal works in the seventies and eighties, such as the Fodor's (1976) contribution to dubbing, it was not until the nineties that research contributions marked the inception of AVT theory. AVT draws from the disciplines of film and media studies, translation studies, sociology, cognitive studies, and cultural studies. All this shows the complexity of studying AVT, which brings under its pulsing domain of enquiry a variety of neighbouring research fields.

This chapter takes a view of AVT from the perspective of translation studies and does not refer to technological advances, which are detailed in the next chapter. The first part describes AVT and its field of study (Sec. 2.1), the translation challenges brought about by multimodality (Sec. 2.1.1) and the various types of AVT (Sec. 2.1.2). Then, the reader can delve into the specific characteristics of the prominent AVT type explored in this thesis, subtitling (Sec. 2.2).

2.1 Audiovisual translation and its remit

AVT has gone through successive stages of development before being established as one of the most vibrant and dynamic fields of translation studies (Díaz-Cintas and Massidda, 2019). Before reaching a consolidation under the term AVT, a plethora of terms were used in the past to describe the translation of audiovisual texts: 'film translation' and 'film and TV translation', closely-bound to the channel through which audiovisual (AV) products were transmitted at the time (Snell-Hornby, 1988; Delabastita, 1989), or 'constrained translation' (Mayoral et al., 1988) which emphasised the medial constraints, e.g. maximum length. Later, the proliferation of different types of digital media was accompanied by other nomenclatures, such as 'multimedia translation' (Gambier and Gottlieb, 2001), 'transadaptation' (Gambier, 2003), or 'screen translation' (O'Connell, 2007; ESIST). Today, AVT is recognised as a mature discipline drawing from practical and theoretical aspects.

Audiovisual Translation is defined as "*a branch of translation studies concerned with the transfer of multimodal and multimedial texts into another language and/or culture*" (Pérez-González, 2008). From this definition we can draw specific elements that make AVT a special type of translation and the challenges it poses for the translator. First, the content to be translated is *multimodal*. This means that it deploys multiple semiotic modes, such as language, image, music, perspective. Semiotics is the study of signs, and a sign is anything that conveys meaning. The study of signs is particularly relevant for AVT. Translation cannot be seen only as a transfer of words into another language, but as a transfer, and possibly interpretation, of a compound message built up from multiple signs.

Second, AV content is *multimedial*, in the sense that it is delivered to the viewer through several media (Pérez-González, 2008, 13). To reach our

senses, AV content, such as films, educational videos, meeting recordings etc., travels through speakers and screens at the same time, and appears on different types of devices, from cinema screens to smartwatches. The medium and device through which the message is transferred impose particular constraints and restrictions on the translation. For example, the size of the screen will determine the length of the translation and the way it is displayed. The fact that a message is perceived through image and language simultaneously requires them to be presented in synchrony and under time restrictions. It is thus left to the ability of the translator to convey a multimodally-rich message but in a spatially and temporally constrained context. In this case, the optimal translation solution is not simply the one that maximises adequacy and fluency, but also conformity to external constraints posed by each scenario.

Third, the transfer is performed not only across *languages* but also across *cultures*. Priority is therefore given to the pragmatic rather than the referential level of the language, to the communicative intention rather than the textual and semantic details. AVT is concerned with linguistic and extra-linguistic elements, and is affected by the cultural norms of the target community. One critical point is the translation of culture-specific elements (Newmark, 1988) or, in the context of AVT, extralinguistic culture-bound references (Pedersen, 2005), present in the source text. The lack of an obvious official equivalent for these references constitutes a translation problem. The translator needs to actively decide whether to mirror source text elements or to adapt them to the target preferences, expectations and culture codes. The strategies vary from completely literal renditions to utter recreations. This choice towards domestication or foreignisation is a classic dilemma in translation studies (Venuti, 1995). Apart from this, the norms of the target culture have an impact on the viewers' reception of humour (Zabalbeascoa, 1996; Antonini and Chiaro,

2005), offensive language (Fernández Fernández, 2009), politeness, dialectal and linguistic variation (Rittmayer, 2009). All in all, multimodality, multimediality and a strong presence of cultural elements are three elements that make AVT a particularly challenging translation area.

Research in AVT is just as industry-oriented as it is academy-oriented. In a later definition, Pérez-González (2014) mentions that "*Audiovisual translation focuses on the practices, processes and products that are involved in or result from the transfer of multimodal and multimedial content across languages and/or cultures*". This definition notes the wide remit of AVT as the study of *practices, processes and products*, each of them corresponding to a category of audiovisual translation models: causal, process and comparative (Pérez-González, 2014). Causal models attempt to determine the reasons why a translation is the way it is. This includes the study of norms in the AVT industry, discourse and ideology, and the impact of different systems (Even-Zohar, 1979). Process models draw from psycholinguistic, cognitive and neurolinguistic approaches to study how the translator's mind works during the translation process. The last category is comparative models, which are concerned with the comparison of features and the correspondence between the source and the target texts. These models focus on the AVT product through the study of translation strategies and corpus-based approaches. In this thesis, we study the application of speech technologies and machine translation in AVT, drawing from all three categories of AVT models (see Chapter 6 on evaluation).

2.1.1 Multimodality

In the previous section, I referred to multimodality as one of the factors posing translation challenges in AVT. This section describes multimodality in detail and how the transfer of multimodal content is performed in AVT.

Multimodal texts are defined as "*texts which combine and integrate the*

meaning-making resources of more than one semiotic modality – for example, language, gesture, movement, visual images, sound and so on – in order to produce a text-specific meaning" (Thibault, 2000, p. 311). Audiovisual texts are thus typical instances of multimodal texts. In audiovisual communication, meaning is carried through several semiotic modes, which sometimes function complementary or in isolation. For example, the phrase *Let's go* can be complemented by the actor exiting the room. On the contrary, parts of a video without dialogue still convey meaning through images or sounds, a notable example being silent films. Today's AV products make use of a rich gamut of meaning-making resources.

Meaning in audiovisual communication is conveyed through two signs and two channels: through verbal and non-verbal signs, and through the aural and the visual channel. Figure 2.1 shows how the combination of signs and channels leads to four semiotic elements, which compose the message in audiovisual texts (Delabastita, 1989; Zabalbeascoa, 2008):

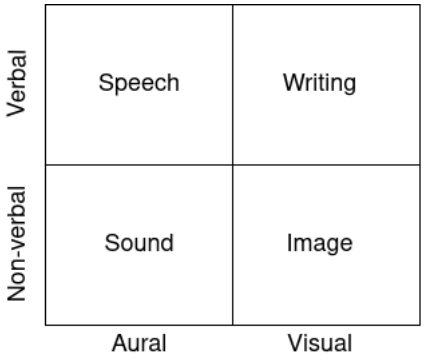


Figure 2.1: The combination of signs (verbal, non-verbal) and communication channels (aural, visual) results in semiotic elements.

- **Image** (the visual-nonverbal channel): moving or static pictures, colours, perspective, photography, gestures.
- **Writing** (the visual-verbal channel): includes text appearing on screen, regardless of whether it appears in the form of images, e.g. displays,

road signs, banners, letters, messages on phones and computer screens, newspaper headlines etc., or as written text inserted in post-production (captions, credits, subtitles, notices).

- **Sound** (the aural-nonverbal channel): music, sound effects, noises.
- **Speech** (the aural-verbal channel): dialogues, narration, monologue, songs, voice-off (off-camera speech).

The construction of the complete meaning from different modes is what makes AVT a particular type of translation. Equivalence is not achieved through a mere correspondence between written signifiers (text) but between the interpretation of a semiotically complex sign in the source-target text. The translator deals with the way words interact with sounds and images to form cohesion, intertextuality and other features of structure and meaning. This requires a semiotic approach, where translator can deviate from "*the semantic meaning of the ST [source text] by creating a 'new' target text which is nevertheless significationally equivalent to the original*" (Kruger, 2001, p. 185).

To fully grasp the semiotic nature of AVT as a translation type, it is necessary to refer to the seminal work of Jakobson (1959), who distinguished three ways of translating a sign. *Intralingual translation* (or rewording) is defined as the interpretation of verbal signs through other means in the same language. For example, intralingual translation is performed when rewriting a text into easy-to-understand language. This process includes simplification and visual presentations to make information easy to understand by everyone. *Interlingual translation* (or translation proper) is the interpretation of verbal signs through means in another language. It confers the classical notion of translation with the linguistic transfer from one language to another. *Intersemiotic translation* (or transmutation) is the third way of translating a signal, and consists in the interpretation of

verbal signs through non-verbal means (or vice-versa). Intersemiotic translation requires a transfer from one semiotic mode to another. For example, sign language interpreting involves interpreting the verbal signs (speech) into visual signs.

All these ways of translating a message manifest themseves in AVT. AVT can be polysemiotic, isosemiotic or diasemiotic, depending on whether the translation uses the same semiotic channel as the original (Gottlieb, 1998). Subtitling, for instance, is diasemiotic, because the verbal signs are transformed into written text on screen in the form of subtitles. Dubbing, on the other hand, is isosemiotic; both the original and the translation are represented through the aural channel.

The inability of current machine translation systems to take into account the multimodal and multisemiotic nature of AV texts is one of the main criticism points against the adoption of machine translation in AVT. In this thesis, we propose automatic subtitling approaches based on speech translation. Contrary to previous approaches to automatising AVT which incorporated only the semiotic element of writing, our methods extend the semiotic elements considered to writing and speech. In this way, we make a first step towards automatically addressing multimodality in AVT texts.[1]

2.1.2 Types of AVT

For most laypeople, the notion of audiovisual translation brings to mind subtitling and dubbing. However, there is a plethora of AVT types, accommodating different needs and addressing different target audiences. Several typologies of AVT modes have been proposed during the years. Among the first typologies was the one proposed by Gambier (2003), who identified ten AVT types and proposed a division by dominant and challenging types.

[1]We acknowledge the importance of the visual channel, however, due to its relation to the field of computer vision, incorporating elements from the image is out of the scope of this thesis.

Since then a lot of research has gone into AVT and the boundaries between dominant and challenging types have become less distinct. A later categorisation (Pérez-González, 2014) proposed dividing AVT types into three main groups: subtitling, revoicing, and assistive forms of AVT. Matamala (2021a) took an integrated approach between AVT and media accessibility where assistive forms do not form a category of their own, by proposing a categorisation into sound modalities and visual modalities. Sound modalities are the ones where the product is delivered in spoken form (dubbing, voice over, interpreting, audiodescription, audio subtitiling), while visual modalities include subtitling and sign language interpreting.

This thesis does not focus on assistive forms of AVT, therefore we adopt here the categorisation of Pérez-González (2014), to provide the reader with a clearer distinction between subtitling and other AVT forms. Table 2.1 shows the forms of AVT based on this typology.

Subtitling	Revoicing	Assistive
Interlingual	Dubbing (lip-synchronised)	Audio description
Intralingual	Voice-over	SDH
Respeaking	Narration	Sign language interpreting
	Simultaneous interpreting	Audio subtitling
	Free commentary	

Table 2.1: Types of AVT, adapted from Pérez-González (2014).

In terms of its linguistic properties, **subtitling** can be interlingual or intralingual. Interlingual subtitling is a written rendition of the source speech in the target language, therefore it involves translation from the source to the target language. Intralingual subtitling entails only a shift from oral to written mode, since the captions are in the same language as the speech. In contrast to offline subtitling which is produced in advance, the increase in live events has given rise to another type of subtitling, live

subtitling, produced through the technique of respeaking (Romero-Fresco, 2011). Live subtitling was initially intralingual, aimed at providing access to people with hearing impairments, but recently the growing need for subtitles in multilingual events has led to interlingual respeaking becoming established as a new profession to produce interlingual live subtitles. A broader description of the types of subtitles is given in Section 2.2.1.

Revoicing (Luyken et al., 1991) is a general term to describe a variety of spoken translation modes, including dubbing, voice-over, narration, free commentary and interpreting. Dubbing consists in the replacement of the original soundtrack with a new soundtrack in the target language. It is mainly limited to film and TV productions where dubbing is used to create the illusion of original dialogues. For this reason, dubbing requires the translation to have the same duration as the source utterance and synchronisation with the lip movements of the actors when the face is clearly visible on screen. On the contrary, in voice-over the original and translated soundtrack coexist. The volume of the original soundtrack is faded away a few seconds after its start and the translated soundtrack is overlapped. The coexistence of original and translated version, together with the possibility of a fuller translation, makes voice-over suitable for programmes not requiring a realistic effect, such as interviews and documentaries. Voice-over is often used to translate films in small markets with limited budgets. Narration provides a prepared, translated and possibly condensed version of the original speech, which is then read out, respecting the synchronisation with the visual information presented. Free commentary is an adaptation of a programme for "*a new audience, with additions, omissions, clarifications and comments*" (Gambier, 2003, p. 174). As with narration, synchronisation follows the image rather than the soundtrack, but the delivery is more spontaneous and catered to the needs of the target audience. Narration and free commentary are more common in children's

programmes and documentaries. Simultaneous interpreting or sight interpreting is used in live events, when time and fund constraints do not allow for another type of translation.

Assistive forms of AVT include audio description (AD), subtitling for the deaf and hard-of-hearing (SDH), sign language interpreting and audio subtitling. In AD, visual elements of the product are orally described and inserted in the soundtrack in a way that does not overlap with the dialogues and other sound effects for blind and visually impaired persons (Matamala, 2021b). Audio subtitles are voiced subtitles and often interwoven with AD (Matamala, 2022). SDH is a written rendition of the source speech in the source or target language for persons with auditory impairments (Szarkowska, 2020). The difference with subtitling is that aural non-verbal components (sound effects, music) are also rendered in a written form. Sign language interpreting is another assistive form for DH persons.

A sibling discipline to AVT is video game localisation. The audio assets in video games can be dubbed (a process called voice-over in the video game industry) and/or subtitled. Even though video game localisation and AVT have some language transfer modes in common, the norms applied in localisation are not the same as in AVT, since the expectations and needs of gamers differ from those of viewers (O'Hagan, 2019).

All these modes of AVT have blurred the traditional borders between translation and interpreting, and between written and oral code. Different forms of audiovisual translation and other translation modes are converging, creating new hybrid forms and sometimes catering for very different and well-defined target audiences. Proposals to automatise AVT should thus take into consideration the particular characteristics of each AVT type. In the next sections, we delve deeper into the particular characteristics of subtitling as the scope of enquiry of this thesis.

2.2 Subtitling

Subtitling could be possibly considered as the most prominent AVT type. Subtitles are short snippets of timed, written text superimposed on the video that convey a target language version of the source speech (Pérez-González, 2014). According to Díaz-Cintas (2020), "*subtitling may be described as a translation practice that consists of rendering in writing, usually at the bottom of the screen, the translation into a target language of the original dialogue exchanges uttered by different speakers, as well as all other verbal information that appears written on-screen (letters, banners, inserts) or is transmitted aurally in the soundtrack (song lyrics, voices off)*". This implies the cohabitation of the original text in spoken form, the original image, and the target text in written form, the latter adding an extra layer of information. Subtitling is a constrained type of translation. Subtitlers need to achieve a balance between the visual and aural dimension and leave enough time for the viewers to focus on the image while reading the text.

2.2.1 Types of subtitling

Subtitles take up multiple roles and serve a variety of purposes in our society. It is thus no wonder that there are so many different types of subtitles. Several parameters for their classification have been proposed (Luyken et al., 1991; Carroll and Ivarsson, 1998; Bartoll, 2004) but, due to the velocity of change in AVT, some of these parameters lose their importance while others come to life. We follow the most recent classification (Cintas and Remael, 2021), which groups subtitles according to the following criteria: (1) linguistic, (2) time available for preparation, (3) display mode, (4) technical parameters, (5) methods of projection, and (6) medium of distribution. Figure 2.2 shows the classification of subtitles.

Based on their linguistic dimension, subtitles can be intralingual or in-

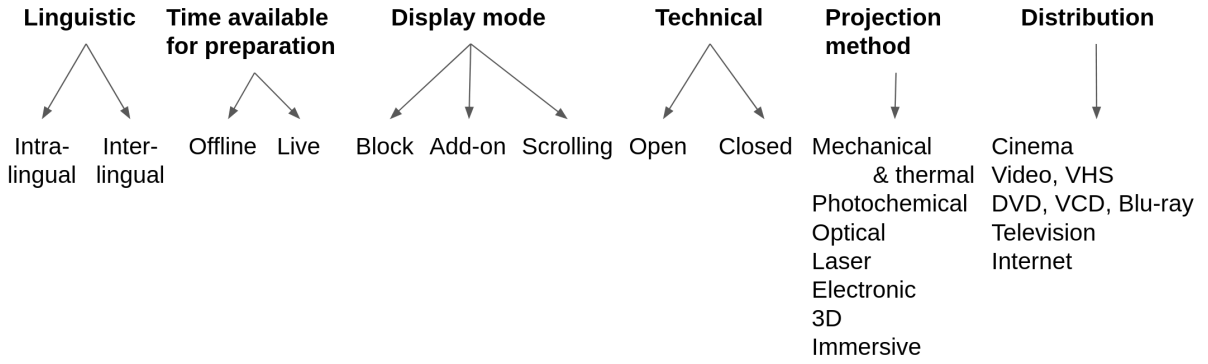


Figure 2.2: Subtitles and their categorisation acc. to (Cintas and Remael, 2021).

terlingual. A note on terminology: In this thesis I use the term 'captions' to refer to the product of intralingual subtitling and 'subtitles' for the product of interlingual subtitling. Intralingual subtitles (or captions) are in the same language as the source speech and involve a shift from oral to written form. They are used when the source speech is difficult to discern, for example with bad quality recordings, dialects and accents, or for watching material in silent mode. Interlingual subtitles involve translation from the source language to one or multiple target languages, along with a shift from oral to written. Both types of subtitles can have a didactic purpose, meaning that they aid language learners understand the speech. Captions have been found to increase learners' L2 vocabulary (Sydorenko, 2010) and improve listening comprehension (Guichon and McLornan, 2008). Subtitles in the learners' native language (L1) are an indispensable tool for comprehension and access to information, especially for beginners. Subtitles and captions have an inclusive function when they are tailored to the needs of the deaf and hard-of-hearing (SDH).

Subtitles can further be monolingual, bilingual or, in few cases, even multilingual. In bilingual subtitles different language versions of subtitles co-exist, with each version occupying one line of text on the screen (Gambier, 2003). They appear in bilingual communities and countries, such as

Belgium (French and Flemish Dutch) and Finland (Swedish and Finnish). Recently, however, captions along with subtitles have been employed in other countries to attract wider audiences. For instance, in Mainland China English captions are displayed along with Mandarin subtitles. One group which undoubtedly benefits from the parallel presence of captions and subtitles are language learners. In bilingual subtitles, the captions support learners in understanding the speech and acquiring terminology in their L2, while subtitles serve as a dictionary, facilitating bilingual mapping (García, 2017). Interestingly, despite doubling the amount of text that appears on the screen and the high redundancy, it has been shown that bilingual subtitles do not significantly increase users' cognitive load (Liao et al., 2020). In Section 5.3, we propose automatic subtitling models which produce bilingual subtitles.

Subtitles can be pre-recorded or live. Pre-recorded or offline subtitles are prepared after the AV content has been shot and before its release, while live (or online) subtitles as prepared as the content is broadcast. In offline subtitling, subtitlers have the time to carefully prepare the transcription/translation, to synchronise it with the speech and segment it based on linguistic rules (for more details on subtitling constraints, see Section 2.2.2). Live subtitling allows for obtaining subtitles in real time and is recently witnessing an upsurging demand in a range of occasions; from news and TV programmes, to online meetings, conferences, live events, shows and university lectures. Live subtitles are a simpler and more customisable alternative to other ways of translating speech, such as simultaneous interpreting (Marsh, 2004). They were produced initially with standard keyboards, but the need to reduce latency led to resorting to stenography or to the invention of a customised syllabic keyboard, called Velotype[2]. Another method to produce live subtitles is the dual keyboard system, which

[2]https://www.velotype.com/en/homepage-eng/

employs a group of subtitlers working in collaboration. With the adoption of automatic speech recognition (ASR) technologies, respeaking became the most popular technique for live subtitling (Lambourne, 2006). With this technique, a respeaker listens to the original sound of a (live) event and respeaks it to an ASR software, which turns the recognised utterances into subtitles on the screen (Romero-Fresco, 2011). Originally, respeaking was used to produce intralingual subtitles, especially to enable the deaf or hard-of-hearing to follow live TV programs (Lambourne, 2006). With the widening definition of accessibility to include persons without full access to the audiovisual content, such as new immigrants, older adults and language learners, respeaking is used to provide interlingual subtitles. One difference with simultaneous interpreting is that respeakers repeat the speech using a 'robotic voice' to increase ASR accuracy and enunciate necessary punctuation (Dawson, 2019). In its interlingual mode, live subtitling is a newly established practice and thus the industry is experimenting with different profiles for interlingual live subtitlers (Pöchhacker and Remael, 2020). Interlingual live subtitling requires skills from three disciplines: respeaking, subtitling and simultaneous interpreting. Driven by the growing needs in real-time multilingual communication, in Section 5.4, we experiment with simultaneous speech translation for automatising live subtitling.

As far as display mode is concerned, display in blocks is the standard display mode in offline subtitling, where subtitles are prepared beforehand. A full subtitle block is displayed with the text organised in one or more lines appearing together. This mode produces subtitles that are easier to read (Ofcom, 2005). Add-on or cumulative subtitles appear one line at a time, usually when two speakers are involved in the event. They produce a dramatic effect by not revealing information before it is spoken by the speaker, e.g. in the case of a joke. Scrolling, or roll-up subtitles, are the preferred display mode for live subtitles, where a small delay is a

desideratum. They can be displayed in blocks, line-per-line or word-for-word, the latter being used in most automatic simultaneous translation applications (Ma et al., 2019; Bojar et al., 2021). However, eye-tracking research has shown that a word-for-word display is not optimal, since the presence of the word to the right of fixation is vital for fluent reading (Rayner et al., 2006) and its absence leads to more re-reading (Sharmin et al., 2016). In a line-per-line display, every time a line is filled, the previous line moves to the upper row and the new line occupies the lower row. Because of their high level of instability, this type of subtitles makes reading more difficult and leads to higher cognitive load and more time spent looking at the text rather than the image (Ofcom, 2015; Romero-Fresco, 2010; Rajendran et al., 2013).

As regards the technical parameters, open subtitles are not encoded into the video signal but burned on the image and as a result, cannot be turned off. On the contrary, closed subtitles are overlaid on the video and can thus be turned on and off. The higher personalisation required when watching AV content today has made closed subtitles the norm (Díaz-Cintas, 2020, 26–27). Other taxonomies for classifying subtitles include the technical aspects of methods to project them on the video, as well as the medium through which they are distributed. These technical and commercial aspects are not analysed here since they do not affect the technology presented in the thesis.

2.2.2 Subtitling constraints

As we have seen in previous sections, subtitles are written text snippets representing the source speech and possibly including elements from the visual image. The coexistence of image, sound and text creates medial restrictions on two axes; the spatial and the temporal. These medial restrictions brought the need for establishing conventions to ensure easy in-

formation processing and a pleasant user experience. Conventions mainly focus on the length of subtitles, their duration and their layout, and aim at standardising the form of subtitles. Notable early attempts to provide subtitling guidelines are the *Code of good subtitling practice* by Carroll and Ivarsson (1998), also endorsed by the European Association for Studies in Screen Translation (ESIST), and the subtitling guidelines of Karamitroglou (2001). Later, language-specific guidelines were progressively proposed by professional associations in several countries.[3] Besides guidelines focusing on the specificities of each target language, company-specific guidelines exist. The Netflix guidelines[4] now serve as reference for a wide range of projects. TED Talks, a collaborative translation project, has also released guidelines[5] to ensure the quality of the subtitles produced by its volunteers.

The scope of this thesis largely focuses on corpora created from TED Talk subtitles (Sec. 4), as among the very few subtitling resources where the source audio/video are available and published under a Creative Commons license. Moreover, TED contains a very specific type of AV content (single speaker, good audio quality, minimal background noise), which, even though not representative of the wide range of AV content types, makes it ideal for a depth-first experimentation of techniques for automatic subtitling, before passing to more challenging content types. For this reason, the TED guidelines are adopted as the main reference, even though other guidelines are mentioned in the following paragraphs. A summary of the constraints according to the TED guidelines is shown in Table 2.2.

Spatial dimension Spatial restrictions constrain subtitles to a specific length in two respects: number of characters per line (CPL) and number of lines.

[3]One example of such language-specific guidelines can be seen for Norwegian: www.sprakradet.no/globalassets/aktuelt/retningslinjer-for-god-teksting.pdf.

[4]https://partnerhelp.netflixstudios.com/hc/en-us/articles/215758617-Timed-Text-Style-Guide-General-Requirements

[5]https://www.ted.com/participate/translate/guidelines

Constraint	Realisation
Length	Characters per line (CPL) ≤ 42 Lines per subtitle: max. 2
Reading speed	Characters per second (CPS) ≤ 21
Segmentation	Linguistic units Equal length of lines

Table 2.2: Subtitling constraints acc. to the TED guidelines.

The allowed CPL of a subtitle depends on different factors, such as the size of screen, the font used, the client, the platform, the age of the audience and the country. For TV, cinema and DVD, the norm in interlingual subtitling is between 35-39 characters including spaces and typographical signs, while Video-On-Demand platforms, such as Netflix, and TED allow as many as 42 characters per subtitle line. These lengths are valid for single-byte languages (e.g. languages with Latin, Cyrillic and Semitic alphabets), while for double-byte scripts the maximum is lower, with around 14 CPL for Japanese and Korean and 16 for Chinese. There is normally no clear limit on how short a subtitle can be, but subtitles with less than four or five characters are rare (Díaz-Cintas, 2020, 98-99). As for the maximum number of lines per subtitle, the space allowed for a subtitle is about 20% of screen space. Therefore, a subtitle block should not exceed 2 lines.

Temporal dimension The temporal dimension concerns the display time of a subtitle. The first temporal constraint is reading speed, most commonly measured with a display rate in characters per second (CPS) or words per minute (WPM). Reading speed gives the maximum number of characters a viewer must read per second in order not to miss any information in the subtitle. The on-air time of a subtitle should be sufficient for the viewers to read and comfortably assimilate its content, while at the same time their

attention should not be monopolised by the subtitles. Recommendations for a comfortable reading speed depend on the user group (Szarkowska et al., 2016) and language (Ferriol, 2013). For example, 15 CPS are recommended for live interlingual English SDH (Ofcom, 2005), 12-15 CPS for offline interlingual subtitling in Central Europe (Szarkowska, 2016), 17–20 CPS in global online streaming services (Netflix, 2021) and 21 CPS for TED Talks (TED, n.d.). The duration of the utterance (measured in seconds or in feet/frames) is directly equivalent to the space a subtitle should occupy. This means an utterance of 4 seconds can accommodate the longest subtitle allowed for TED, that is 84 characters (two lines of 42 characters each). With the proliferation of online content, faster reading speeds are observed. Recent research found that most viewers could read the subtitles as well as follow the images, coping well even with fast subtitle speeds (Szarkowska and Gerber-Morón, 2018). On the other hand, a subtitle should not remain too long on the screen as this triggers more rereading, increasing the cognitive load and reducing enjoyment (Szarkowska and Gerber-Morón, 2018). According to Carroll and Ivarsson (1998), no subtitle should appear for less than one second or, with the exception of songs, stay on the screen for longer than seven seconds. Moreover, at least four frames should be left between subtitles to allow the viewer's eye to register the appearance of a new subtitle.

Another important aspect is synchronisation with the speech. Because of the co-existence of source audio and target text, there must be a close correlation between spoken utterance and subtitle content; source language and target language should be synchronised as much as possible (Carroll and Ivarsson, 1998). The synchronisation is performed through spotting (or timing), which consists in setting the start and end time of a subtitle in the form of time codes. The start and end times represent the time a subtitle will appear and disappear from the screen. The subtitle should

match as much as possible the start and the end of an utterance. In other words, "*the in and out times of subtitles should follow the speech rhythm of the dialogue and take cuts and sound bridges into consideration*" and "*spotting must reflect the rhythm of the film*" (Carroll and Ivarsson, 1998). Poor spotting can greatly hurt user experience as it is confusing and detracts the attention from who is saying what. On the contrary, a good synchronisation leads to a pleasant user experience and semiotic integrity.

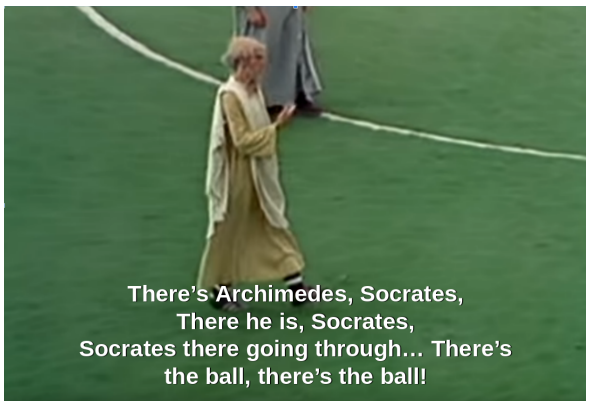
Segmentation and line breaks Subtitles have a fragmented nature; each subtitle is disconnected from the preceding and the following one. A long sentence needs to be split across multiple subtitles. Moreover, subtitle exceeding the CPL limit should be split in two lines, while often a long subtitle not exceeding the CPL limit is also segmented in two lines. From this we can define two types of breaks in subtitling: line breaks *across* subtitles, which delimit **subtitle blocks**, and line breaks *within* subtitles, delimiting **subtitle lines**. But how are segmentation decisions taken? There are two criteria defining a good segmentation. One focuses on achieving readability, meaning that the content should be presented in a way that facilitates comprehension, and the second one on legibility, referring to how easy it is for the eyes to read the content. Readability is based on semantic and syntactic criteria, while legibility on aesthetic criteria.

Readability is achieved through the preservation of 'linguistic wholes' (TED, n.d.). Subtitle segmentation should not rely only on the maximum length, but subtitle splits should not 'break' semantic and syntactic units, leading to subtitles which are "*semantically and syntactically self-contained*" (Carroll and Ivarsson, 1998). In an ideal case, every subtitle line (or at least subtitle block) should represent a coherent linguistic chunk *i.e.* a sentence or a phrase. In cases when this is not possible, subtitles should be arranged "*in sense blocks and/or grammatical units*" (Carroll

and Ivarsson, 1998). For example, a noun should not be separated from its article. When possible, segmentation should be performed at the highest syntactic node (Karamitroglou, 2001). A segmentation which respects linguistic units has been shown to facilitate comprehension and to lead to more readable subtitles in recent studies using eye-tracking (Perego, 2008; Rajendran et al., 2013; Szarkowska, 2016). Lastly, subtitles should respect natural pauses and not cross shot changes, even though empirical evidence has not been found suggesting that viewers re-read subtitles over shot changes (Krejtz et al., 2013).

The aesthetic criterion for splitting subtitles relates to the length of the upper and lower line. There is no consensus about whether the upper line should be longer or shorter. Wherever two lines of unequal length are used, a pyramidal structure tends to be favoured. The upper line should preferably be shorter to keep free as much of the image as possible and to reduce unnecessary eye movement in left-justified subtitles (Carroll and Ivarsson, 1998). 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 (Perego, 2008). For this reason, TED recommends an equal length of lines. In general, readability is prioritised over legibility, meaning that a segmentation decision should first be taken based on syntactic and semantic properties and if these allow multiple segmentation options, the aesthetic criterion can inform the optimal one. While subtitle length and reading speed can be controlled directly by the subtitle software typically used by translators, subtitle segmentation is left completely to the subtitler's jurisdiction. Subtitlers often have to either compromise the aesthetics in favour of the linguistic wholes or resort to omissions and substitutions. This means that a lot of technical, linguistic and extra-linguistic expertise goes into the segmentation process. In Section 4.3, we train segmentation models on human segmentation decisions

to automatically segment full sentences into subtitles.



(a) Original subtitle



(b) Subtitle adapted based on the subtitling constraints.

Figure 2.3: Example of a subtitle not conforming to the subtitling constraints and the same subtitle as it should be ideally displayed.

Figure 2.3 shows an example of a subtitle that does not conform to the constraints (left) and the same subtitle as it should be ideally displayed on the screen (right).[6] As it can be seen, the original subtitle (Figure 2.3a) is spread across four lines, covering almost 1/3 of the screen. Furthermore, the last two lines of the original subtitle are not split in a way such that linguistic wholes are preserved: "*There's the ball*" should in fact be displayed in a single line. In the right subtitle (Figure 2.3b) instead, the lines have been kept at two, by removing redundant information and unnecessary repetitions (*"Socrates, There's the ball"*). Each subtitle line consists of a full sentence, therefore logical completion is accomplished in each line and linguistic wholes are preserved. Lastly, *"going through"* has been substituted with a synonym (*"passing"*) in order not to exceed the 42-character limit. This all shows that the task of subtitlers is not limited to simply translating the speech on the screen, but they are required to creatively compress and adapt their translation to match the temporal, spatial and

[6]Screenshot taken from: `https://www.youtube.com/watch?v=i21OJ8SkBMQ`

aesthetic constraints. Translating subtitles can hence be seen as a complex optimisation process, where one needs to find an optimal translation based on parameters beyond semantic equivalence.

2.2.3 Quality

Antonini and Chiaro (2009, 99) remark that "*media translation is an area in which quality control becomes extremely difficult due to the complexity of several semiotic systems functioning simultaneously*". In Section 2.1.1 I have described how multimodality affects the translation. As a result, multimodality also affects the perceived quality. For example, in text translation omissions are normally treated as errors but in subtitling they are a condition *sine qua non*. In subtitling, quality does not depend only on semantic and semiotic parameters. Because of the form of the subtitles, quality additionally requires respecting the subtitling constraints described in Section 2.2.2. The initial attempts to ensure quality revolved around prescriptive guidelines, style guides and textbooks which included non-binding recommendations on linguistic and technical aspects, both formal and textual, such as punctuation and other conventions (Carroll and Ivarsson, 1998; Karamitroglou, 2001; Ofcom, 2005; Díaz-Cintas, 2020; BBC, 2021).

Later, classifications of quality criteria were proposed, either based on the author's own views (Bogucki, 2013; Klaus, 2015) or based on extensive surveys on professionals' opinions (Kuo, 2014; Robert and Remael, 2016) and professional vs. viewers' opinions (Szarkowska et al., 2020). In broad strokes, quality in subtitling can be grouped into three categories: technical parameters, factors derived from linguistic transfer, and factors related to the target culture, norms.[7] Technical aspects focus on the legibility of sub-

[7]Other important factors which determine quality are related to the working conditions of subtitlers, such as remuneration, deadlines, training and available materials (Díaz Cintas, 2001). In this thesis we focus on product-based factors and thus do not refer extensively to work-related factors.

titles and can be subdivided into i) presentation (length, number of lines, position, colour etc.) and ii) reading speed (subtitle speed, rhythm, pauses between subtitles) (Klaus, 2015). The second group is about translation-related aspects, with the emphasis being placed on comprehensibility. For this, semantic and syntactic accuracy needs to be ensured. The linguistic transfer parameters can be distinguished between a) language (simple wording and phrasing, correct spelling and punctuation, properly-placed line breaks); and b) content (condensation and coherence between the polysemiotic source text and the subtitle). The third category focuses on stylistic aspects, for example adequate register (Bogucki, 2013), appropriateness (Robert and Remael, 2016), and acceptability based on the norms of the target language and culture.

This tripartite of criteria is used as the basis of theoretical models for subtitle evaluation. Pedersen (2017) proposed the "FAR model", which stands for **F**unctional equivalence, comprising semantic and stylistic aspects; **A**cceptability, which concerns how well subtitles adhere to target-language norms and which looks at grammar, spelling, and idiomaticity; and **R**eadability, i.e., technical aspects related to how easily can the subtitles be processed, with the sub- categories a) segmentation and spotting, b) punctuation and graphics, and c) reading speed and line length. This model provides a numerical way of assessing subtitling quality by assigning penalties to the errors, based on their severity and potentially disruptive impact on viewers' comprehension. Similarly, Künzli (2020) conducted a large scale study of quality perceptions of subtitlers in German-speaking countries, based on which he proposed the CIA model. CIA stands for **C**orrespondence (denotational, connotational, pragmatic, aesthetic), **I**ntelligibility, focusing on legibility and readability, and **A**uthenticity, focusing on orality and idiomaticity.

For live intralingual subtitling via respeaking, the NER model (Romero-

Fresco, 2011) has been the most popular model for quality assessment, and widely applied both in the industry and academia. The NER (Number of words, Editions, Recognition) model draws from the WER (Word Error Rate) calculation, and allows for the calculation of an 'accuracy rate'. This rate is not estimated at the level of words, but of 'idea units', defined as a "*unit of intonational and semantic closure*" (Chafe, 1985, p. 106). To obtain the score, the number of idea units contained in the original dialogue is compared to the number of idea units found in the subtitle, and edition and recognition errors are substracted from the number of words. The model does not penalise correct editions. Correct editions (CEs) are deviations from the original that do not involve a loss of information, for example omissions of redundancies and hesitations, and are not factored in the accuracy score. The errors are classified as 'minor', 'standard' or 'serious', scoring 0.25, 0.5 and 1. Subtitles need to reach an accuracy rate of at least 98% to be considered suitable for users.

The NER model was later adapted to interlingual live subtitling, as the NTR model, drawing from the evaluation of simultaneous interpreting (Pöchhacker, 2001). The formula substitutes edition errors with translation errors as follows:

$$NTR = \frac{N - T - R}{N} * 100 \tag{2.1}$$

with N corresponding to the number of words, T to translation errors, and R to recognition errors. This model rewards positive solutions by counting the number of Effective editions (EE) and calculates the accuracy rate to a more standard 10-point scale.

All these models and criteria have been tailored for evaluating human subtitling, but can possibly accommodate scenarios of evaluating automatic subtitles. However, they all rely on human annotations for estimating the scores, a costly and long process. In this thesis, we propose

automatic metrics for evaluating the readability and intelligibility of subtitles (Sec. 5.2 and 6.2).

2.2.4 Tools and master templates

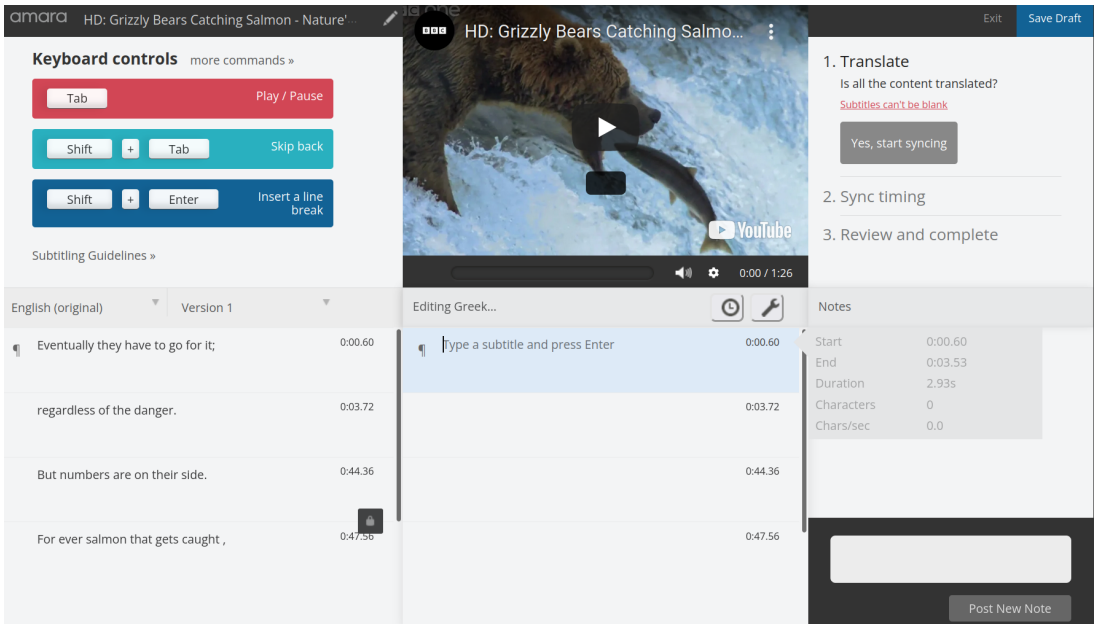
Moving into the professional aspects of subtitling, this section presents the tools subtitlers use in their workflows. Subtitling has always been a technical profession, in the sense that subtitlers needed to work with specialised software, from the analogue subtitle workstations during the teletext era to modern file conversion and video editing programs. With the DVD and later the streaming era, more and more computer-assisted translation (CAT) tools appeared for subtitling. These are called subtitling editors and their goal is to simplify the subtitling process, mainly by aiding in the technical aspects of the process, e.g. spotting and conforming to the length constraints. Figure 2.4 shows screenshots from two popular subtitle editors, Amara[8] and Matesub[9]. Subtitle editors normally contain a video area, a timeline of the subtitles and a text editing area. Common features include spell checkers, quality control on subtitles exceeding the length or reading speed constraints, a waveform which allows the subtitler to identify video segments containing speech and shot change detection. With the fragmentation of the subtitling process among different agents, the increasing need of collaboration between multiple agents involved in the subtitle creation process led to more cloud-based editors (Bolaños-García-Escribano and Díaz-Cintas, 2020; Bolaños-García-Escribano et al., 2021b). Recently, subtitle software has started incorporating MT and speech technologies, while editors are being tailored for post-editing.

An instance of an automatic subtitling tool customised for post-editing is Matesub. Matesub is a novel subtitling tool featuring a state-of-the-

[8] amara.org/

[9] https://matesub.com/

art speech translation system, automatic generation of timestamps for the translated subtitles – a process called automatic spotting (or auto-spotting) – and automatic segmentation of the translated audio into subtitles.



(a) The Amara subtitle editor



(b) The Matesub subtitle editor

Figure 2.4: Two typical instances of subtitle editors. Both are cloud-based.

Figure 2.4b shows a screenshot of the tool. The subtitlers are presented with a list of the automatically generated subtitles (upper left box) and the video on which the subtitles appear (upper right). The boxes corresponding to each subtitle appear at the bottom of the screen, superimposed on a waveform which allows the subtitler to identify parts of the video corresponding to the selected speech segments. The position and length (duration) of the boxes can be adjusted to match the beginning and the end of the spoken utterance and to accommodate the time the subtitle will appear on screen. Automatic constraint control, along with other useful features, such as keyboard shortcuts and positioning or colour settings, are implemented as well, therefore post-editing subtitles in Matesub is representative of subtitlers' real working settings. The tool is free, tested in real-life use cases and is already being used by professional subtitlers.

Subtitlers should ideally always have access to the source video when translating. However, working directly from the video has several drawbacks. The subtitler needs to make sense of what is said on the screen, deal with regional accents, noise, unintelligible speech etc. One way to speed up this labour-intensive process, especially in settings where several target languages are involved, is creating a subtitle template. A subtitle template is a source language/English version of a subtitle file already segmented and containing timestamps (time codes), which is used to directly translate the text in target languages while preserving the same structure (Díaz-Cintas, 2020; Georgakopoulou, 2019; Netflix, 2021). In other words, it is an enriched transcript of the speech where the text is already compressed, timed and segmented into proper subtitles. Optimal duration, length and proper segmentation are ensured, since the change of code between oral and written is curated by an expert, teasingly called the 'templator'. Figure 2.5 shows an instance of a template. The translation needs to fit in the space allowed in column D. This process reduces the cost, turn-around times and

	A	B	C	D	E
1	In-times	Out-times	Source	Target	Subtitle ID
2	00:00:01,122	00:00:03,663	Immaginate di svegliarvi con uno sconosciuto,		1
3	00:00:03,687	00:00:06,133	a volte più di uno sconosciuto,		2
4	00:00:06,939	00:00:09,777	che mette in discussione il vostro diritto di esistere		3

Figure 2.5: Example of a subtitling template opened in a spreadsheet.

effort needed to create a separate timed version for each language, and opens up the pool of translators able to perform the task, since it does not require special skills for spotting and segmentation. Another benefit is that it facilitates quality assurance since errors can be spotted across the same blocks (Nikolić, 2015). This template can serve as a basis for translation into other languages, whether the translation is performed by a human or an MT system. These benefits motivated our work in Section 5.3 for exploring architectures for simultaneously generating two language versions (captions and subtitles) with the maximum consistency, where the caption file can further serve as a template for multilingual localisation.

However, templates have downsides too. They are often fixed to the constraints of the template language, meaning that translators are not allowed to adjust the timing or structure of the subtitles, even if needed (Nikolić, 2015). As a consequence, the translation has to fit to the template language constraints. Considering potential cross-linguistic differences in subtitling conventions for different target languages, regions, broadcasters or streaming platforms etc., templates may sometimes in fact hamper or prevent optimal solutions for the target language. To accommodate the needs of the target language, subtitlers should have the flexibility to adjust subtitling segmentation, timestamps as well as delete, merge and split the subtitles (Oziemblewska and Szarkowska, 2020). Moreover, reliance on (sometimes poor-quality) templates and de-skilling of subtitle translation, by removing the need for specialised skills such as spotting, may have

adverse effect on overall subtitling quality in the long term (Kapsaskis, 2011). In this thesis we explore alternative workflows to increase efficiency in subtitling without relying on templates (Section 5.2).

2.2.5 Volunteer subtitling at TED

With the recent advances in digital technologies, subtitling is no longer a task practised exclusively in professional settings, but has become extremely popular among volunteer and fan communities (Bogucki, 2009; Díaz-Cintas and Sánchez, 2006; Simó and Rosario, 2005; Orrego-Carmona and Lee, 2017), who take it upon themselves to produce subtitles. Volunteer subtitling communities come together for a wide range of different purposes, varying from entertainment to activism or the democratisation of information. This type of subtitling is also described as Non-Professional Subtitling (NPS), since there is no monetary remuneration involved. Due to this large variation, NPS initiatives are difficult to categorise: non-professional subtitles can be produced by individuals through their own initiative or by structured communities with well-defined hierarchies, guidelines and workflows (Orrego-Carmona, 2016).

The TED Translators programme is one of the most well-established volunteer subtitling initiatives currently operating. TED is a media organisation producing freely accessible online talks under the slogan "ideas worth spreading".[10] TED was launched in 1984 as a conference covering topics on technology, entertainment, and design, but has since expanded to many other areas, including science, humanitarian issues, and politics. The dissemination of knowledge is at the core TED's mission to 'spread ideas'. TED Talks are brief and feature leading figures from around the planet. The talks are normally single-speaker, scripted and prepared but rehearsed to sound natural and spontaneous, with a conversational style. They have

[10]https://www.ted.com/

limited disfluencies, hesitations or background noise. Talks are recorded and uploaded to the TED website, where they become part of an internationally accessible repository. As of March 2022, the website advertises more than 3900 TED Talks, as well as other videos, for example TED-Ed (the TED education initiative) and TEDx (independently organised events including local speakers under the TED umbrella).

Integrating translation to a project like TED as part of an international expansion strategy was a natural step for an organisation that promotes access to knowledge. The TED Open Translation Initiative, which later became TED Translators, was launched in 2009. As part of its crowdsourced translation project, TED provides a highly controlled environment for the development of volunteer activities. Over the years, TED Translators has had three cloud-based technology partners: dotSUB was used until 2012, when it was replaced by Amara. The Amara subtitle editor allows for translating, timing and quality check of subtitles, as well as communication between the volunteers. Currently, the project uses CaptionHub,[11] a subtitling platform that integrates automated solutions such as speech recognition and MT. Each one of these organisational shifts had a direct effect on the workflows of the agents involved in the production of the subtitles and, very likely, the final product (see Sec. 4.2). With almost 39,000 registered translators, TED has developed a solid fan base of users who are willing to support the project by volunteering their time and skills.

To ensure high quality, a staggered system of tasks is implemented, where responsibilities are clearly defined and controlled. The talks are first transcribed and synchronised in the source language. The resulting captions are then translated by volunteers. To become engaged in the translation of TED talks, volunteers are required to meet three main criteria: they need to be fluently bilingual in the source and target languages,

[11] https://captionhub.zendesk.com/hc/en-gb

they need to be 'fluent in the transcription language' (source language of the video) and they should be 'knowledgeable of subtitling best practices'. The subtitles are revised by more experienced volunteers who have already published at least 90 minutes worth of translations into their target language.[12] Before publication, the reviewed translations need to be approved by higher-ranking volunteers, for example a language coordinator, or by a staff member. TED has released transcription and translation guidelines and subtitling training material for the volunteers. Due to the richness of the topics and speakers and the high quality of subtitles, TED provides excellent material for creating corpora for training and testing automatic subtitling solutions, as we show in Section 4.2.

2.3 Conclusion

In this chapter I have presented the general scope of AVT as a dynamic field of translation. The challenges related to the translation of AV texts stem from the multimodal character of the source and the formal requirements of the target (Sec. 2.1.1). Different types of AVT exist, with the most prominent being subtitling and revoicing (Sec. 2.1.2). Subtitling, whether intralingual or interlingual, offline or live, is bound by spatial and temporal constraints, expressed as a desired conformity to length and reading speed, segmentation based on semantic and syntactic properties, and synchronisation with the utterances (Sec. 2.2). Satisfying these criteria is an integral part of quality for subtitlers, who are supported by specialised software and file formats (templates) in the process of subtitling (Sec. 2.2.4). The advances in digital technologies are opening up the subtitling practice to non-professionals, leading to volunteer subtitling projects such as the TED Translators program (Sec. 2.2.5). These characteristics make AVT

[12]TED Get started, https://www.ted.com/participate/translate/get-started

a challenging translation type, acting as a test bed for probing the limits of machine translation and other language technologies. The next chapter is dedicated to machine translation, providing the basic background on neural MT and speech-to-text translation, as well as their application to subtitling.

Chapter 3

Neural Machine Translation

The second part of the background is more technology-oriented and focuses on the state of the art of neural machine translation (NMT). After a more general introduction to NMT (Sec. 3.1) and the prevailing deep architecture, the Transformer (Sec. 3.1.2), in Section 3.2, I move to the workings of speech-to-text translation (ST). The two dominant approaches, cascade and direct, are explained in detail and compared. Then, I present specific ST scenarios relevant to this thesis, namely joint generation of transcription and translation (Sec. 3.2.3) and simultaneous ST (Sec. 3.2.4). Transitioning to more subtitle-related aspects, Section 3.3 outlines previous methods for the task of subtitle segmentation, that is segmenting a text into a series of subtitles. The related work on MT for subtitling is also part of this chapter. A comprehensive review is presented in Section 3.4, showing how the technological advancements lead up to the establishment of the new paradigm of automatic subtitling. The last part of this chapter is dedicated to evaluation (Sec. 3.5), both human and automatic, as well as particularly in the case of MT for subtitling.

3.1 NMT and sequence-to-sequence learning

In recent years, with the advent of deep learning, NMT has become the dominant concept in machine translation. NMT systems are trained on large amounts of data, up to millions of parallel sentences. A single model is trained in an end-to-end manner, learning complex characteristics autonomously and directly from the data. From a computational point of view, translation can be viewed as a statistical problem of predicting the probability of a target sentence y given a source sentence x.

$$p(y|x) = \frac{p(x,y)}{p(x)} \tag{3.1}$$

In NMT, the translation model is factorised as a product of word-level translation probabilities,

$$P(y|x;\theta) = \prod_{j=1}^{J} P(y_j|x, y_{<j};\theta) \tag{3.2}$$

where θ is a set of model parameters and $y_{<j} = y_1, ..., y_{j-1}$ denotes a partial translation.

Most of the NMT approaches follow the encoder-decoder architecture (Sutskever et al., 2014; Bahdanau et al., 2015). In this architecture, a neural network (NN), the 'encoder', encodes the sentence in the source language in a continuous representation. Then, a second neural network, the 'decoder', decodes the continuous representation obtained by the encoder into the translation of the source sentence in the target language. A third component is the attention mechanism, which functions similar to an alignment model between source and target. In the next paragraphs, I present the details of these components.

NMT can be framed as a sequence-to-sequence (seq2seq) learning problem. Seq2seq models generate a sequence of observations given another

sequence of observations as input (Sutskever et al., 2014). Contrary to structured prediction tasks, such as POS-tagging, in sequence-to-sequence learning the two sequences can have different lengths. This is a fundamental property for translation, where the length of the source and target sequences do not necessarily match. These models are trained by minimising the cross-entropy loss as the training objective. Given X as a sequence of source words $x^1, x^2, ..., x^N$ and Y as a sequence of target words $y^1, y^2, ..., y^M$, then seq2seq learning optimises the cross-entropy loss:

$$L_\theta(\tilde{Y}, Y) = \sum_{i=0}^{N} logP(\tilde{y}_i = y_i | X, y_{<i}; \theta) \tag{3.3}$$

where θ is the vector of model's parameters and $\tilde{y}_i$ is the predicted output. During decoding, the prediction of the new words is conditioned on the source representation and the previously predicted tokens. During training, the prediction is conditioned on the source and the correct previous tokens instead of the predicted tokens.

The encoder is a NN which transforms the input sequence into an input sequence embedding. The input sequence is a sequence of words which are encoded as word embeddings through an embedding matrix $E(w_j)$. The embedding matrix can be seen as a lookup table for the vocabulary, mapping the one-hot vector to a lower-dimensional space. The sequence of word embeddings $\mathbf{X}$ is processed through the encoder to obtain a hidden representation $\mathbf{H}$.

$$H = encoder(X) \tag{3.4}$$

Then, the representation of the input text is passed to the decoder, which generates the target sequence, one token at a time. It takes the input representation $\mathbf{H}$, the previous decoder hidden state s_{i-1} and the embedding of the previous output word Ey_{i-1} and generates a new output word prediction Ey_i and a new hidden decoder state s_i:

$$s_i = decoder(s_{i-1}, Ey_{i-1}, H) \tag{3.5}$$

The output word is predicted from this hidden state. The prediction is in the form of a probability distribution over the output vocabulary. For example, if the target vocabulary contains 40,000 words, the prediction is a 40,000-dimensional vector, where each element contains the probability of each word in the vocabulary. The highest value in the vector is the output word prediction. The raw output vector is transformed into a probability distribution through the softmax function, as follows:

$$t_i = softmax(W(s_i) + b), \tag{3.6}$$

where W is a weight matrix and b a bias term.

The encoder and the decoder are normally composed of several stacks of NN layers, hence *deep* architectures. Researchers have experimented with different types of NNs for building the encoder-decoder architecture. In the implementation by Sutskever et al. (2014), they were computed through a long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997; Graves, 2013). LSTMs are a type of recurrent neural networks (RNN). An RNN is a network which conditions each hidden state not only on the previous prediction but also on the previous state, in a recurrent way. RNNs allow for modelling arbitrarily long sequences and a sequential processing of their input. In this way, each step takes information from the new input word and the full sentence history. However, RNNs have several disadvantages. First, the importance of previous words diminishes as the context becomes longer, with more recent information 'drowning' older information. Second, back-propagating all updates to the beginning of the sequence can lead to the problem of vanishing or exploding gradients, especially for longer sequences (Hochreiter et al., 2001). LSTMs address these problems by incorporating a memory cell which controls how

much of the previous history needs to be remembered or forgotten. In both cases, the encoder can be built by two networks, one processing the sequence left-to-right and one right-to-left to take into account both sides of the context. In this case the network is called bidirectional.

The explicit time dependencies over each word in the sequence make training of deep RNNs computationally expensive (Wu et al., 2016; Miceli Barone et al., 2017). In an attempt to find more efficient architectures, Gated-Recurrent Units (GRUs) (Cho et al., 2014) were proposed, which simplified the LSTM cell by removing one gate and one state vector. Other alternatives are non-recurrent approaches, such as the convolutional sequence-to-sequence model (ConvS2S) (Gehring et al., 2017) which is built from convolutional neural networks, often used in vision. These take advantage of the full parallelisation capabilities of GPUs and are computed in parallel over word sequences, speeding up the training by one order of magnitude. In fact, deep and non-recurrent architectures soon outperformed LSTM-based NMT in both translation quality and training time (Gehring et al., 2017; Vaswani et al., 2017).

3.1.1 Attention

Attention mechanisms in Neural Networks (NN) were first introduced in the field of computer vision (Larochelle and Hinton, 2010). Attention is motivated from information processing in humans. Intuitively, the attention mechanism can be described as a method which focuses on just some specific information, instead of the entire context. Humans tend to focus on a certain region (whether it is inside an image or in a sentence during reading) with "high resolution" while perceiving the surroundings in "low resolution", and adjust the focal point over time.

In translation, some words in the source can be particularly important for translating a specific target word. Between typologically diverse lan-

guages of different word order and in long sentences, knowing the previous (or following) n words may not be sufficient. One example is the verb-subject agreement, where the subject may be several words apart from the verb. Another bottleneck of the seq2seq models is that the source sentence is encoded as a single state vector, regardless of its length. This means that i) this state vector needs to contain all the information of the source sequence, without forgetting the first words when reaching the end of the sentence, and ii) the decoder state is overloaded because it keeps track of the source sentence as well as of the previously translated words.

Attention creates a 'bridge' between the encoder and the decoder. Intuitively, it weights each input word based on how important it is for generating the next target word (at each decoder state). The attention mechanism proposed by Bahdanau et al. (2015) is informed by the input word representation h_j and the previous hidden state of the decoder s_{i-1}, and it produces a context vector c_i, computed as:

$$c_i = \sum_{j=0}^{N} a_{ij} hj \tag{3.7}$$

The attention a is a feed-forward network which computes the association between the input word and the encoder states as:

$$a(s_{i-1}, hj) = v_a^T tanh(W_a s_{i-1} + U_a hj), \tag{3.8}$$

where Wa and Ua are weight matrices, v_a is a weight vector and $tanh$ is the hyperbolic tangent non-linear activation function. The output of this computation is a scalar value, which is then normalised using softmax, so that all attention values add up to one:

$$a_{ij} = \frac{exp(a(s_{i-1}, hj)}{\sum_k exp(a(s_{i-1}, h_k)} \tag{3.9}$$

The context vector is different for each decoder time step i. With attention, the input representation H is replaced by the context vector c_i in Equation 3.5 as follows:

$$s_i = f(s_{i-1}, Ey_{i-1}, c_i) \tag{3.10}$$

Later, the attention mechanism was refined with alternative alignment functions and simpler computation paths (Luong et al., 2015) and different types of attention appeared.[1]. The attention mechanism was an important improvement for NMT and led to notable improvements, establishing NMT as the main paradigm in automatic translation.

Self-attention Self-attention is only one of the many proposed attention architectures. There is, however, a main difference between self-attention and other attention types. Attention is a mechanism operating outside the input (in NMT it is found in the decoder) which attends either to the whole input state (Global/Soft) or to a part of the input state (Local/Hard). Self-attention on the other hand, connects several positions of the same sequence with each other. With attention, when the decoder needs to output the word $y4$ of the target sequence it may attend to positions $x1$ and $x4$ in the source sequence. With self-attention, however, each word in the encoder is weighted for its relevance to the rest of the words in the source sequence. The same occurs in the decoder with the target sequence.

Self-attention-based architectures have achieved state-of-the-art results in many Natural Language Processing (NLP) tasks and domains. Architectures such as the Image Transformer (Parmar et al., 2018), the Transformer (Vaswani et al., 2017) and the BERT (Devlin et al., 2019) language

[1]Others include Soft vs. Hard Attention, Global vs. Local Attention, Causal Attention in SNAIL and many others. However, the comparison of all these attention types is outside the scope of this thesis. For a full comparison of attention mechanisms, see Lilian Weng's blog post: https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

model have pushed the field forward, achieving state-of-the-art results in computer vision, NMT and several NLP tasks respectively.

3.1.2 Transformer

Definitely one of the most impactful papers of the previous years, 'Attention is all you need' (Vaswani et al., 2017) acknowledged the importance of self-attention for NMT. The proposed architecture, the Transformer, enabled sequence-to-sequence modelling based entirely on self-attention and completely eliminated the need for recurrent connections. The encoder and the decoder of the Transformer are actually feed-forward networks. This is counter-intuitive, since a feed-forward NN cannot capture information about the position of the words in a sequence, as it is the case with RNNs. Therefore, positional encodings are added to the word embeddings, which contain information about the position of each word in the sentence. The architecture of the Transformer is shown in Figure 3.1. The architecture can be summarised in the following components:

Scaled Dot-product Attention The attention mechanism in the Transformer is based on the notions of *keys* (K), *queries* (Q) and *values* (V). The output of the attention is computed using scaled dot-product:

$$Attention(Q, K, V) = softmax\frac{QK^T}{\sqrt{d_k}}V \tag{3.11}$$

where d_k is the dimension of the input (keys). Scaled dot-product is nothing more than a dot product with an added scaling factor $1/\sqrt{d_k}$. The scaling factor is added to facilitate learning, since for long inputs, the softmax function can have extremely small gradients.

The computation of attention can be described in the following steps:

1. Obtain weight matrices Q, K and V by multiplying the embedding

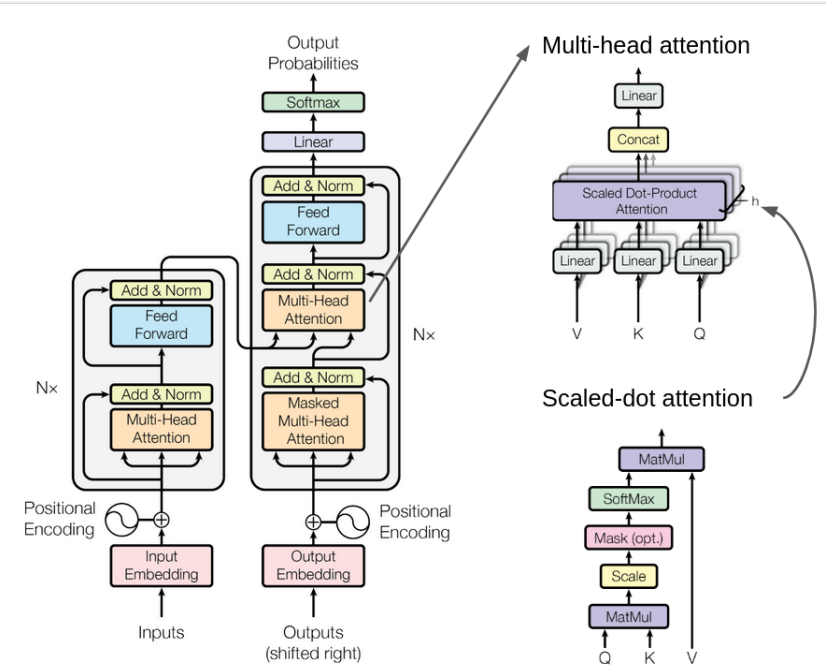


Figure 3.1: The Transformer model architecture. Figure from (Vaswani et al., 2017).

matrix X (concatenation of input embeddings for each word in the sequence) with a different weight matrix W^Q, W^K and W^V.

2. Compute the dot product of QK.

3. Divide by the square root of the input dimension.

4. Normalise the number obtained in Step 3 by softmax to output values between 0-1.

5. Multiply V by the softmax output to weigh important words higher and eliminate non important ones.

Multi-Head Self-Attention Instead of having only one 'round' of attention, Transformer applies multiple passes of the scaled dot-product attention. Each pass is performed by a different *head*. Since the weight matrices Q, K and V are initialised randomly, each head will compute a different representation of the input. Equation 3.12 shows how Multi-Head Attention

is computed.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_n)WO \tag{3.12}$$

Each head outputs a matrix $head_i$, all of which are then concatenated and multiplied by a weight matrix W_h. The result, which contains information from the entire sequence, is then passed to the feed-forward network.

Encoder The encoder is comprised by several encoder stacks. Each encoder stack is split in two subcomponents: 1) multi-head attention + layer normalisation (Add & Norm); and 2) feed-forward network + layer normalisation. A residual connection is applied around each sublayer for easier back-propagation and to avoid gradient vanishing.

Decoder The decoder is also a stack of decoders, similar to the encoder, with one difference: it contains a third subcomponent, called masked multi-head attention. During decoding, the decoder should not 'cheat' by looking at future positions, therefore a mask is applied, which hides future positions (words). The output of the top encoder, in the form of a K_{enc} and V_{enc} matrix is passed to every decoder to generate the output word x_1. After generating x_1, the output is passed again to the decoder, which performs self-attention on the target sequence in order to continue generating words up to the end of the sentence.

The Transformer is the main architecture used for building our segmentation models in Section 4.3 and the NMT models in Chapter 5.

3.2 Speech translation

Speech translation (ST) is the translation of a spoken signal and can have two types of output: speech (speech-to-speech) or text (speech-to-text).

Speech-to-text translation, its application to AVT being the main focus of this thesis, is the task of translating a speech signal in a source language S into text in a target language T. ST has been traditionally approached with cascade architectures, combining a pipeline of different subtasks. In the past few years, direct architectures have been on the rise, translating the speech signal in an end-to-end fashion. In both cases, the audio signal, which is stored as a waveform, is split into short-term frames using overlapping windows and log Mel filterbanks are extracted as features to be given to the system, whether direct or cascade. In the next sections, I describe each of these two approaches.

3.2.1 Cascaded speech-to-text translation

Traditionally, the task of Speech-to-Text Translation has been addressed with cascade systems consisting of two components: an Automatic Speech Recognition (ASR) system, which transcribes the speech into text, and an MT system, which translates the transcribed text into the target language (Stentiford and Steer, 1988; Waibel et al., 1991; Eck and Hori, 2005). Figure 3.2 shows a scheme of a typical ST cascade architecture. Because of its pipelined architecture, other intermediate modules can be added depending on the target application, such as text normalisation, punctuation restoration, text condensation etc. In speech-to-speech translation, the pipeline is complemented by a text-to-speech (TTS) component which synthesises the text into speech. Even though applicable for revoicing, TTS is out of the scope of this thesis, therefore I will here focus only on the ASR and MT components.

By concatenating ASR and MT components, cascade ST architectures represent an intuitive solution to achieve reasonable performance and high adaptability across languages and domains. Cascades have the benefit that they can take advantage of state-of-the-art technology for both individual



Figure 3.2: Typical ST cascade architecture.

components, each component can be evaluated independently, and they can leverage the large amounts of data available for both tasks. At the same time, however, they are susceptible to well-known problems related to the concatenation of multiple systems. First, they require *ad-hoc* training and maintenance procedures for the ASR and MT modules. Second, they suffer from the propagation of ASR errors to the MT system, and from the loss of speech information (e.g. prosody, speakers' vocal traits) which might be useful to improve the final translations, especially in the case of AV texts.[2] Research has focused on mitigating error propagation by: *i)* feeding the MT system with ASR data structures (e.g. ASR n-best, lattices or confusion networks) which are more informative than the 1-best output (Lavie et al., 1996; Matusov et al., 2005b; Bertoldi and Federico, 2005; Beck et al., 2019; Sperber et al., 2019), and *ii)* making the MT robust to ASR errors, for instance by training it on parallel data incorporating real or emulated ASR errors (Peitz et al., 2012; Ruiz et al., 2015; Sperber et al., 2017; Cheng et al., 2019; Di Gangi et al., 2019c; Martucci et al., 2021). Although the former solutions have shown to be effective to some extent, state-of-the-art cascade architectures (Pham et al., 2019; Bahar et al., 2020) prefer the latter, as they are simpler to implement and maintain.

One further drawback of the cascaded approach, particularly relevant for the task of subtitling, is that any transcript, no matter how accurate, is subject to information loss in the semiotic shift from the richer audio

[2]In Section 5.2 we will show that speech pauses are useful for predicting the correct segmentation of the text into subtitles.

representation to the poorer text representation. Attempts to preserve speech information have been made in speech-to-speech translation cascades mainly for improving the naturalness of the synthesised speech and for resolving ambiguities. This has been performed through acoustic feature vectors related to different prosodic elements, such as duration and power (Kano et al., 2013), emphasis (Do et al., 2015, 2016) and intonation (Aguero et al., 2006; Anumanchipalli et al., 2012).

3.2.2 Direct speech-to-text translation

Direct ST approaches overcome the limitations of cascade ST models by training a system that translates directly from the audio, bypassing any intermediate representations, e.g. transcription. Direct ST can be defined as a seq2seq task, where a sequence X of audio features in a source language S is mapped into a sequence of words Y in a target language T, Y being the translation of the word sequence uttered in the audio signal represented by X. In this way, a single neural network can be trained to maximize the probability of $P(Y|X)$.

Bérard et al. (2016) and Weiss et al. (2017) proposed the first direct ST solutions. These were encoder-decoder architectures based on recurrent neural networks. Currently, more effective solutions (Di Gangi et al., 2019b; Potapczyk and Przybysz, 2020; Bahar et al., 2020; Gaido et al., 2020b) rely on ST-oriented adaptations of the Transformer (Vaswani et al., 2017), where the encoder is integrated with: *i)* convolutional layers which reduce the input length to make the input computationally manageable, and *ii)* penalties which bias the model to attend to local context in the encoder self-attention layers (Povey et al., 2018; Sperber et al., 2018; Di Gangi et al., 2019b).

These architectures are more effective and have shown encouraging results (Sperber and Paulik, 2020). However, one arduous problem in direct

ST is training data paucity, a critical bottleneck for neural solutions in general. In ST, the problem has been mainly tackled with data augmentation and knowledge transfer techniques. Data augmentation increases the amount of training data either by producing artificial training corpora or by augmenting the existing datasets. In the first case, the generation of new data – (*audio*, *translation*) pairs – can be performed in two ways: 1) the text side of existing speech corpora (audio, transcription) is automatically translated into a target language through MT, or 2) existing translation corpora (source-target text) are speech synthesised through TTS (Bahar et al., 2019b; Nguyen et al., 2020; Ko et al., 2015; Jia et al., 2019). The second direction consists in increasing the data size by adding copies of the original data where the audio has been 'corrupted' in various ways (e.g. by speeding up, slowing down, adding noise). One such technique is SpecAugment (Park et al., 2019), an online data augmentation method, which does not augment the waveform itself but rather the spectrogram by warping it in the time direction and masking blocks of utterances and frequencies.

Knowledge transfer (Gutstein et al., 2008) consists in leveraging the knowledge from another trained model. In the case of ST, the knowledge learnt from a neural network trained on closely related tasks (ASR and MT) is passed on to the ST model. On one side, existing ASR models have been used for encoder pre-training (Bérard et al., 2018; Bansal et al., 2019; Bahar et al., 2019a) and multi-task learning (Weiss et al., 2017; Anastasopoulos and Chiang, 2018a; Indurthi et al., 2020). On the other side, existing neural MT models have been used for decoder pre-training (Bahar et al., 2019a; Inaguma et al., 2020), joint learning (Indurthi et al., 2020; Liu et al., 2020b) and knowledge distillation (Liu et al., 2019; Gaido et al., 2020b).

As mentioned above, direct ST receives audio as input. The model receives two types of information from the spectrogram, along its two axes:

1) information about the temporal dimension of the speech, e.g. duration, and 2) information related to the frequency, such as pitch, power and other prosodic elements. While intonation and stress are mostly related to semantic properties, speech tempo directly affects the compression rate of subtitles, and pauses often correspond to prosodic chunks which can determine the subtitle segmentation. Given that, in this thesis I will explore whether access to this information can lead to better modelling of the subtitling constraints and subtitle segmentation (Sec. 2.2.2).

Comparisons Most of the works on direct ST also evaluate the proposed solutions against a cascade counterpart. The conclusions, however, have long been discordant. Only to mention a few examples, Pino et al. (2019) showed similar scores for the two approaches, Indurthi et al. (2020) reported higher results for their direct model, while Inaguma et al. (2020) reported higher results for the cascade. Besides the inconclusiveness, other problems of these comparisons included: *i)* not all the architectures were equally optimised (e.g. hyperparameter search in the cascade), *ii)* in order to have a fair comparison in terms of training data, cascade systems were restricted to unrealistic settings with a limited amount of training dataand *iii)* evaluation relied only on automatic metrics computed on single references. The IWSLT campaigns (Niehues et al., 2019; Ansari et al., 2020; Anastasopoulos et al., 2021, 2022) have set up a shared evaluation framework where systems built on large training data are optimised to achieve the best performance, independently of the underlying architecture. In the 2020 edition, direct models approached, and in one case (Potapczyk and Przybysz, 2020) even outperformed, the cascade solutions. This tendency continued in the last two editions, where direct models approached the traditional cascades. In comprehensive study based on multiple human references and error annotation of the outputs, Bentivogli et al. (2021)

showed that the performance gap between the two approaches is closed. The two approaches make different types of errors, the cascade making less morphology and word order errors, while the direct being more robust to audio understanding and taking advantage of prosodic features in disambiguating the translation. In Section 5.2 we perform comparisons of cascade and direct architectures in automatic subtitling.

3.2.3 Joint generation of transcription and translation in ST

The idea of generating both the transcript and the translation of the same input speech has been previously addressed in (Weiss et al., 2017; Anastasopoulos and Chiang, 2018b). These papers presented different solutions (both cascade and direct) with the goal of improving translation performance by leveraging both ASR and ST data in direct ST. Later, Sperber et al. (2020) reviewed different cascade, direct and joint architectures with the focus of jointly producing consistent source and target language texts. Their underlying intuition is that, since in cascade solutions the translation is derived from the transcript, cascades should achieve higher consistency than direct solutions. Their results, however, showed that triangle models (Anastasopoulos and Chiang, 2018b), in which the translation decoder attends to the output embeddings of the ASR encoder, achieve the highest consistency among the architectures tested (considerably better than that of cascade systems) and have competitive performance in terms of translation quality. Direct independent models and models sharing only the audio encoder with two different decoders (Weiss et al., 2017), instead, do not achieve the translation quality and consistency of cascades. However, all these previous efforts have fallen outside the domain of automatic subtitling and ignore the inner structure of the subtitles and their relevance when considering consistency. We are experimenting with joint generation of captions and subtitles in Section 5.3.

3.2.4 Simultaneous speech translation

Simultaneous Speech Translation (SimulST) is the task in which the generation of the translation starts before the audio input becomes entirely available. As with offline ST, cascades have long dominated the simultaneous translation landscape. In cascade architectures, a streaming automatic speech recognition (ASR) module transcribes the input speech into the corresponding text (Wang et al., 2020b; Moritz et al., 2020). Then, a simultaneous text-to-text translation module translates the partial transcription into target-language text (Gu et al., 2017; Dalvi et al., 2018; Ma et al., 2019; Arivazhagan et al., 2019b). Recently however, direct architectures are gaining momentum, having the benefit of lower latency (translation delay) over cascaded approaches.

SimulST systems are trained with a Prefix-to-Prefix Framework, meaning that the generation of the target word y_t is based on the source prefix $x_{g(t)}$ and not on the entire source sequence X as in seq2seq training. $g(t)$ is a monotonic non-decreasing function of t that denotes the number of source words processed by the encoder when deciding the target word y_t.

In simultaneous settings, a model receives input incrementally, and at each time step chooses between a read or a write action. A READ action corresponds to receiving new information from the input, while the WRITE action means outputting the translation using the information received until that step. Determining the next action is performed through a decision policy. Decision policies can be divided into *fixed* and *adaptive*. In fixed policies, the decision of when to write is taken based on the elapsed time. Among the most popular fixed policies, the wait-k policy (Ma et al., 2019), initially proposed for simultaneous MT, consists in reading k source words before starting to generate the target sentence. Figure 3.3 shows an example of the policy with $k = 2$. With the wait-k strategy, the target

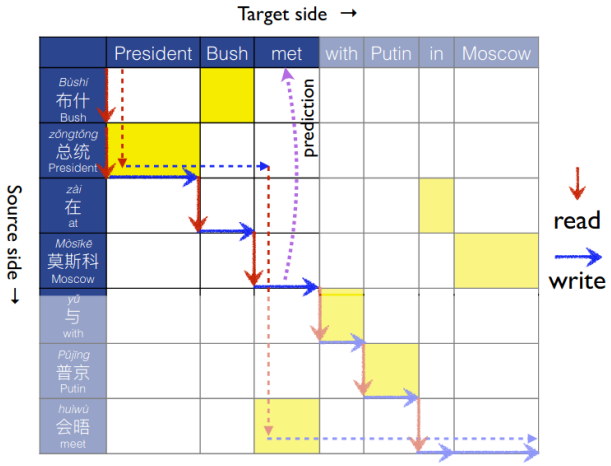


Figure 3.3: An example of the wait-k policy with k=2. Figure from (Ma et al., 2019).

sentence is always k words behind the input. This gives controllable latency and an implicit anticipation on the target side when decoding. The output is generated in a greedy manner, since each word is committed on-the-fly, however some attempts have been made to apply beam search (Zheng et al., 2019c; Zeng et al., 2021; Nguyen et al., 2021).

While the original wait-k implementation is based on textual source data, (Ma et al., 2020) adapted this strategy to audio input by waiting for k fixed amount of time (step size) instead of k words. The best step size resulting from their experiments was 280ms, corresponding to, approximately, the length of a word – on average 271ms – motivating the equivalence between the MT and the ST policies.

With *adaptive* policies (Cho and Esipova, 2016; Grissom II et al., 2014; Gu et al., 2017; Alinejad et al., 2018; Arivazhagan et al., 2019b; Zhang et al., 2022; Dong et al., 2022), a WRITE decision is taken by considering also the contextual information extracted from the input. Such a policy makes sure there is enough source information before generating a target word. Even if fixed policies are not optimal since they disregard the input context, they are still more popular than adaptive policies (Gu et al., 2017; Zheng et al., 2019a, 2020), because they are easier and less

time-consuming to train (Zheng et al., 2019b; Arivazhagan et al., 2019b). Due to their low latency and reduced maintenance, direct SimulST systems can be promising for the application scenario of live subtitling, which we explore in Section 5.4.

3.3 Automatic subtitle segmentation

We have seen in Section 2.2 that creating readable subtitles comes with several challenges. The difficulty imposed by the transition to a different semiotic means, which takes place when transcribing or translating the original audio into text (Sec. 2.1.1), is further exacerbated by the limitations of the medium (time and space on screen). Moreover, chuncking or segmentation, i.e. the way a subtitle is split across the screen, has a great impact on comprehension (Sec. 2.2.2). Where and if to insert a subtitle break depends on several factors such as speech rhythm, pauses but also semantic and syntactic properties. This all makes segmenting a full sentence into subtitles a complex and challenging problem.

Automatically segmenting text into subtitles has long been addressed as a post-processing step in a translation/transcription pipeline. In the subtitling industry, language-specific rules and simple algorithms are employed for this purpose. Among these are splitting based on simple character counts, e.g. before a specific length is reached, or splitting the sentence at the highest syntactic node (Karamitroglou, 2001). Other approaches on subtitle segmentation use classifiers to predict subtitle breaks. One of these approaches used Support Vector Machine and Logistic Regression classifiers on correctly/incorrectly segmented subtitles to determine subtitle breaks (Álvarez et al., 2014). Extending this work, Álvarez et al. (2017) trained a Conditional Random Field classifier for the same task, but in this case distinguishing between line and subtitle breaks. A more

recent, neural-based approach (Song et al., 2019) employed an LSTM to predict the position of the period for improving the readability of automatically generated YouTube captions, but without focusing specifically on the segmentation of subtitles. Specifically for subtitling, Matusov et al. (2019) customised an NMT system to subtitling and introduced a segmentation neural classifier based on human segmentation decisions from OpenSubtitles data and established penalties in the subtitling industry. In Section 4.3 we improve over these approaches by proposing a neural seq2seq segmenter model, while in Section 5.2, we propose an end-to-end ST solution which translates and predicts subtitle segmentation boundaries in one.

3.4 Machine translation for subtitling

In this section, I present research in the field of MT for subtitling. Compared to other, more popular applications of MT, subtitling has only been tackled in a handful of academic studies during the past 20 years. These works can be conveniently subdivided into three periods, for the categorisation of which I will use a three-age system according to the dominant MT paradigm defining each era: rule-based systems, statistical systems and neural systems.

The rule-based age: Early works employing MT for the multilingual generation of subtitles were optimistic in the degree of automation that could be achieved. Pipeline systems were constructed to perform the task from end to end. The architectures were based on several subcomponents, designed on the basis of rules. The MUSA project (Piperidis et al., 2005) automatised multiple steps in the subtitling process, by creating a pipeline comprising ASR, text condensation and constraint formulation, followed by rule-based MT. Similarly, the eTITLE project (Melero et al., 2006) created a web-based subtitling service and made steps towards automatising

the full workflow. The architecture was composed of an audio-to-transcript alignment module, translation with translation memories (TM) and MT, text compression and subtitle segmentation based on geometric rules (similar length of upper and lower line). The main limitation of these projects, as well as of other rule-based (Popowich et al., 2000) and example-based (O'Hagan, 2003; Armstrong et al., 2006) MT approaches to subtitling, was that MT technology at the time was not ripe enough to be fully deployed in professional settings.

The statistical age: Statistical methods led the efforts of the scientific community to move away from rule-based subcomponents in favour of data-driven approaches. As a result, the initial enthusiasm for fully automatising subtitling was abandoned and more realistic expectations were set, focusing on improving translation performance. The age of Statistical Machine Translation (SMT) gave birth to two large projects using SMT to translate subtitles with moderately encouraging outcomes. The Stockholm MT System for Film Subtitles (Volk et al., 2010) reported a success story. Its success can be attributed to a large corpus of proprietary data covering Scandinavian languages, the similarity of the languages involved and the collaboration of professional subtitlers. The EU-funded project SUbtitling by MAchine Translation (SUMAT) (Etchegoyhen et al., 2014; Bywood et al., 2017) trained SMT systems on professional and crowd-sourced data for eleven translation pairs. They reported productivity gains for some languages but recognised that the ground was not ready for the establishment of MT in the subtitling world. Unfortunately, the data used to train the systems in both projects were to a large degree proprietary, so they were not released to the research community, which impeded further studies on SMT for subtitling. Similar productivity gains were reported from smaller scale studies comparing academic SMT systems and commercial systems available at the time (de Sousa et al., 2011; Aziz et al., 2012). The

approach employed in this era can be regarded as automatising the translation of templates in the subtitling industry (Georgakopoulou, 2019). The soundtrack is transcribed by a subtitler, with the necessary degree of text compression, and spotted in order to create the source language subtitles which are then automatically translated. In this way, there is a one-to-one correspondence between source and target subtitles. Still, significant manual effort needs to be invested in the transcription and spotting of the source text speech.

The neural age: The unprecedented improvements in quality brought about by NMT resparked the interest in MT for subtitles. Researchers from AppTek (Matusov et al., 2019) customised a general purpose NMT system for the domain of subtitling. Along with the translation of the subtitles, the system incorporated a segmentation module which predicted line breaks. Despite the productivity gains reported, this approach also relied on the existence of a high quality template. The MeMAD project (Koponen et al., 2020c) employed a similar approach, applying MT at the level of sentence or document. This approach extended the context received by MT, leading to improved translation performance, but introduced the problem of aligning the translated text back to the source timestamps. The expansion of text in the target language and the segmentation based on punctuation often caused the alignment heuristics to fail. As a result, some subtitles were pushed to later frames, leading to the text falling out of sync with the soundtrack. The project also carried out experiments with ASR-MT cascades both for fully automatic subtitling and as a tool for subtitlers (cf. Section 6.4).[3] However, the results were rather discouraging, since considerable editing was needed and the subtitlers' user experience was overall rather negative. Cherry et al. (2021) treated subtitle translation as mark-

[3]https://memad.eu/wp-content/uploads/D6.9-Evaluation-Report-final-version-version-1.0-merged.pdf

up translation, by projecting the subtitle boundaries from the source to the output, using a projection algorithm based on word alignments from source to target. In this way, the properties of the source (timestamps, colours and other metadata) can be directly transferred to the output subtitles. It should be noted that all these approaches rely on the existence of high-quality source language captions.

All these works have provided valuable inspiration for the rising paradigm of automatic subtitling proposed in this thesis. Chapter 5 is dedicated to the direct ST models and their application to subtitling.

3.5 Evaluation

An indispensable part of system development is evaluation. Evaluation has been a thorny issue in MT since its nascence. Recent research has demonstrated flaws in widely-used MT evaluation methods, bringing about controversies in the presumably solid preconceptions on MT quality (Toral et al., 2018; Läubli et al., 2018; Fomicheva and Specia, 2019; Freitag et al., 2019, 2020). In this section, I describe MT evaluation methods, focusing on human (Sec. 3.5.1) and automatic evaluation (Sec. 3.5.2).

3.5.1 Human evaluation

Human evaluation is the 'gold' method for evaluating MT, even though the high costs and time required for the task make it unfeasible for frequent use. The natural intuition to judge the goodness of a translation is to ask humans to rate the quality of the output, comparing it to the input sentence. This requires bilingual evaluators, highly fluent in both the source and target language. Since the question '*how good is this translation?*' can be interpreted in various ways by evaluators, a common practice is to use two criteria: **Fluency**, which corresponds to the question '*Is the output*

good fluent English/Spanish etc?', and **Adequacy**, corresponding to '*Does the output convey the same meaning as the input sentence?*'. Evaluators judge the output on five or seven-point scales (Callison-Burch et al., 2007; Denkowski and Lavie, 2010). However, research has shown that there is broad variability in human judgements of fluency and adequacy (Callison-Burch et al., 2008, 2009). Instead of evaluating a system on an absolute scale, another evaluation method is to **rank** two system outputs against each other (Callison-Burch et al., 2012). This method leads to higher agreement among evaluators and is useful in evaluation campaigns where different systems compete with each other. The most recent way of evaluating MT output is **Direct assessment** (DA) (Graham et al., 2013). In reference-based DA, evaluators are presented with the reference translation and the system output and evaluate the goodness of the translation on a continuous scale (0-100). The benefits of DA are that monolingual evaluators can preform the task, it scales better and considers the variability between evaluators. DA is now the established evaluation method in WMT campaigns (Barrault et al., 2020; Akhbardeh et al., 2021).

Another perspective on evaluation is through quality assessment. Human annotators annotate the errors in an MT output based on an error typology. Several error annotation frameworks have been proposed with the most widely-used being Multidimensional Quality Metrics (MQM) (Burchardt, 2013). This framework provides a list of error types, hierarchically arranged in groups, with accuracy, verity and fluency as core concepts.

Lastly, task-based evaluation is aimed at whether a translation is good enough for allowing the user to accomplish a task. Such cases can be real-world tasks, for example, whether the translation of a form allows one to successfully submit an application, or content understanding (gisting) tasks, which consist in asking an evaluator to re-write the output based on what they have understood of it. Another form of task-based evaluation

focuses on the use of MT as a tool to increase productivity, through the task of post-editing, which is described in detail in the following section.

Post-editing Human post-editing (PE) consists in manually correcting the system output according to the input. In PE-based evaluation, instead of evaluating how good the output is, the focus is given to how much work is needed to fix it. The original output is compared against its post-edited version using an automatic metric, Human translation edit rate (HTER) (Snover et al., 2006). HTER calculates the minimum number of insertions, deletions, substitutions and shifts to convert the system output into the edited translation. This allows for counting only the actual errors made by a system, without penalising differences due to linguistic variation as it happens when exploiting independent references.

HTER is a product-based metric, meaning that it cannot measure the exact effort made by translators. For example, a translator may completely erase the MT suggestion and re-write the exact text. This would give a HTER of 0 even though the translator re-typed the entire text. For this reason, process-based metrics are also useful for evaluating productivity. Productivity gains are assessed by comparing temporal effort, as the time spent editing an MT output compared to translating the input from scratch. Another indicator of productivity is the technical effort, consisting in how much typing the translator actually performs, usually by counting the number of keystrokes with keystroke-logging software (e.g. InputLog (Leijten and Van Waes, 2013)). Even though more interpretable from a practical point of view, process-based evaluations are subject to high variability, because of the subjectivity among the translators. In addition, time is a sensitive measure and often affected by outliers, e.g. when a translator gets distracted or due to technical reasons related to the software. Still, with a careful experimental setting and proper data cleaning,

productivity-based evaluation gives a realistic estimation of the usefulness of MT in translators' workflows. Product- and process-based analyses are employed in our investigation of PE in automatic subtitling in Section 6.3.

3.5.2 Automatic evaluation

Automatic evaluation is based on automatic metrics, which measure the goodness of the output. Automatic evaluation is appealing because it is easier, faster and cheaper compared to human evaluations. It is reproducible, and can be used multiple times during system development to drive the implementation decisions towards performance gains. The goodness of each metric is judged by its correlation with human judgements.

Reference-based evaluation consists in comparing the system output (hypothesis or candidate translation) against one or several human translations (reference translations). The automatic metrics measure some degree of similarity with the reference. For years, string-matching algorithms prevailed, with the most prominent and long-standing being BLEU (Papineni et al., 2002). BLEU is a precision metric, which computes the ratio of n-grams in the translation that match the reference, multiplied by a brevity penalty which penalises too short translations. To address the fact that BLEU is based only on surface features, other metrics proposed using deeper linguistic features. METEOR (Lavie et al., 1996; Banerjee and Lavie, 2005) incorporates stemming and synonyms giving partial credit to near matches, but requires linguistic resources for each language.

Recently, semantic metrics have been gaining ground, which compute scores by comparing (contextual) embeddings extracted from pre-trained models. For example, Yisi (Lo, 2019) and BERTscore (Zhang et al., 2020b) measure the token similarity between the output and the reference(s). COMET (Rei et al., 2020) has the additional benefit that it also encodes the source, along with the MT hypothesis and reference, in the same cross-

lingual space. PRISM (Thompson and Post, 2020) casts MT evaluation as a zero-shot paraphrasing task, producing segment-level scores by force-decoding between output and reference. Most of these metrics have a high computational cost, since they rely on large pre-trained models.

Evaluation of MT for subtitling Evaluation for MT for subtitling has followed the steps of MT evaluation, since BLEU (Papineni et al., 2002) is the dominant automatic metric used in all studies on MT for subtitling. However, the majority of studies used other string-based metrics along BLEU, such as TER (Snover et al., 2006), characTER (Wang et al., 2016) and chrF (Popović, 2015). Moreover, some studies (Volk et al., 2010; Etchegoyhen et al., 2014) used multiple references for calculating the scores, to account for the variability in translation solutions. It is worth noting that the tendency of using multiple automatic metrics and references suggests that researchers recognise that string-based metrics are not sufficient for representing the functional equivalence required in creative and multimodal texts such as subtitles.

Recently, the need for evaluating the technical constraints of subtitling (e.g. segmentation, spotting) has lead to new automatic metrics. Aziz et al. (2012) computed how much the output text increases in length compared to the reference (in characters), however this does not give any indication for the degree of conformity to the subtitling constraints. In this thesis we employ conformity metrics to evaluate these technical aspects (Sec. 5.2). In an attempt to consider segmentation quality, Matusov et al. (2019) report MT scores in the *S-mode*, where line breaks in a subtitle are marked with a separator symbol. However, in their setting there is a one-to-one correspondence between source-target subtitles, which is not always the case with end-to-end generation systems. Cherry et al. (2021) propose two metrics again related to BLEU, to include timing quality: 1) *Timed-BLEU*,

where target to reference alignments necessary for the evaluation are created by linear temporal alignment, over which BLEU is calculated as usual, and 2) *T-BLEU Headroom*, calculated as the difference between an upper bound of T-BLEU and the actual T-BLEU. At the time of writing this thesis, Wilken et al. (2022) proposed Subtitle Edit Rate (SubER), a metric which considers all subtitling aspects, namely subtitle text, segmentation and timing. SubER is based on TER, which is extended by including break edit operations and time-alignment constraints. Both metrics can only be applied on outputs containing timestamps e.g. in an .srt format, which is not addressed with the models proposed in this thesis.

Evaluation of Subtitle Segmentation Automatic subtitle segmentation has been previously evaluated in the case of interlingual and intralingual subtitles, by comparing the automatically generated output against a reference (Álvarez et al., 2014). For interlingual subtitles, Álvarez et al. (2016) proposed a segmentation algorithm based on logistic regression and support vector machine classifiers. The evaluation was performed with precision-recall-F1 measures, based on the ability of the algorithm to insert a segmentation boundary, without however distinguishing between line and subtitle breaks. Later, Álvarez et al. (2017) compared rule-based and machine-learning segmentation methods with metrics considering either subtitle breaks only or both line and subtitle breaks. These metrics count the segmentation errors (F1, NIST), the number of incorrectly segmented portions (DSER) or the edit distance between sequences of reference positions and hypothesis positions (SegER). In Section 4.3.2 except for F1, we evaluate segmentation accuracy with BLEU as a similarity measure, and characters per line (CPL), as the percentage of segmented subtitles conforming to the length constraint. However, most of the segmentation metrics can only be applied when the output text perfectly matches the reference text, which

is not the case with end-to-end generation systems. In Section 6.2, we address this gap by proposing new subtitle segmentation metrics.

3.6 Conclusion

In this chapter I have set out the technical background for this thesis. I have described NMT (Sec. 3.1) and the Transformer, as well as the state of the art in speech translation (Sec. 3.2) and MT for subtitling (Sec. 3.4). Last but not least, evaluation metrics and methodologies were explained in Section 3.5. Having the main ingredients at hand, the reader is now ready to delve into the main part of this thesis, which addresses automatic subtitling from a computational perspective.

Chapter 4

Speech Subtitling Data

In this chapter, we present the first contribution towards the new paradigm of automatic subtitling, that is the compilation of suitable data for training automatic subtitling systems. We start with an examination of existing subtitling corpora and their suitability for training systems under the new paradigm (Sec. 4.1). The analysis shows that existing subtitling corpora lack several elements required for building end-to-end systems. This motivates the compilation of the MuST-Cinema corpus, a speech-to-subtitles corpus, based on TED talks. This corpus is an extension of a speech-to-text corpus (MuST-C) where subtitle boundaries are annotated as special symbols. The corpus, the steps for its compilation and an evaluation based on formal subtitling criteria is presented in Section 4.2. In Section 4.3, we demonstrate the applicability of MuST-Cinema for the task of subtitle segmentation, i.e. the segmentation of long segments of text into subtitles, compliant to specific guidelines. This technique, apart from the mentioned application, has the advantage of allowing for annotating existing (speech) translation corpora to generate synthetic subtitling data, both for languages seen during training and on unseen languages in a zero-shot fashion. In this way, we contribute towards alleviating the bottleneck of the lack of subtitling data for training automatic subtitling systems.

4.1 Are Subtitling Corpora really subtitle-like?

Building an end-to-end automatic subtitling system is interleaved with several challenges, the most prominent one being the fact that neural systems need large amounts of high-quality training data representative of the target application scenario (subtitling in our case). Automatic subtitling systems would require audiovisual data as input and timed text segmented into subtitles as output.[1] There are large amounts of available parallel data extracted from subtitles (Lison and Tiedemann, 2016; Pryzant et al., 2018; Božović et al., 2018; Di Gangi et al., 2019a; Serrat Roozen and Martínez Martínez, 2020). But are subtitling corpora really subtitle-like? Subtitling corpora are usually obtained by collecting files in a subtitle specific format (usually SubRip *.srt*) in several languages and then parsing and aligning them at sentence level. Sentence-level segmentation and alignment has advantages and disadvantages. On the one hand, MT training at sentence level generally achieves good performance as the system receives longer context, which can be useful, for instance, to disambiguate words and resolve coreference. On the other hand, this process compromises the subtitle format by converting the separate subtitle blocks into full sentences. As shown in Table 4.1, two subtitle blocks with two subtitle lines each have been merged to form three sentences. With this 'merging', information about subtitle segmentation (line breaks) is often lost. Therefore, restoring the MT output to a proper subtitle format has to be performed subsequently, either as a post-editing process or by using hand-crafted rules and boundary predictions.

In this section, we explore whether the large, publicly available parallel data compiled from subtitles are suitable for training and testing automatic subtitling systems. Specifically, we test whether they conform to the

[1]The incorporation of visual information in NMT systems, although promising, is outside the scope of this thesis.

1 00:00:14,820 −− > 00:00:18,820 Grazie mille, Chris. É un grande onore venire 2 00:00:18,820 −− > 00:00:22,820 su questo palco due volte. Vi sono estremamente grato.	Grazie mille, Chris. É un grande onore venire su questo palco due volte. Vi sono estremamente grato.

Table 4.1: Subtitle blocks (left, 1-2) as they appear in an .srt file and the processed output for obtaining aligned sentences (right).

temporal and spatial constraints necessary for achieving quality subtitles (see Section 2.2.2). For evaluating length and reading speed, we employ character counts, while for proper line breaks we inspect sequences of Part-of-speech (POS) tags. Based on the analysis, we discuss limitations of the existing data and present a roadmap towards creating resources faithful to the subtitling format for training automatic subtitling systems.

4.1.1 Existing subtitling data

Several projects have attempted to collect parallel subtitling corpora. The most well-known one is the OpenSubtitles[2] corpus (Lison and Tiedemann, 2016), extracted from 3.9 million subtitle files across 60 languages. Since subtitle blocks do not always correspond to sentences (see Table 4.1), 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 an algorithm based on overlapping timestamps (Tiedemann, 2008) and bilingual dictionaries. Although the 2018 version of OpenSubtitles (Lison et al., 2018) has sentence alignments of higher quality, it does not resemble the realistic subtitling scenario de-

[2]http://www.opensubtitles.org/

scribed above, since time and line break information are lost in the merging process. 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 desired scenario. On the other hand, non-matching alignments are discarded, which might hurt the integrity of the subtitling documents. As we will show in Section 4.3, this is particularly important for learning proper line breaks between subtitle blocks.

Deviating from the domain of films and series, TED Talks are a valuable resource for compiling ST corpora. MuST-C (Di Gangi et al., 2019a) is among the largest multilingual ST corpora, compiled from TED talk subtitles. It contains {*audio, source language transcription, target language translation*} triplets, aligned at segment level. The process of creation is as follows; the authors first aligned the transcription with the translation and force-aligned the audio to the transcribed segments. MuST-C is a promising corpus for training automatic subtitling systems which translate the audio directly into subtitles. However, since the subtitles are merged to create sentences, they are far from the desired subtitle format. Given the challenges discussed above, we perform a systematic study of the suitability of the existing corpora for building automatic subtitling systems.

4.1.2 Conformity criteria

To assess the conformity of the existing subtitle corpora to the formal subtitling constraints, we consider the following criteria. The criteria were defined on the basis of the TED guidelines. Even though the corpora come from different domains using different sets of guidelines, the TED guidelines are in general more 'relaxed' (see Sec. 2.2.2) and thus represent

the minimum acceptable requirements to assess the 'subtitle-likeness' of the corpora.

1. **Characters per line (CPL):** As mentioned above, 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 conforms to the length constraint by calculating the maximum possible length for a subtitle block. According to the TED guidelines, this is 2 * 42 = 84 characters for Latin scripts and 2 * 14 = 28 for Japanese. If $CPL > max_length$ then the subtitle does not conform to the length constraint.

2. **Characters per second (CPS):** This criterion relates to reading speed. For the corpora where time codes and duration are preserved, we calculate CPS as follows: $CPS = \frac{\#chars}{seconds}$. If $CPS > 21$ the subtitle is non-conforming.

3. **Chink-Chunk (CC):** For assessing the preservation of linguistic wholes we use Chink-Chunk (Liberman and Church, 1992), a low-level parsing algorithm, as a simple but efficient way to detect syntactic boundaries. It uses POS information to detect the end of a unit only at punctuation marks (logical completion) or when an open-class or content word (chunk) is followed by a closed-class or function word (chink). For example, in the sentence of Table 4.1, the subtitle break comes between *venire*, VERB-a content word, and *su*, PREP-a function word, and thus counts as a valid break. Here, we use Chink-Chunk to compute statistics about the type of block breaks in the data (punctuation break, content-function break or other). A subtitle is conforming if the block break before and after fall into the punctuation or content-function break.

For our investigation we consider the corpora which are large enough to train neural systems; OpenSubtitles, JESC and MuST-C. We focus on

3 language pairs, Japanese, Italian and German, paired with English at the source, as languages coming from different families and having a large portion of sentences in all corpora. We tokenise and then tag the data with Universal Dependencies (Nivre et al., 2020) using the Stanza toolkit(Qi et al., 2020) to obtain POS tags for the Chink-Chunk algorithm.

4.1.3 Results

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	-	95 / 94	-	99 / 99	91
DE	MuST-C	segment	✓	51 / 47	77 / 66	99 / 99	42
	OpenSubtitles	segment	-	95 / 95	-	99 / 99	92
JA	OpenSubtitles	segment	-	96 / 93	-	99 /98	91
	JESC	subtitle	-	97 / 94	-	88 / 87	85

Table 4.2: Percentage of data conforming to each of the quality criteria on the corpora. Percentages are given on source and target side (*s/t*), except for the *Total* where source and target are combined.

Table 4.2 shows the percentage of conforming subtitles/segments after applying each criterion. Both OpenSubtitles and JESC conform to the quality criterion of length in at least 93% of the cases. Despite the merging operations to obtain sentence alignments, OpenSubtitles still preserves a short length of lines, possibly because most subtitles come from the film domain and therefore contain short and quick dialogues. On the other hand, since time information is not present in JESC or in the parallel version of OpenSubtitles, assessing reading speed conformity is not possible. For MuST-C, length conformity is poor. The merging operations in MuST-C create long sentences that do not resemble the subtitling format. As for reading speed, except for Japanese, where the allowed number of charac-

ters per line is lower, the CPS for Italian and German is ≤21 in 66%-78% of the subtitles. MuST-C and OpenSubtitles conform to the criterion of linguistic wholes in 99% of the sentences, which does not occur in JESC, despite its subtitle format. JESC is compiled by removing captions based on unmatched time codes, therefore the integrity of the documents is possibly broken. An inspection of the types of breaks based on CC shows that OpenSubtitles and MuST-C end in a punctuation mark in 99.9% of the cases. Even though logical completion is preserved, these corpora do not contain a large amount of examples with line breaks inside the sentences, a necessary ingredient for generating readable subtitles.

To sum up, in this section we explored whether the existing parallel subtitling resources conform to subtitling constraints. We found that subtitling corpora generally conform to length and proper segmentation, despite the merging operations for aligning parallel sentences. We isolated some missing elements: the lack of time information (utterance duration, timestamps), which does not allow for modelling spotting, and the insufficient representation of line breaks other than at punctuation marks, which is a bottleneck for proper segmentation.

This raises several open issues for creating automatic subtitling corpora; 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 sentence/phrase often occupies several subtitles, 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. Break information could be preserved by inserting special symbols. Based on this roadmap, in the next session we address the lack of representative corpora

for automatic subtitling by creating MuST-Cinema, a speech-to-subtitles corpus.

4.2 The MuST-Cinema corpus

In the previous section, we found that existing subtitling corpora are insufficient for developing automatic subtitling solutions for at least two reasons; first, because they do not provide access to the source language audio or to information about the duration of a spoken utterance, and second, because line breaks between subtitles were removed during the corpus compilation process to create parallel sentences. In this section, we address these limitations by presenting MuST-Cinema, a multilingual speech translation corpus where subtitle breaks have been annotated with special symbols. The corpus is unique of its kind, since, compared to other subtitling corpora, it provides the source language audio, which is indispensable for automatically modelling spatial and temporal subtitling constraints.

4.2.1 Corpus compilation

MuST-Cinema is built on top of MuST-C (Di Gangi et al., 2019a), by annotating the transcription and the translation with special symbols, which mark the breaks between subtitles. MuST-Cinema is compiled for 7 language directions: English→Dutch, French, German, Italian, Portuguese, Romanian and Spanish. Here we describe the steps for its compilation.

Mapping sentences to subtitles As described in (Di Gangi et al., 2019a), the MuST-C corpus contains audio segments, transcriptions and translations obtained by processing the TED videos and the source and target language SubRip files (.srt). It was compiled by concatenating the textual parts of all the .srt files for each language, splitting the resulting text according to

strong punctuation and then aligning it between the languages. The source sentences were then aligned to the audio by using the Gentle software.[3] Although this sequence of steps allowed the authors to release a parallel corpus aligned at sentence level, the source and target subtitle structure and time information were not preserved. To overcome this limitation of MuST-C and create MuST-Cinema, we implemented a procedure to restore the subtitle structure by projecting the subtitle breaks from the .srt files on the MuST-C sentences:

- All the subtitles in the original .srt files obtained from the ted2srt website[4] are loaded in an inverted index (text, talk ID and timestamps).
- For each sentence in MuST-C, the index is queried to retrieve all the subtitles which: i) belong to the same TED talk of the sentence query and ii) are fully contained in the sentence query.
- An identical version of the query is created by concatenating the retrieved subtitles, and by adding special symbols to indicate the breaks.[5]

Inserting breaks When reconstructing the sentences from the .srt files, we insert special symbols to mark the line and block breaks. As explained in Section 2.2.2, we distinguish between two types of breaks: 1) subtitle block breaks, i.e. breaks denoting the end of the subtitle on the current screen, represented by the tag <eob>, and 2) line breaks, i.e. breaks between two consecutive lines (only for 2-liners) inside the same subtitle block, represented by the tag <eol>. In the case of one-liners <eob> is

[3] https://github.com/lowerquality/gentle

[4] https://ted2srt.org/

[5] An alternative to this process would be to re-align the subtitles according to the MuST-C sentences and insert the breaks based on this alignment. However, not all sentences present in the .srt files are included in MuST-C. This would lead to text which does not correspond to the audio files being inserted in MuST-Cinema.

used. Figure 4.5 shows an example of a sentence after inserting the breaks based on the corresponding .srt file.

164
00:08:57,020 → 00:08:58,476 164
I wanted to challenge the idea

165
00:08:58,500 → 00:09:02,060
that design is but a tool
to create function and beauty.

I wanted to challenge the idea <eob> that design is but a tool <eol>
to create function and beauty. <eob>

Figure 4.1: Subtitle file (top) and the full sentence with subtitle breaks (bottom).

Development and test sets For the development and test sets of MuST-Cinema, our goal was to provide a representative, high-quality dataset to promote research on automatic subtitling. For this, we aimed at 1) ensuring the translation quality by manually checking the text/audio data against the video, 2) avoid having subtitles above the character limit or subtitles with missing breaks and 3) preserve the integrity of the talks by including all segments. For these reasons we created new development and test sets for MuST-Cinema based on talks not included in MuST-C, for which we obtained the subtitles offered through the Amara repository.[6] We selected the subtitle files from Amara, because, compared to ted2srt, they contain the most updated .srt files actually used in the TED videos.

For the test set, we manually selected 9 talks with subtitles available in all 7 languages, published after April 2019 to avoid any overlap with the training data. We hence obtained a common test set for all the languages. The talks were translated by different TED translators and reviewers. The

[6] https://amara.org/en/teams/ted/videos/

.srt files were transformed to .txt files annotated with subtitle breaks. All the subtitles were merged and the text in the source (English) was sentence-split at strong punctuation. Then, each language text was separately aligned to the English transcription using InterText (Vondřička, 2014) to obtain parallel sentences and the alignment was manually corrected. As in MuST-C, the source sentences were aligned to the audio using Gentle.

The same steps were performed for the development set (10 talks), with the difference that we manually selected talks for each language without the requirement that the English talk has translations for all the languages. Therefore the development sets may differ among languages in a couple of talks, but the size of the set was kept similar for all languages, except for Romanian for which only five talks not included the training data were translated at the time of compilation.

4.2.2 Corpus structure

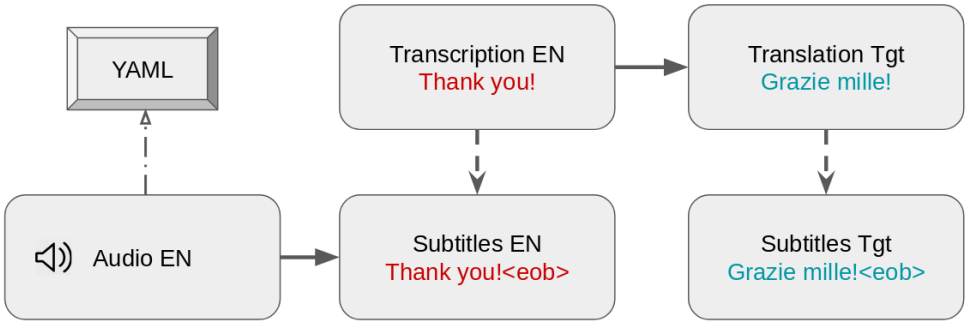


Figure 4.2: Structure of MuST-Cinema.

The structure of MuST-Cinema is shown in Figure 4.2. There are connections at two levels; first, on the horizontal axis, the corpus is composed of triplets of audio, source language transcription annotated with subtitle breaks (*Subtitles EN*), target language translation annotated with subtitle

breaks (*Subtitles Tgt*). Both the transcription and the translation are full sentences, aligned between English and the target language, as in MuST-C. On the vertical axis, there is a monolingual mapping between sentences with and without subtitle breaks: the English transcription is paired to the English subtitles of the same sentence and the target language (Tgt) translation is paired to the Tgt subtitles of the same sentence. The utterance duration (start and end time) of each sentence is stored as metadata in a YAML file. The structure of MuST-Cinema makes it optimal for several tasks related to subtitling, from subtitle segmentation (Section 4.3) to end-to-end automatic subtitling (Chapter 5). The statistics of the corpus (train, development and test sets) are presented in Table 4.3. The corpus has been officially released under a Creative Commons licence.[7]

	Train		**Dev**		**Test**	
Tgt	**sents**	**tgt words**	**sents**	**tgt words**	**sents**	**tgt words**
De	229K	4.7M	1088	19K	542	9.3K
Es	265K	6.1M	1095	21K	536	10K
Fr	275K	6.3M	1079	22K	544	10K
It	253K	5.4M	1049	20K	545	9.6K
Nl	248K	4.8M	1023	20K	548	10K
Pt	206K	4.2M	975	18K	542	10K
Ro	236K	4.7M	495	9K	543	10K

Table 4.3: Number of sentences and target side words for each language of MuST-Cinema for the training, development and test set for all 7 languages.

An initial analysis of the corpus showed that almost 70% of the sentences in MuST-Cinema contained at least one subtitle above the 42-character limit (first column in Table 4.4). In addition, the subtitle blocks in the .srt files downloaded from ted2srt are only one-liners. When comparing the subtitles in the .srt files with the video published on the TED website,

[7] `https://ict.fbk.eu/must-cinema/` - License: CC BY-NC-ND 4.0

Language	Conforming 42	Conforming 84	Lines with <eol>
De	28%	91%	2512 (0.01%)
Es	31%	91%	6103 (0.02%)
Fr	27%	90%	7563 (0.02%)
It	29%	93%	1009 (0.004%)
Nl	33%	94%	377 (0.002%)
Pt	30%	94%	1218 (0.006%)
Ro	34%	94%	522 (0.002%)

Table 4.4: Percentage of sentences containing only conforming subtitles based on the 42 and 84 CPL criterion, number of sentences containing <eol>symbols and the percentage of these sentences in the corpus (in parentheses).

we noticed that the line breaks inside subtitle blocks were removed from the .srt files (for reasons not explained). We observed though, that in some talks the subtitle blocks contained a double space, which corresponds to the <eol> in the videos. We replaced the double spaces with the <eol> symbol. We concluded that the reason for this lack of conformity is that the subtitle lines inside the blocks were collapsed into one line. In fact, when we analysed the conformity with a maximum CPL of 84 (2∗42), more than 90% of the sentences were conforming (second column Table 4.4). The third column in the table shows the number of sentences where the <eol> symbol is present. This is a small portion of the data, however, in Section 4.2.3 we will show that it can be valuable for learning subtitle line segmentations. This is an issue present only in the training data, as the development and test sets were downloaded from the TED Amara website, where the subtitle lines inside the .srt files are not collapsed. One way to deal with this limitation would be to attempt to recreate MuST-C from the Amara website. However, this would require building and testing a new pipeline from scratch and weeks for downloading and processing the dumps. Instead, we opted for a more scalable solution, also re-applicable

to new data; we re-annotate the existing training data with a model that learns to insert the <eol> symbol.

4.2.3 Iterative re-annotation of line breaks

We address the lack of <eol>in MuST-Cinema with a self-learning approach (Blum and Mitchell, 1998; Zoph et al., 2020; Lakew et al., 2021), by applying a segmenter model to iteratively re-annotate the training data only with the missing <eol> symbols. The segmenter is a seq2seq2 model, which receives input sentences already containing the <eob> symbols and outputs the same sentences additionally annotated with <eol>. This is because we wanted to preserve the existing block segmentation while splitting the blocks into lines. More details about the model and training can be found in Section 4.3. The process is carried out as follows:

1. The segmenter is first trained on all available MuST-Cinema data separately per language (*base*) and then fine-tuned on the small portion of sentences containing <eol> symbols (last column in Table 4.4) (*ft_eol*).

2. Using the *ft_eol* models, we annotate sentences in the training data which do not respect the 42 CPL constraint with the <eol>symbol.

3. We filter the annotated data with the CPL$\leq$ 42 criterion and select the length-conforming sentences containing <eol> symbols. We also make sure that no sentences contain two or more consecutive <eol> symbols.

4. Finally, we concatenate the selected data from step 3 with the initial data containing <eol>and fine-tune the *base* model again. The procedure can be iterated until no more sentences can be successfully annotated. Our intuition is that, even though sub-optimal due to possible insertion errors done by the system, the new training instances collected at step t can be used to improve the model at step *t+1*.

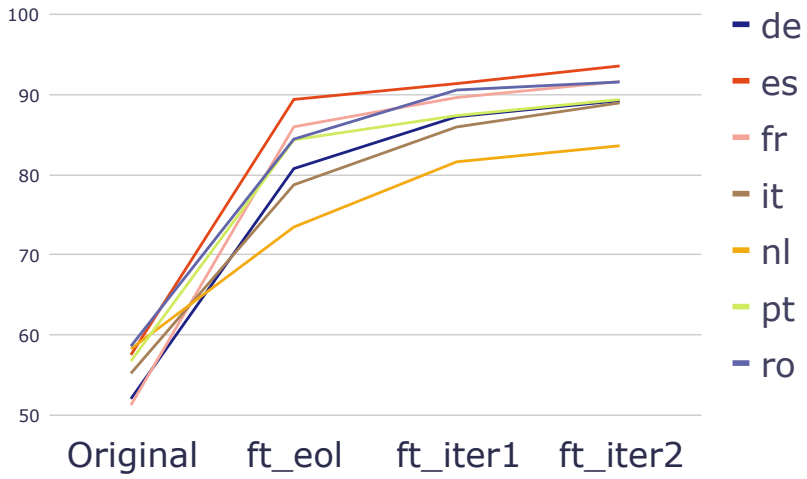


Figure 4.3: Percentage of subtitles in the MuST-Cinema training data conforming to CPL$\leq$ 42 after segmenting with different models.

Figure 4.3 shows the percentage of length-conforming subtitles for the MuST-Cinema training data at 4 stages; at the initial corpus compilation (*Original*), after segmenting the data with the model fine-tuned on small amounts of data with <eol> (*ft_eol*), and after 1 and 2 iterations of the method proposed above (*ft_iter*). There is a clear increase in the number of length-conforming subtitles, with all languages reaching a CPL conformity above 83% after only one iteration. This shows that training data can be efficiently and automatically annotated with the missing line breaks.

4.2.4 Evaluation and analysis

We perform an analysis of the MuST-Cinema corpus, focusing on the three formal aspects of subtitling quality: subtitle length, reading speed and preservation of linguistic wholes. We explore this in breadth, for all the 7 target languages of MuST-Cinema and the English source.[8] Table 4.5 shows the conformity to these constraints.

[8]The English side has been taken from the French parallel set, since it is the one with the largest size.

Language	CPL (%)	CPS (%)	CC (%)
De	89	80	87
Es	94	86	77
Fr	92	79	84
It	89	84	88
Nl	84	87	91
Pt	89	92	92
Ro	91	87	91
En	95	89	90

Table 4.5: Subtitling conformity statistics for the 7 target languages of MuST-Cinema and the source (English).

For length, all languages reach a conformity above 84%, reaching up to 95% for English. This is quite a high percentage given that the <eol>symbols have been annotated automatically. The percentages are close to the conformity of 94% for Italian and 95% for German of OpenSubtitles in Table 4.2. A loss in conformity is due to a small number of talks containing a large percentage of non-conforming subtitles. In (Karakanta and Orrego-Carmona, in press) we investigated length conformity diachronically for the subtitles of TED Talks. We found that talks with long subtitles come from the earlier years of TED translators programme and the non-conformity is due to the fact that their subtitles 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 the subtitling tool Amara and a stricter quality assurance process. In fact, all talks after 2012 – the year Amara was introduced – have almost 100% conformity to the length constraint (Figure 4.4).

Conformity to the criterion of reading speed is achieved to a lesser degree, as shown by CPS. Conforming to reading speed seems to be less

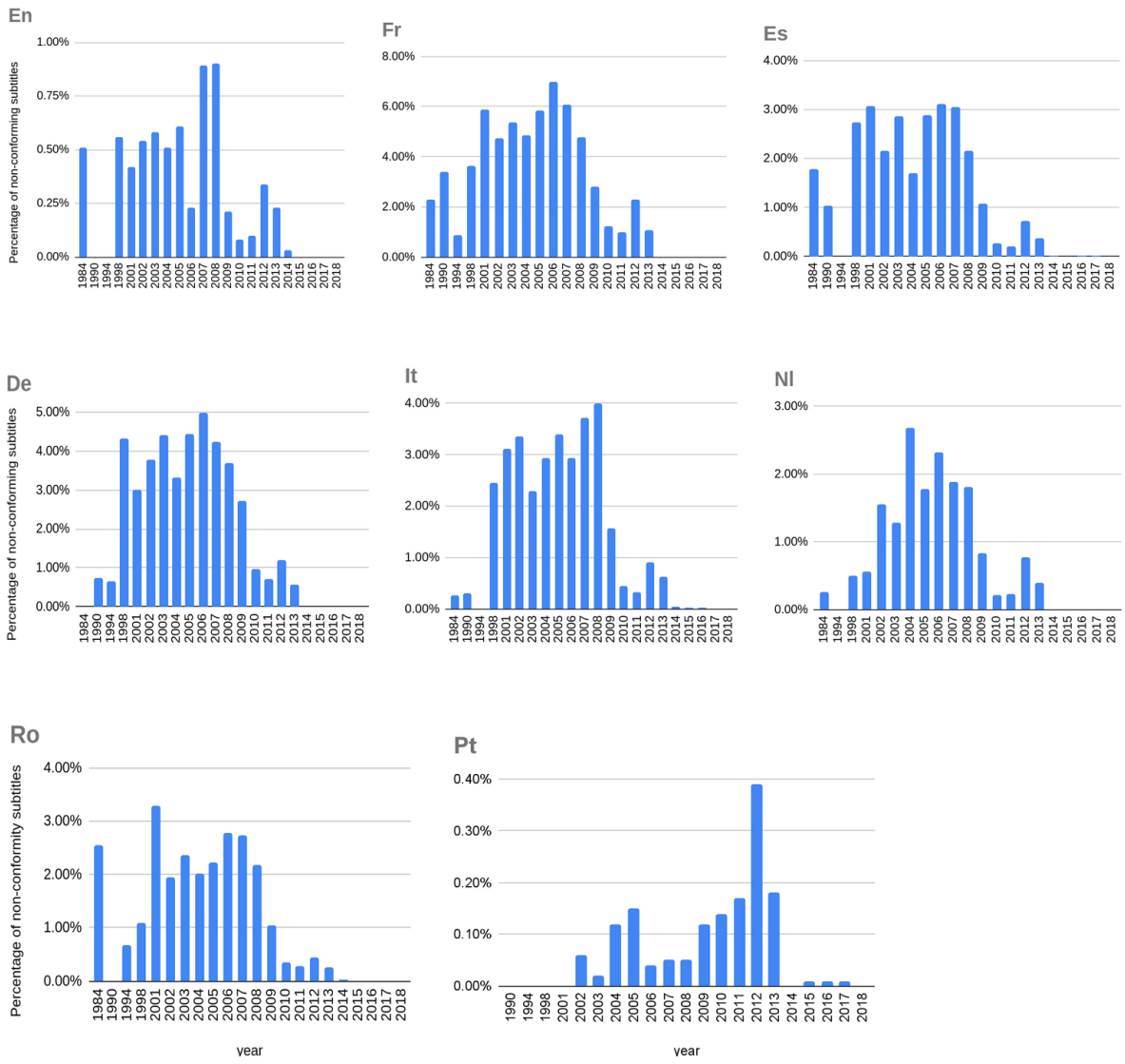


Figure 4.4: Percentage of subtitles longer than 84 characters per subtitle block for each year. Years containing no talks have been removed from the graphs.

important or more difficult to respect by volunteer subtitlers. Moreover, estimating the reading speed is laborious and difficult to be performed outside a subtitling tool. This is a possible explanation for the higher non-conformity for reading speed compared to length. As for the preservation of linguistic wholes, MuST-Cinema shows an acceptable segmentation with above 77% of breaks being in valid positions. It should be kept in mind that segmentation is affected by multiple factors which cannot be fully con-

sidered by Chick-Chunk as a low-level algorithm. In addition, this score could be affected by the performance of the POS-tagger.

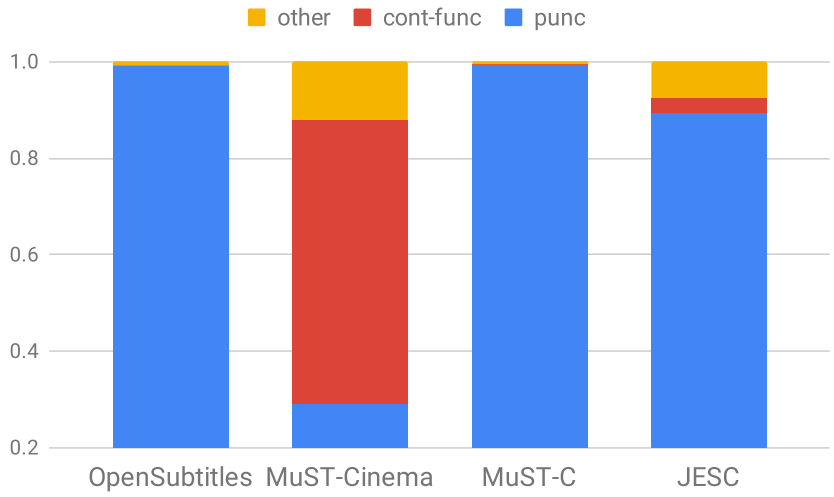


Figure 4.5: Percentages of the positions where breaks occur in the total breaks in each corpus. *punc*: the break occurs after a punctuation mark. *cont-func*: the break occurs between a content and function word). *other*: the break occurs in another position.

In Table 4.2 we saw that OpenSubtitles and MuST-C scored above 98% conformity according to Chink-Chunk. To provide an explanation for this high score, we analyse the categories of breaks based on the Chink-Chunk algorithm. The percentages for the German and Italian parts of OpenSubtitles, MuST-C and MuST-Cinema and Japanese for JESC are shown in Figure 4.5. As mentioned before, segments in OpenSubtitles and MuST-C end in a punctuation mark in 99.9% of the cases. A similar tendency is shown for JESC, where 90% of the sentences end in a punctuation mark, but the content-function category represents only 3%. In 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 properly segment.

To conclude, MuST-Cinema incorporates elements missing from other subtitling corpora as described in Section 4.1, by: i) providing the duration of the utterance, both through the alignment to the audio and as metadata, ii) showing high conformity to the subtitling constraints, iii) including subtitle breaks as annotations, iv) preserving the integrity of the documents, since the talks in the dev and test sets were used as a whole and without shuffling the sentences, and v) offering a reliable gold standard for evaluation of subtitling-related tasks. MuST-Cinema provides us with the necessary ingredient to experiment with automatic subtitling systems for different scenarios (Chapter 5) and improve evaluation methodologies (Chapter 6). In the next section, we describe how MuST-Cinema can be used for the task of subtitle segmentation.

4.3 Automatic subtitle segmentation

In the previous section, a segmenter model was used to automatically annotate MuST-Cinema with the missing <eol> symbols. In this section, we develop segmenter models for automatic subtitle segmentation, which consists in the segmentation of a full sentence into a sequence of subtitles, by inserting both <eol> and <eob> symbols. These models, apart from the task of converting sentences into readable subtitles, can be used for annotating existing translation corpora with subtitle breaks with the goal to synthesise new subtitling resources.

Previous work has performed subtitling segmentation with rule-based or statistical machine-learning approaches (see Section 3.3 for a full description of automatic segmentation approaches). Here, we address this task as a sequence-to-sequence problem in a monolingual fashion; the input to

the segmenter is a full sentence and the output is the same sentence annotated with <eol> and <eob> symbols. We present three experiments using segmenter models. The first experiment (Sec. 4.3.1) compares the performance of a segmenter trained on the original MuST-Cinema data (only human segmentation retrieved from the .srt files) with the same segmenter trained on the final MuST-Cinema data (human segmentation plus automatic <eol> annotation). The second (Sec. 4.3.2) explores heterogeneous data mixes of MuST-Cinema and OpenSubtitles for identifying the data mix leading to the highest segmentation performance, also under low-resource conditions. The third experiment (Sec. 4.3.3) extends the textual segmenter to a multimodal and multilingual one, with the aim of *i)* improving performance by exploiting audio information and the human segmentation decisions from all available data, and *ii)* increasing efficiency by maintaining only one model, able to accurately segment even in zero-shot conditions.

4.3.1 MuST-Cinema annotation

In Section 4.2.3 we showed that automatically annotating MuST-Cinema with the missing <eol> symbols led to an increase in length conformity. However, this increase does not necessarily imply a good segmentation. In this experiment, we compare the accuracy of a segmenter trained on the original retrieved MuST-Cinema segmentations compared to the released version, where <eol> have been automatically annotated. The goal is to show that the automatic annotation of <eol> can still lead to high segmentation accuracy when the corpus is used as training data.

Experimental settings We train a sequence-to-sequence model based on the Transformer architecture (Vaswani et al., 2017) using fairseq (Ott et al., 2019). For the original data, the model is first trained on all the avail-

able data before automatically annotating the <eol>symbols (*All*) and then fine-tuned on the small portion of sentences (third column in Table 4.4) containing the <eol> symbols (*ft_eol*). This two-step procedure was applied to force the NMT system to take into consideration also the <eol> symbols. This was necessary considering the mismatch in the number between sentences with and without <eol> symbols. For the automatically annotated data, we train the same model only on MuST-Cinema (*MC*). For optimisation, we use Adam (Kingma and Ba, 2015) (betas 0.9, 0.98), and dropout (Srivastava et al., 2014) is set to 0.3. We train the initial model (*All*) for 12 epochs and fine-tune for 6 epochs, while (*MC*) is trained for 18 epochs.

For each of the source and the 7 target languages of MuST-Cinema, we segment the data into subword units with SentencePiece (Kudo and Richardson, 2018) with a different model for each language and 8K vocabulary size. The break symbols are kept as a single word-piece. Accuracy scores are computed on the Must-Cinema test set.

We compare the performance of the Seq2seq segmenter against three baselines. First, a rule-based method (*Count Chars*) where a break is inserted in the first space before a 42-character limit is reached, as per TED guidelines. If the previous break is an <eob>, it randomly selects between <eob> and <eol>, while, if the previous break is an <eol>, it inserts an <eob>. This choice is motivated by the fact that, due to the constraint of having only two lines per block, we cannot have a segmentation containing consecutive <eol> symbols. The last break is always an <eob>.

The second baseline is a segmenter based on the *Chink-Chunk* algorithm, incorporating length and syntactic rules. Given a sequence of POS-tags, the algorithm finds all possible segmentation positions. These are positions after a punctuation mark and between a content and a function word. Then, if the length of two consecutive subtitles separated by these

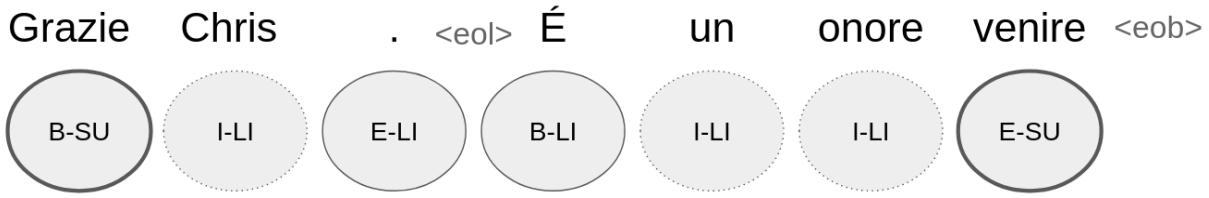


Figure 4.6: Class labels for the NCRF model. B-SU: beginning of subtitle, I-LI: in-line word, E-LI: end of line, B-LI: beginning of line, E-SU: end of subtitle.

segmentation positions does not exceed 42 characters, the subtitles are merged. This process ends when no more subtitles can be merged without exceeding the 42 limit. The choice between <eol> and <eob> is done as in the *Count Chars* baseline.

The third baseline is a multi-class classifier based on the Conditional Random Field (CRF) segmenter by Álvarez et al. (2017). Here, subtitle segmentation is seen as a sequence labelling problem, where each word is assigned a class label depending on its position in a subtitle: beginning of subtitle/line, end of subtitle/line, in-line. An example is shown in Figure 4.6, while see (Álvarez et al., 2017) for a full description of the eight class labels. The classifier is a neural CRF (NCRF), based on the NCRF++ toolkit (Yang and Zhang, 2018). It is composed of a CNN character sequence layer and a bi-directional LSTM word sequence layer followed by a CRF inference layer. During decoding, the label sequence with the highest probability is found using the Viterbi algorithm (Viterbi, 1967). The word embedding size is set to 300 while the character embedding size to 30. The hyperparameters are as follows: 4 CNN 50-dimensional character layers and 1 bidirectional 200-dimensional LSTM word layer, drop-out of 0.5 and a batch size of 10. For optimisation, Adam is used with an initial learning rate of 0.0001. The models are trained until performance on the dev set does not improve for four consecutive epochs. The NCRF model is trained only on the automatically annotated MuST-Cinema. The data is

tokenised using the Moses tokeniser and converted to CoNLL 2003 format. Pre-trained word embeddings were obtained from FastText (Grave et al., 2018). The output is detokenised and subtitle breaks are added in the predicted positions for a comparable evaluation with the other models.

Evaluation The evaluation of segmentation accuracy, i.e. how good the model is at inserting a sufficient number breaks and at the right places, is performed through **precision**, **recall** and **F1**-score, as follows:

$$Precision = \frac{\#correct_breaks}{\#total_breaks(output)} \tag{4.1}$$

$$Recall = \frac{\#correct_breaks}{\#total_breaks(reference)} \tag{4.2}$$

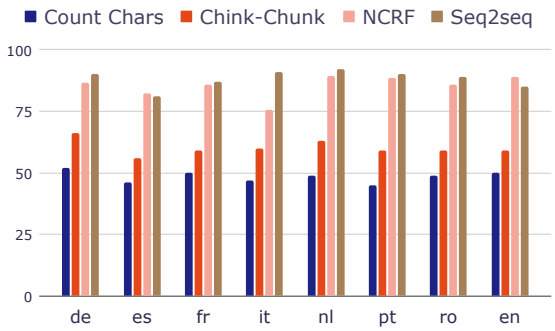
$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{4.3}$$

However, F1 score has the disadvantage that it cannot distinguish between the two types of breaks. For this reason, <eol>and <eob>are treated as the same segmentation token in the computation of F1. To address this limitation of F1, we additionally report BLEU (Papineni et al., 2002) between the human reference and the output of the system. BLEU is estimated on the raw output containing break symbols (**BLEU**$_{br}$). In BLEU$_{br}$every correctly placed segmentation token contributes to the n-gram overlap between reference and output. Higher values indicate a higher similarity between the system segmentation and the desired segmentation.

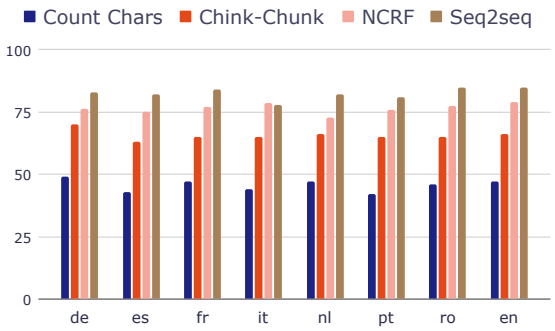
Finally, we want to test the ability of the system to segment the sentences into properly formed subtitles, that is measuring how well the system conforms to the constraint of length and number of lines. We compute **CPL** as the percentage of subtitles conforming to a length of ≤ 42.

Results We first compare the accuracy of the different segmentation methods (*Count Chars*, *Chink-Chunk*, *NCRF* and *Seq2seq*) to identify the highest-performing segmenter architecture. The results are shown in Figure 4.7. All segmentation metrics (Prec, Rec, F1, BLEU_{br}) agree in that the Count Chars baseline performs poorly, barely reaching scores of 50. Incorporating syntactic information with the Chink-Chunk method benefits segmentation accuracy, as demonstrated by an increase of 15 F1 points and 11 BLEU_{br}points on average across languages. Despite achieving perfectly length-compliant subtitles as shown by a CPL conformity of 100%, in terms of segmentation accuracy both rule-based methods are greatly outperformed by the models learning from data, whose segmentation scores are above 75. When comparing the NCRF and Seq2seq methods, Seq2seq achieves higher F1 scores for each of the languages and on average (83.4 versus 80.6 for the NCRF). In terms of BLEU_{br}, which considers the type of break, the results are mixed, with Seq2seq scoring higher for four out of the eight languages (es, nl, pt, ro) and on par for Italian, while the average score across languages is on par (83.7 versus 84 for the NCRF). The decisive factor over selecting the Seq2seq method as the best segmentation strategy is length conformity, where the subtitles generated by the Seq2seq model conform to the length constraint on average in 92.5% of the cases, compared to only 84.7% for NCRF, a score which is corroborated by a lower Recall. Given the comparable accuracy and the fact that length conformity is important for user experience, we select the Seq2seq segmenter as the architecture to experiment with subtitle segmentation.

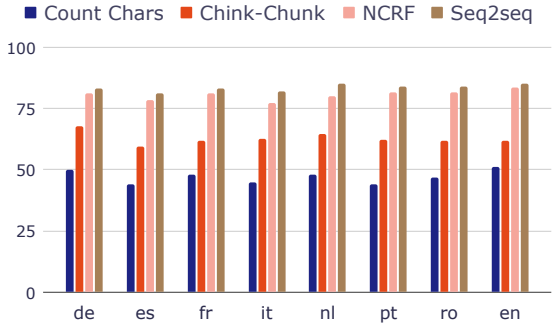
Having detected the best architecture for subtitle segmentation, we compare the accuracy of the Seq2seq segmenter trained on the original retrieved MuST-Cinema segmentations compared to the released version containing the automatic <eol> segmentation. The results are shown in Figure 4.8. In terms of F1, the fine-tuned model (ft_eol) is better than *All* with 6



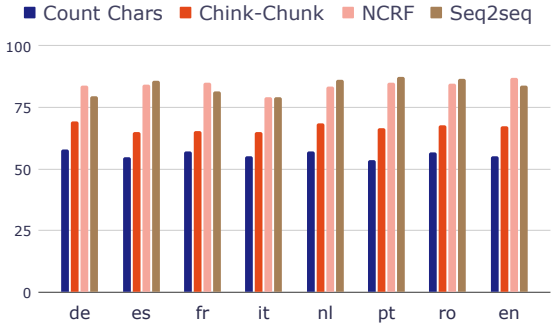
(a) Precision



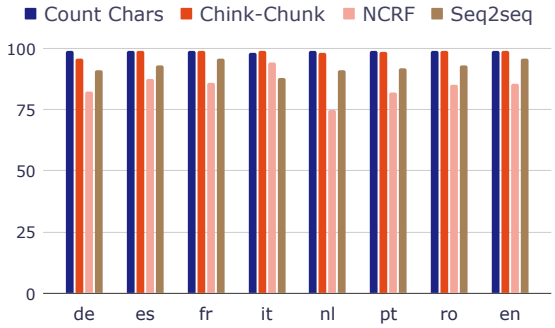
(b) Recall



(c) F1 score



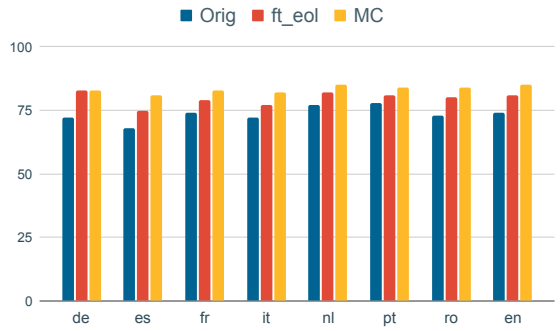
(d) BLEU$_{br}$score



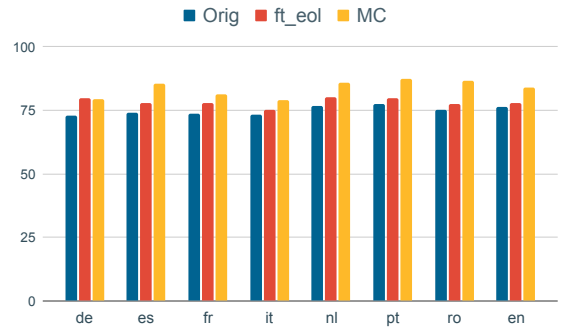
(e) Length conformity CPL≤42

Figure 4.7: Accuracy of different segmentation methods based on Precision, Recall and F1, BLEU$_{br}$and length conformity (CPL).

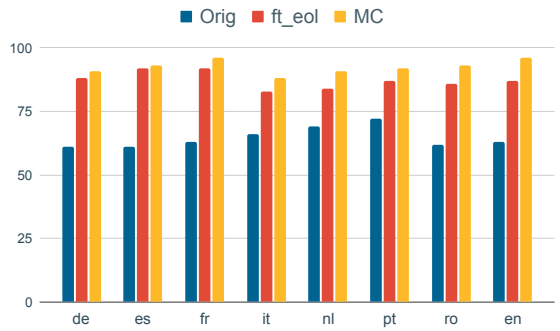
points on average. It is interesting noticing the trade-off between precision and recall (Figures 4.7a and 4.7b); while precision drops by 7 points on average when fine-tuning compared to the non-fine-tuned model, recall increases by 13 points. In fact, for every fine-tuning epoch, precision drops



(a) F1 score



(b) BLEU_{br}score



(c) Length conformity CPL≤42

Figure 4.8: Accuracy for different data mixtures of MuST-Cinema based on F1, $\text{BLEU}_{br}$and length conformity (CPL).

by about 1 point compared to the previous epoch while recall increases at a double rate. This shows that fine-tuning helps in segmenting the sentence into smaller pieces by inserting more breaks, even though small differences in positioning them are highly penalised in terms of precision. The model trained on MuST-Cinema after annotating it with the missing <eol> (MC) achieves the highest accuracy for all languages, showing that the automatic annotation of <eol> not only does not harm performance, but allows for synthesising subtitling data which lead to improved segmentation accuracy.

The picture is similar for BLEU_{br}, shown in Figure 4.8b. with the model fine-tuned on <eol> achieving an improvement over *All* (up to 7 points for

1.	La forma del bottone non è variata molto <eol> da quella che era nel Medioevo. <eob> *The design of a button hasn't changed much since the Middle Ages.*
2.	Sie lief aus dem Haus des Mannes fort, <eol> den sie nicht heiraten wollte, <eob> und heiratete schließlich <eol> den Mann ihrer Wahl. <eob> *She ran away from the house of the man she did not want to marry and* *ended up marrying the man of her choice.*
3.	C'est donc toujours plus difficile. <eob> *So it gets to be more and more of a challenge.*
4.	Je m'enroule en une petite <eol>boule comme un foetus. <eob> *And I curl up into a little fetal ball.*

Table 4.6: Examples of sentences output by the segmenter. Text in blue indicates insertions by the segmenter.

de). Still, *MC* performs better for all languages and on par for de, reaching scores above 85 for es, pt, nl and ro. BLEU_{br}, although useful to indicate segmentation differences of both break types compared to the reference, is not very informative regarding the type of error (e.g. undersegmentation-oversegmentation), as we will further show in Section 6.2.

Figure 4.8c shows the percentage of subtitles $\leq$42 characters long. The model trained on all data (*All*) does not insert sufficient breaks to reduce the length of the subtitles inside a sentence. Here the fine-tuned model achieves on average a 21% improvement, reaching up to a 92% length conformity for French and Spanish. In spite of the limited amount of data, fine-tuning leads to large improvements. Again, the model trained on all MuST-Cinema data (MC) achieves the highest conformity, showing that the corpus is a valuable resource towards achieving outputs conforming to the subtitle format.

Table 4.6 contains some examples of the segmenter output. Sentence 1 shows that the segmenter is able to insert breaks in the correct place. In Sentence 2, the segmenter inserts both <eol> and <eob> symbols, correctly alternating between the two. Sentence 3 is a sentence shorter than 42

characters, which was not changed by the segmenter. This shows that the segmenter does not unnecessarily segment subtitles unless needed. Lastly, Sentence 4 is already annotated with breaks. In this case, the segmenter does not insert any breaks because the sentence is already segmented.

These examples together with the high accuracy and conformity scores obtained by all metrics for the Seq2seq *MC* segmenter confirm the high quality of MuST-Cinema and its suitability for subtitling tasks. Despite the automatic annotation of <eol> symbols, segmenters trained on MuST-Cinema successfully segment long sentences into readable subtitles.

4.3.2 Heterogeneous data for subtitle segmentation

In the first experiment (Sec. 4.3.1) we demonstrated that segmenters trained on MuST-Cinema achieve high segmentation accuracy. However, the size and language coverage of MuST-Cinema is small compared to other subtitling corpora, such as OpenSubtitles whose size is at least ten-fold. Since size matters for training resource-hungry neural systems and data with human segmentation decisions are hard to obtain, one may wonder whether using the readily available OpenSubtitles data can lead to higher segmentation accuracy. In this experiment, we leverage available subtitling resources in different resource conditions to train models which automatically segment sentences into readable subtitles. The goal is to exploit the advantages of the available resources, i.e. size for OpenSubtitles and quality for MuST-Cinema, for maximising segmentation performance, but also taking into account training efficiency and cost.

Data and settings As training data we use the English portions of MuST-Cinema (from the En-Fr split) and OpenSubtitles. The monolingual files for OpenSubtitles come in XML format, where each subtitle block forming a sentence is wrapped in XML tags. We are therefore able to insert the

<eob> symbols for determining the end of a subtitle block. However, as mentioned in Section 4.1, line breaks are not present in OpenSubtitles. We hence proceed to creating artificial annotations for <eol>. We filter all sentences for which all subtitles have a maximum length of 42 characters (*OpenSubs-42*). Then, for each <eob>, we substitute it with <eol> with a probability of 0.25, making sure to avoid having two consecutive <eol>, as this would lead to an invalid display of three lines. Since this length constraint results in filtering out a lot of data, we also relax the length constraint by allowing sentences with subtitles with up to 48 characters (*OpenSubs-48*). The motivation for this relaxation is that, if a sequence-to-sequence model is not able to learn the constraint of length from the data but instead learns segmentation decisions based on patterns of neighbouring words, having more data will increase the amount and variety of segmentation decisions observed by the model. This may result in more plausible segmentation, possibly though to the expense of length conformity. We compare the models on OpenSubtitles with the segmenter trained on all MuST-Cinema data (*MC*) from Section 4.3.1 and on a segmenter trained on only 100k sentences for testing in low-resource conditions (*MC-100*).[9] Dataset sizes are reported in Table 4.7.

We are interested in the real application scenario where high-quality data containing human segmentation decisions are not available or scarce. According to our hypothesis, a relatively limited size of high-quality data can be compensated by OpenSubtitles. Therefore, we fine-tune each of the OpenSubtitle models on 10k and 100k English sentences from MuST-Cinema, which contain high-quality break annotations.

OpenSubtitles and TED Talks have been shown to have large differences and to constitute a sub-classification of the subtitling genre (Müller and Volk, 2013). For this reason, we experiment with 2 test sets for cross-

[9]Training a model with 10k data did not bring any good results.

Data	Sentences
MuST-Cinema	275,085
OpenSubs-42	185,758
OpenSubs-48	13,713,708

Table 4.7: Dataset sizes in sentences for English.

domain evaluation. The first set is the English test set released with MuST-Cinema, containing 9 single-speaker TED Talks (545 sentences). The second test set (*Industry*) is much more diverse and contains 782 sentences. To create it, we selected a mix of public and proprietary data, more specifically, excerpts from a TV series, a documentary, two short interviews and one advertising video. The subtitling was performed by professional translators and the .srt files were processed to insert the break symbols in the positions where subtitle and line breaks occur.

The segmenter model is the same model used in Section 4.3.1 and we followed the same subword pre-processing. Models were trained until convergence on 1 Nvidia GeForce GTX1080Ti GPU.

	MuST-Cinema					Industry					
Model	$BLEU_{br}$	Prec	Rec	F1	CPL	$BLEU_{br}$	Prec	Rec	F1	CPL	Time
Count-Chars	63.73	50	47	48	100	51.45	46	43	44	100	-
MC-all	**84.00**	85	**85**	**85**	**96**	**66.38**	72	**64**	**69**	**97**	305
MC-100	81.77	84	83	83	94	65.38	76	**64**	68	96	**210**
OpenSubs-42	72.24	86	66	73	74	61.41	**84**	56	65	79	270
+MC-10	77.99	83	76	79	88	63.53	76	60	66	93	+26
+MC-100	80.09	**87**	78	81	88	65.30	77	62	67	94	+250
OpenSubs-48	76.00	77	67	68	72	63.37	63	56	59	81	6980
+MC-10	82.46	86	80	82	91	65.66	78	61	67	94	+24
+MC-100	83.34	**87**	82	84	91	65.96	79	61	**69**	94	+260

Table 4.8: Segmentation results on the two test sets. Training time in minutes. Best score in bold.

Results Table 4.8 shows the results for the MuST-Cinema and the Industry test set respectively. As in the first experiment, the simple Count-Chars baseline achieves almost perfect conformity to the length constraint, it is however not accurate in inserting the breaks at the right positions, as shown by the very low BLEU_{br} (63.73 and 51.45) and F1 scores (48 and 44). The best performance for all metrics and both test sets is achieved when using all available MuST-Cinema data (*MC-all*). For this in-domain test set, BLEU_{br} and F1 are higher than for the Industry out-of-domain test set, however the number of subtitles conforming to the length constraint is consistently high (96% and 97%). This suggests that the systems trained on high-quality segmentation are able to produce readable subtitles in terms of length in diverse testing conditions even without massive amounts of data. Even with 100k of training data (*MC-100*) the performance of the model, which is the fastest model to train (210 minutes), drops only slightly, with -2% for all metrics on the MuST-Cinema test set and -1% on the second test set. This shows that high efficiency can be achieved without dramatically sacrificing quality. This is particularly important for industry applications where tens of languages are involved and training data for a domain might not be abundant.

The models trained only on OpenSubtitles show a large drop in performance for the MuST-Cinema test, which is to be expected because of the different nature of the data. However, the drop is present also for the second test set, which shows that these models are not robust to different domains. Surprisingly, the larger model (*OpenSubs-48*) does not perform much better than the model with less data (*OpenSubs-42*) even though it is trained on almost 10 times as much data. This could be an indication of a trade-off between data quality and data size. *OpenSubs-48* with more noisy data has similar recall to *OpenSubs-42*, but it is much less accurate in the position of the breaks, as shown by the drop in precision (86 vs. 77

and 84 vs. 63). We conjecture that the procedure of artificially inserting <eol> symbols by changing the existing <eob> does not reflect the distribution of the type of breaks in real data. Interestingly, the *OpenSubs-42* model, despite containing only subtitles of a maximum length of 42, is not able to generate subtitles which respect the length constraint (74% and 79%). It is therefore possible that the segmenter does not learn to take into consideration the constraint of length, but the segmentation decisions are based on lexical patterns in the data, as we will also show in the case of ST for subtitling in Chapter 5.

Fine-tuning, even on a minimum amount of real data, e.g. 10k of MuST-Cinema, can significantly boost the performance compared to OpenSubtitles and is a viable and fast solution towards readable subtitles. This corroborates the claim in favour of creating datasets which are representative of the task at hand. Surprisingly though, fine-tuning the *OpenSubs-42* model on MC-100 does not improve over training the model from scratch on MC-100 for neither test set. For the case when only a small amount of MuST-Cinema data is available (MC-10), having a larger base model on which to fine-tune (*OpenSubs-48*) is beneficial, as shown by an improvement for all metrics and in both testing conditions compared to all other models trained on OpenSubtitles or fine-tuned on them. Therefore, we conclude that, in the presence of little data containing human segmentation decisions, a model trained on more data, even though possibly noisier, is a more robust base on which to fine-tune using the high-quality data. One considerable drawback is that the improvement comes at a training time of x25 over the other base model (*OpenSubs-42*), which raises significant considerations for cost and efficiency. Such a model however, once trained, could be re-used for fine-tuning on several domains and for different client specifications.

We further perform a manual inspection to identify particular issues. A

low precision is connected to over-splitting or splitting in wrong positions, while low recall suggests under-splitting (not inserting a sufficient number of breaks). Indeed, we observe that the OpenSubtitle models tend to over-segment short sentences, but under-segment longer ones:

Reference: Let's turn our attention to the hows. <eob>(37 characters)
OpenSubs-42: Let's turn our attention <eol>to the hows. <eob>(25 + 12 characters)

Reference:
My family's traditions <eol>
and expectations for a woman <eob>
wouldn't allow me to own a mobile <eol>
phone until I was married. <eob>
(22 + 28 + 39 + 20 characters)
OpenSubs-42:
My family's traditions and expectations <eol>
for a woman wouldn't allow me to own a mobile phone until I was married. <eob>
(39+72 characters)

In the following example, fine-tuning on MC increases length conformity, splitting the first subtitle in two, while MC-100k succeeds in segmenting all subtitles exceeding 42 characters, matching the reference segmentation.

Reference:
Meditation is a technique <eol>
of finding well-being <eob>
in the present moment <eol>
before anything happens. <eob>
OpenSubs-42:
Meditation is a technique of finding well-being <eob>
in the present moment before anything happens. <eob>
(47+46 characters)
OpenSubs-42 + MC 10K:
Meditation is a technique <eol>

of finding well-being <eob>
in the present moment before anything happens. <eob>
(25+21+46 characters)
MC-100K:
Meditation is a technique <eol>
of finding well-being <eob>
in the present moment <eol>
before anything happens. <eob>

The examples above confirm our results which showed that the models do not explicitly learn the constraint of length, but rather patterns of segmentation. From a syntactic point of view, the break symbols are inserted after a noun (e.g. attention, expectations) and before a preposition/conjunction (to, for, in, before), regardless of the model. The break symbols, even though do not overlap with the human segmentation decisions, are inserted at plausible positions. This leads to subtitles that present logical completion, *i.e.* each subtitle is formed by a phrase or syntactic unit, even though they do not respect the constraint of length. The conformity to the length constraint seems to be forced only with the high-quality MuST-Cinema data. It is possible that the artificial break symbols in OpenSubtitles clash with the real break symbols in MuST-Cinema, which creates confusion for the model. Replacing some <eob> symbols with <eol> in OpenSubtitles to simulate data where human-annotated line breaks exist means that the models trained on OpenSubtitles observe a line break at positions where normally a subtitle break is present. Given the different functions of the two types of breaks, this is a possible explanation why fine-tuning OpenSubtitle models on MC-100 performs worse than training on MC-100 from scratch and provides us with insights on future design of artificial segmentation decisions to augment subtitling data.

In this experiment, we have presented methods to combine heteroge-

neous subtitling data in order to improve automatic segmentation of subtitles. We leveraged large data containing noisy segmentation decisions from OpenSubtitles and combined them with smaller amounts of high-quality data from MuST-Cinema to generate readable subtitles from full sentences. We found that even limited data with reliable segmentation can improve performance. We conclude that quality matters more than size for determining the break points between subtitles.

4.3.3 Multimodal and multilingual segmentation

In this experiment, we aim to improve both segmentation performance and efficiency. As far as performance is concerned, the segmenters presented in the previous sections were only text-based and did not consider audio information. Since speech phenomena, such as pauses and silences, can strongly influence the structure of the subtitles (Carroll and Ivarsson, 1998), we expect that information from the speech modality could improve segmentation performance. To explore this hypothesis, we extend the textual segmenter of the previous experiments with a multimodal architecture (Sulubacak et al., 2020), which receives audio and text input.[10] As for efficiency, we experiment with multilingual training to create one model able to segment multiple languages, drawing from the fact that subtitling constraints are similar among languages. Multilingual training (Johnson et al., 2017; Aharoni et al., 2019; Zhang et al., 2020a) allows for exploiting the scarce human segmentation decisions from multiple languages with one segmenter model. Specifically, our segmenter exploits, for the first time in this scenario, the source language audio (En) and segmented target text already available in a few languages (here: De, En, Fr, It). Its key strength

[10]Images and videos with subtitling material are often protected by copyright and thus not publicly available. Improving the segmenter with data from the visual modality is thus dependent on the availability of such resources.

is the ability to segment not only target languages for which high-quality segmented data is available but also unseen languages having some degree of similarity with those covered by the original ST resource(s), in a zero-shot fashion. This opens up the possibility to automatically obtain synthetic automatic subtitling training data for previously not available languages.

Segmenter To improve segmentation quality, we extend the textual segmenter of the previous sections in two ways. Our first extension is multimodal training, motivated by the multimodal nature of subtitling. The **multimodal segmenter** is built using an architecture with two encoders: one for the text (with the same structure as the textual segmenter) and one for the audio. We combine the encoder states obtained by the two encoders using parallel cross-attention (Bawden et al., 2018), as it proved to be effective both in speech and machine translation (Kim et al., 2019; Gaido et al., 2020a). Parallel attention (Figure 4.9) is computed by attending at the same intermediate representation (the decoder self-attention); then, the audio encoder cross-attention and the text encoder cross-attention are summed together and fed to the feed-forward layer.

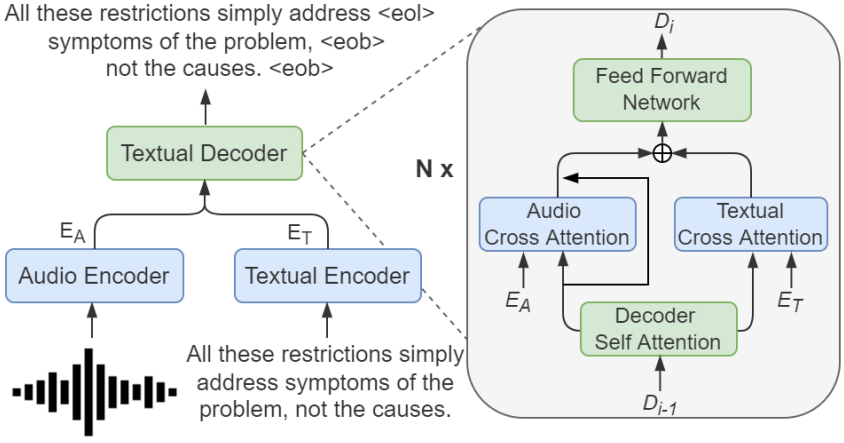


Figure 4.9: Parallel multimodal segmenter architecture.

Since subtitling constraints are the same across several languages, our second extension is to learn segmentation multilingually. To this aim, we follow the standard approache used in MT and ST; we combine samples from multiple languages in the same training step and add a prefix language token to the target text. As in MT, multilingual training in ST has been shown to enhance performance (Wang et al., 2020a) while maintaining only one model for multiple languages.

Data and evaluation We use three sections of MuST-Cinema, French, German, and Italian. For English, we concatenate the source captions of the French, German and Italian sections (757K sentences). To test the segmenters in zero-shot conditions we select two target languages: Dutch (an SOV – Subject-Verb-Object – language) and Spanish (SVO).[11]

For evaluating the segmentation, in the previous experiments we used Precision, Recall and F1. Using these metrics was possible because the segmenters did not introduce significant changes to the textual content. However, in this experiment, we noticed that the output was often significantly changed in zero-shot settings. For this reason, except for BLEU_{br}, we compute BLEU between the reference and the output text after removing the break symbols (BLEU_{nb}). BLEU_{nb} ensures that no changes are performed to the actual text, since the segmentation task consists in only inserting the break symbols, without changing the textual content. Higher values indicate less changes between the original text and the text where the break symbols were inserted. In addition, except for CPL conformity, to investigate when and if the systems over- or under-generate subtitle breaks, we report *Break coverage* computed as follows:

$$Coverage(\%) = \left(\frac{\#\texttt{<break>}_{pred}}{\#\texttt{<break>}_{ref}} \cdot 100 \right) - 100 \tag{4.4}$$

[11]Es and Nl data are only used for testing purposes so as to simulate the zero-shot conditions.

where <break> corresponds to either <eol>or <eob>. EOL and EOB coverage obtains negative values when the segmenter inserts less breaks than the reference or positive values when it inserts more.

Baselines We first evaluate the multilingual and multimodal segmenters against the textual monolingual segmenter of the first two experiments (Sec. 4.3.1 and 4.3.2) on *seen* languages. For evaluating the performance of our proposed segmenter on *unseen* languages, except for the Chink-Chunk baseline, we additionally compare its performance with two supervised Seq2Seq segmenters, trained on the data for the respective language (Es, Nl). The first one is the Seq2seq textual segmenter trained on MuST-Cinema (*MC*), which provides the best possible performance that could be achieved if in-domain data for the respective language was available; the second is the highest-performing segmenter trained on OpenSubtitles data in the second experiment (*OpenSubs-42*), representing the case when only out-of-domain data is available for the respective language. After filtering the OpenSubtitles data, we noted that almost 90% of the sentences obtained contain only one subtitle. This is not very informative for the segmenter, since the only operation learned is inserting one <eob>at the end of the sentence. For this reason, we select only sentences with at least two subtitles (or two subtitle lines). This results in 2,956,207 sentences for Es and 683,382 sentences for Nl. We then add the same number of sentences containing only one subtitle, finally obtaining 5,912,414 sentences for Es and 1,366,764 sentences for Nl.

Models The *Textual* segmenter is a Transformer-based (Vaswani et al., 2017) architecture consisting of 3 encoder layers and 3 decoder layers. We set the hyper-parameters as in the fairseq (Ott et al., 2019) multilingual translation task, both for the mono- and multilingual textual segmenters.

For the multilingual model, a mini-batch for each language direction (here: 4) is built and the model weights are updated after each mini-batch.

The *Multimodal* segmenter is an extension of the textual segmenter encoder-decoder structure with an additional speech encoder composed of 12 Transformer encoder layers as in the original speech-to-text task (Wang et al., 2020a) but with the addition of a CTC (Graves et al., 2006) module to avoid the speech encoder pre-training (Gaido et al., 2021a). The training of the multilingual models is performed by pre-pending the language token (`En`, `De`, `Fr`, `It`) to the target sentence (Inaguma et al., 2019). The encoder and decoder embeddings are shared. We select the hyper-parameters of the original implementation,[12] except for a higher learning rate of $1 \cdot 10^{-3}$, since pre-training was skipped. The vocabulary is generated using SentencePiece (Kudo and Richardson, 2018), setting the size to 10k unigrams both for the mono- and multilingual segmenters.

Segmenter	Training	English		French		German		Italian	
		$BLEU_{br}$	CPL	$BLEU_{br}$	CPL	$BLEU_{br}$	CPL	$BLEU_{br}$	CPL
Textual	mono	82.6	**96**	81.4	**96**	79.5	91	79.2	90
	multi	81.8	88	81.8	94	80.0	91	78.5	**91**
Multimodal	mono	86.6	95	85.5	94	84.4	**91**	82.1	90
	multi	**87.8**	95	**87.1**	94	**85.4**	90	**84.2**	90

Table 4.9: Segmentation results on seen languages. Best results in bold.

Results The results for **segmentation accuracy on *seen* languages** are shown in Table 4.9. In this scenario, $BLEU_{nb}$ always approaches 100 i.e., the system learnt to perfectly copy the input text. Looking at the $BLEU_{br}$ values, the *Multimodal* segmenter outperforms the *Textual* one by at least 5.0 $BLEU_{br}$ points on all languages, confirming the benefits

[12] `https://github.com/pytorch/fairseq/blob/main/examples/speech_to_text/docs/mustc_example.md`

of multimodal training. In contrast, overall subtitle conformity is higher for the *Textual* segmenter in 3 out of 4 languages, where its CPL scores are 1.2-2.6 percentage points above those obtained by the *Multimodal*. As for mono-/multilingual training, multilingual training clearly benefits the *Multimodal* segmenter but the results are mixed for the *Textual*, with minimum differences. In addition, except for one case (Italian), higher CPL values are obtained under the monolingual training regime. However, preliminary experiments revealed a worse transfer learning capability of the monolingual *Textual* segmenter on other languages compared to its multilingual version. The results in Table 4.9 also show that our multilingual *Multimodal* segmenter has the highest segmentation quality and conformity. Moreover, it improves over our textual segmenter in Section 4.3.1 by $\sim$4 BLEU_{br} points for English, keeping CPL at the same value.

After testing the segmenters on seen languages, we evaluate their **segmentation performance on *unseen* languages** (Dutch, Spanish), under zero-shot conditions. The results in Table 4.10 show that the zero-shot *Textual* segmenter achieves the lowest segmentation quality, as shown by a BLEU_{br} score of 52.6 for Nl and 54.5 for Es. In this respect, the *Chink-Chunk* and the supervised out-of-domain *OpenSubs-42* baselines perform slightly better, but this difference comes from less changes to the actual text, as demonstrated by the almost perfect BLEU_{nb} scores. Moreover, both *Textual* and *OpenSubs-42* generate subtitles conforming to the CPL constraint in only 70% of the cases, despite having received only length-conforming subtitles during training. Despite being trained on subtitling data for the particular language, the low performance of *OpenSubs-42* could be attributed to the different domain between OpenSubtitles and the MuST-Cinema test set, as it was shown in Section 4.3.2. For example, MuST-Cinema mainly contains long sentences with multiple breaks, while in OpenSubtitles we rarely come across sentences with more than three

breaks. The negative values of EOB and especially EOL coverage indicate that textual methods largely under-generate subtitle breaks. From these results we can conclude that zero-shot segmentation does not perform satisfactorily with textual input only.

Dutch					
Segmenter	$BLEU_{br}$	$BLEU_{nb}$	CPL	EOL	EOB
Chink-Chunk	68.7	**100**	**98.1%**	-29.9%	+20.2%
OpenSubs-42	60.4	99.5	71.2%	-31.4%	-51.3%
MC	**86.1**	99.6	92.1%	**-15.1%**	-1.9%
Textual	52.6	61.3	77.8%	-23.4%	-9.9%
Multimodal	80.1	**99.9**	91.4%	-27.2%	**+0.4%**
Spanish					
Segmenter	$BLEU_{br}$	$BLEU_{nb}$	CPL	EOL	EOB
Chink-Chunk	64.9	**100**	**99.2%**	-30.3%	+41.1%
OpenSubs-42	62.1	96.6	71.2%	-32.3%	-45.4%
MC	**85.6**	99.6	93.4%	**-15.8%**	+4.8%
Textual	54.5	69.6	70.1%	-47.6%	-19.3%
Multimodal	80.6	99.6	91.8%	-22.4%	**+4.7%**

Table 4.10: Segmentation results on *unseen* languages.

In comparison, the *Multimodal* segmenter performs significantly better. It reaches an absolute gain of 27.5 $BLEU_{br}$ for Nl and 26.1 $BLEU_{br}$ for Es compared to *Textual*, confirming the results on seen languages. It shows a ∼20 point increase compared to the supervised *OpenSubs-42* baseline, while it is only 5-6 $BLEU_{br}$ points behind the strong in-domain supervised segmenter (*MC*). Moreover, contrary to *Textual* and *OpenSubs-42*, the *Multimodal* model has learnt to perfectly copy the text, as shown by the high $BLEU_{nb}$ scores (up to 99.9 on Nl), close to the maximum score of a method – *Chink Chuck* – that by design does not change the original text, and even above the supervised in-domain system (MC). The CPL results are in agreement with $BLEU_{br}$; for both languages, the *Multimodal*

model respects the length constraint in more than 91% of the subtitles, with a minimum drop compared to MC. Strikingly, even if the two target languages were never seen by the model, the CPL results are similar to those obtained on seen languages (see Table 4.9). Despite under-generating <eol> symbols as all other systems, the *Multimodal* has the best scores for EOB coverage.

Our results indicate the higher effectiveness of *Multimodal* segmentation to automatically convert sentences into subtitles. In addition, at least for the Western European languages considered in our experiments, our approach can be successfully applied in zero-shot settings involving languages not present in the training data used to build the segmenter. While probably unrealistic (and hard to verify due to the lack of suitable benchmarks), the possibility of porting our approach to scenarios involving different alphabets was not verified in this work. This would require, at least, a vocabulary adaptation which represents a well-known problem in multilingual approaches to MT/ST (Lakew et al., 2018; Garcia et al., 2021). Nevertheless, even in the worst case in which some degree of similarity across languages is required for zero-shot automatic segmentation, our results propose a method for achieving enhanced segmentation accuracy for multiple languages, with the efficiency of maintaining only one model.

4.4 Conclusions

This chapter is dedicated to subtitling data as the first component towards addressing automatic subtitling from a computational perspective. Automatic subtitling comprises, besides translation, automatic spotting and automatic segmentation. For this reason, we argued that representative subtitling data should contain the audio as the source and timed text as the target. Based on these criteria, in Section 4.1 we performed

an analysis of existing subtitling corpora and their suitability for training automatic subtitling systems. We explored whether audio or timing information is available and the degree of conformity to the subtitling constraints of length, reading speed and preservation of linguistic wholes. Our analysis showed that, even though subtitling corpora conform to some degree to the constraints, several limitations exist, mainly related to the lack of audio and time information for modelling spotting and the loss of line breaks in the target, which are indispensable for proper segmentation.

To address this challenge, in Section 4.2 we discussed the creation of MuST-Cinema, the first speech subtitling corpus, in which subtitle breaks from .srt files are projected onto the MuST-C ST corpus. As with other corpora, the line breaks inside the .srt files were collapsed, which posed a challenge to creating a conforming corpus. However, based on a minimum sample of sentences containing line breaks, we proposed a segmenter which annotates the subtitles with line breaks. We showed that an iterative self-learning process of training and annotation can successfully insert the missing <eol>symbols in the corpus. An analysis of the subtitling properties of MuST-Cinema showed high conformity to all subtitling constraints for all 7 target languages. When exploring the reasons for the remaining non-conformity, we found linguistic features in some earlier TED talks suggesting a transition from a plain translation of the transcript to proper subtitling, with more condensation and omissions. Despite this, MuST-Cinema addresses the limitations of other subtitling corpora by offering audio input and high-quality timed-text on the target.

In the last Section 4.3, we explored the approach used to annotate MuST-Cinema with the missing <eol> symbols to the task of automatic subtitling segmentation. Our first experiment in this respect showed that a segmenter model trained on the annotated MuST-Cinema obtained higher segmentation accuracy compared to one trained and fine-tuned only on

the original human segmentations in MuST-Cinema, confirming the suitability of MuST-Cinema for subtitling tasks. In the second experiment, we explored the question whether data quality or size matters most for automatic subtitle segmentation. Since high-quality human segmentations are scarce and limited in the domains available, we took advantage of the sizeable OpenSubtitles corpus to propose data mixes for achieving high segmentation accuracy under different resource conditions. Our results demonstrated that even a small portion of high-quality data can lead to high segmentation performance both in-domain and out-domain. In the last experiment we improved both performance and efficiency in segmentation, by extending the textual segmenter to a multimodal, multilingual segmenter. This segmenter is trained on the source audio and text and multiple target languages. Our results showed an increase in segmentation performance of 6 BLEU_{br} points compared to the textual segmenter, with the added benefit of training and maintaining only one model for multiple languages. Additionally, the multimodal segmenter has the ability to accurately segment even unseen languages in a zero-shot fashion. This opens up the possibility to annotate existing translation corpora with subtitle breaks, in order to increase the availability of training data for automatic subtitling systems. Having procured the necessary training data and techniques for subtitling data augmentation, in the next chapter we delve into automatic models for different subtitling scenarios.

Chapter 5

End-to-end Models for Subtitling

In the previous chapter, we set the ground for transitioning to the paradigm of automatic subtitling by compiling the necessary speech subtitling corpora to train the systems. Having the first ingredient in hand, we move on to developing automatic subtitling systems for different subtitling scenarios. In Section 5.1 I argue why simply matching the length of the source, as proposed in previous approaches to MT for subtitling, is not sufficient to produce readable subtitles under the new paradigm. Having set the motivation for a more comprehensive approach to automatic subtitling, I present our automatic subtitling solutions for offline subtitling, dual subtitling and live subtitling. For the scenario of offline subtitling (Section 5.2), we explore the feasibility of direct ST models for automatically generating subtitles. We compare our proposed novel approach to NMT from a human template and cascade ST (ASR+MT). In Section 5.3, we propose direct models for dual subtitling, addressing the need to display source language captions and target language subtitles at the same time. Since the joint display poses the additional challenge that the source and target version need to be synchronised with each other, we introduce metrics for evaluating consistency between captions and subtitles. Lastly, in Section 5.4, we experiment, for the first time, with direct simultaneous ST models for live

subtitling. The fact that our systems predict subtitle boundaries allows us to explore different display modes: showing the translation word-for-word as it becomes available, or waiting until a segmentation boundary is predicted. As suggested by the literature and normative recommendations, a display in scrolling lines is a solution that generates the most readable subtitles. We investigate whether alternative modes for displaying the output of simultaneous ST systems can still keep the delay at acceptable levels.

5.1 Is 42 the answer to everything in automatising subtitling?

Subtitles, despite being the fastest and most widespread way of translating audiovisual content, still rely heavily on human effort. As described in section 2.2.4, one way to produce subtitles is when human subtitlers translate directly the source speech of the video using specialised software and making sure the text is properly synchronised with the speech and segmented in a way that facilitates readability and legibility. However, computational approaches to subtitling have rarely started from speech.

Previous work in MT for subtitling has approached subtitling as text translation; the source language speech is transcribed and spotted by a human, who creates a subtitle template (see Section 2.2.4), and then passed through an MT system to be translated, either one subtitle at a time, or with several subtitles merged into longer pieces of text. These source language subtitles (captions) are already compressed and segmented to respect the subtitling constraints of length, reading speed and proper segmentation. In this way, the work of the MT system is already simplified, since it only needs to translate matching the length of the source text with length-matching techniques (Matusov et al., 2019; Lakew et al., 2019; Niehues, 2020; Liu et al., 2020a; Cherry et al., 2021). However, the essence

of a good subtitle goes beyond matching a predetermined length (as, for instance, 42 CPL for TED talks). Apart from this spatial dimension, subtitling relies heavily on the temporal dimension, which is incorporated in templates as timestamps. This approach to subtitling is partial and limited in its efficiency, since it automatises only the step of translation and thus requires considerable human intervention, e.g. for transcription and spotting. Templates, besides often preventing optimal solutions for the target language, are costly and slow to create and as so, not a viable solution for short turn-around times and individual content creators. Therefore, skipping the template creation process would lead to cost and time reductions and make multilingual subtitling more accessible to all types of creators.

Because of the multifarious nature of subtitling, models need to generate subtitles adhering to different requirements and characteristics depending on the application scenario. For example, the source language captions often need to be displayed along target subtitles to aid language learners, the deaf and hard-of-hearing, or for other tasks in the localisation industry (indexing, quality assurance, etc). Consistency between captions and subtitles is a desideratum in this case, so that both versions are synchronised in a bilingual display and that possible errors can be spotted across the same blocks to facilitate quality assurance (Nikolić, 2015). Creating high-quality multiple language/accessibility subtitle versions with one model is another step towards maximising automation and efficiency in subtitling. For live subtitling, generating high quality subtitles should be accompanied with generating them quickly enough to be displayed with a minimum delay compared to the speaker, while the way live subtitles are displayed also impacts viewing experience. Taking all these aspects into consideration, 42 may not be the answer to everything in automatic subtitling.[1]

[1]Speaking of subtitles and their optimum length of 42 characters per line, we could not help but alluding to the book and series *The Hitchhiker's Guide to the Galaxy* by Douglas Adams, where the number 42 is the "Answer to the Ultimate Question of Life, the Universe, and Everything", calculated

5.2 Offline automatic subtitling

In this section, we propose Speech Translation as an alternative to the template creation process. For each audio segment, the task consists in generating its target translation segmented with subtitle breaks.[2] We experiment with cascade systems, i.e. pipelined ASR+MT architectures, and direct, end-to-end ST systems. While the MT system in the pipelined approach receives a raw textual transcription as input, the direct speech translation receives temporal and prosodic information from the source language input signal. Given that several decisions about the form of subtitles depend on the audio (e.g. subtitle segmentation at natural pauses, length based on utterance duration), in this work we ask the question whether end-to-end ST systems can take advantage of this information to better model the subtitling constraints and subtitle segmentation. In other words, we investigate whether, in contrast to text translation (where one mainly focuses on returning subtitles of a maximum length of 42 characters), in speech translation additional information can help to develop a more advanced approach that goes beyond merely matching the source text length. Specifically, our contributions can be summarised as follows:

- We present the first end-to-end ST solution for subtitling, completely eliminating the need of a source language transcription and any extra segmentation component.

- We conduct a thorough analysis of cascade vs. end-to-end systems both in terms of translation quality and conformity to the subtitling constraints, showing that end-to-end systems have large potential for subtitling.

by a massive supercomputer named *Deep Thought* for over 7.5 million years.

[2]Our direct approach does not address timestamp prediction at this stage.

5.2.1 Experimental setting

Data For the experiments we use the MuST-Cinema corpus (Karakanta et al., 2020b). As described in Section 4.2, MuST-Cinema contains (*audio*, *transcription*, *translation*) triplets where the breaks between subtitles have been annotated with the special symbols <eol> and <eob>. For the audio we use the MuST-Cinema gold segmentation. We experiment with 2 language pairs, English→French and English→German, as languages with different syntax and word order.

MT and ST systems The **Cascade** system consists of an ASR and an MT component. The ASR component is a hybrid ASR system based on the KALDI toolkit (Povey et al., 2011). It features a time-delay neural network and lattice-free maximum mutual information discriminative sequence-training (Povey et al., 2016). The audio data for acoustic modelling include the clean portion of LibriSpeech (Panayotov et al., 2015) (∼460h) and a subset of the MuST-Cinema training set (∼450h), from which 40 MFCCs per time frame were extracted. A Maximum Entropy language model with n-gram features (Alumäe and Kurimo, 2010) is estimated from the corresponding transcripts (∼7M words). The MT component is based on the Transformer architecture (big) (Vaswani et al., 2017) with similar settings to the original paper. The system is first trained on the OPUS data (not containing subtitling boundaries), with 120M sentences for En→Fr and 50M for En→De, and then fine-tuned on MuST-Cinema. Considering that the ASR output is lower-cased and without punctuation, we lowercase and remove the punctuation from the source side of the parallel data used in pre-training the MT system. To mitigate the error propagation between the ASR and the MT, for fine-tuning, we use a version of MuST-Cinema where the source audio has been transcribed by the tuned ASR. The target side is text containing subtitle boundaries.

For the **End-to-End** system, we experiment with two data conditions: one where we only use the MuST-Cinema training data (**E2E-small**) and a second one where we pre-train on a larger amount of data and fine-tune on MuST-Cinema (**E2E**). This will allow us to detect whether there is any trade-off between translation quality and conformity to constraints when increasing the amount of pre-training data not representative of the target application (subtitling). The architecture used is S-Transformer, (Di Gangi et al., 2019b), an ST-oriented adaptation of Transformer, which has been shown to achieve high performance on different speech translation benchmarks. We remove the 2D self-attention layers and increase the size of the encoder to 11 layers, while for the decoder we use 4 layers.[3] We use the encoder of the ASR model to initialise the weights of the ST encoder and achieve faster convergence (Bansal et al., 2019). The encoder of the ASR model, instead, has 8 encoder and 6 decoder layers. The additional 3 encoder layers are initialised randomly, similarly to the adaptation layer proposed by Bahar et al. (2019a). We optimise using Adam (Kingma and Ba, 2015) (betas 0.9, 0.98), 4000 warm-up steps with initial learning rate of 0.0003, and learning rate decay with the inverse square root of the iteration number. We apply label smoothing of 0.1, and dropout (Srivastava et al., 2014) is set to 0.2. We further use SpecAugment (Park et al., 2019), a technique for online data augmentation, with augment rate of 0.5.

Since the E2E-small system, trained only on MuST-Cinema, is disadvantaged in terms of the amount of training data compared to the cascade, we utilise synthetic data to boost the performance of the ST system (E2E). To this aim, we automatically translate into German and into French the English transcriptions of the data available for the IWSLT2020 offline speech

[3]This choice was motivated by preliminary experiments, where we noted that replacing the 2D self-attention layers with normal self-attention layers and adding more layers in the encoder increased performance, while removing a few decoder layers did not negatively affect the performance.

translation task[4] (whenever the translation is not available in the respective target language). We use an MT Transformer model achieving 43.2 BLEU points on the WMT'14 test set (Ott et al., 2018) for EN→FR. Our EN→DE Transformer model using similar settings achieves 25.3 BLEU points on the WMT'14 test set, with state-of-the-art methods ranging between 27-30 BLEU. The resulting training data (both real and synthetic) amount to 1.5M sentences on the target side. We use a different pseudo-token to separate the real from the synthetic data. Finally, we fine-tune on MuST-Cinema.

For comparison to the new paradigm overcoming the need of a human transcription, we also report NMT results when starting from a subtitle template (**Template**). In this setting, we use the textual source side of MuST-Cinema, which contains the human-generated captions of the source language audio. In this way, the input to the MT system is already split in subtitles using the special symbols, respecting the subtitling constraints and with proper segmentation. This allows us to have an upper-bound of the performance that NMT can achieve when provided with input already in the form of subtitles. We pre-train large models with the OPUS data (Tiedemann, 2012) used in the cascade but without lowercasing or removing punctuation and then fine-tune them on the full training set of MuST-Cinema. It should be noted that only the MuST-Cinema data contain break symbols. We use the same Transformer architecture as in the cascade system. For all the experiments we use the fairseq toolkit (Gehring et al., 2017). Training is completed when the validation perplexity does not improve for 3 consecutive epochs. Byte-Pair Encoding (BPE) (Sennrich et al., 2016b) is set to 8K operations for the E2E-small and E2E systems while to 50K joint operations for the Cascade and the Template systems.

[4] `http://iwslt.org/doku.php?id=offline_speech_translation`

Evaluation To evaluate translation quality we use the SacreBLEU implementation (Post, 2018) to compute BLEU[5] on the MuST-Cinema test set. As with the segmenters in Section 4.3, we compute two variants of BLEU: with the break symbols (BLEU_{br}) and after removing them (BLEU_{nb}). BLEU_{br} evaluates both translation quality and segmentation, while BLEU_{nb} only the translation quality without taking into account the subtitle segmentation. In Section 6.2 we will present an analysis of these two metrics for quality and segmentation.

For evaluating the conformity to length, we calculate the percentage of subtitles $\leq$42 characters per line (CPL), while for reading speed the percentage of sentences with maximum 21 characters per second (CPS). Since our systems do not generate timestamps along with the subtitles, we compute CPS at the utterance level, by dividing the number of characters in the utterance by its duration. This approach has been applied in two of the largest subtitling research projects in the UK (Ofcom, 2014) and US (Jensema, 1997; Linebarger, 2001) to measure reading speed on full videos. Naturally, the computation of reading speed contains segments of the utterance which may not be accompanied by subtitles, but in our case, the utterances in the MuST-Cinema test set mostly contain continuous subtitles, so we expect this effect to be minimal.

Finally, for judging the goodness of the segmentation, i.e. the position of the breaks in the translation, we compute two metrics: the first one is *Sigma*, a new subtitle segmentation score. *Sigma* is the ratio of BLEU_{br} to the highest BLEU_{br} we could achieve if all subtitle boundaries were placed in correct positions. It produces a score between 0-100, where 100 stands for perfect segmentation. In this way, the metric is able to disentangle the effect of segmentation from that of translation quality (see Section 6.2

[5] `BLEU|#:1|c:mixed|e:no|tok:13a|s:exp|v:2.0.0`

		$BLEU_{br}$↑	$BLEU_{nb}$↑	CPL↑	CPS↑	TER_{br}↓	*Sigma* ↑
En→Fr	Template	32.15	29.50	91%	68%	19.37	89.22
	Cascade	26.34†	25.41†	91%	**72**%	**23.22**	**83.05**
	E2E-small	22.52	21.03	95%	70%	24.48	81.82
	E2E	**26.35**†	**25.46**†	**95**%	70%	23.51	81.46
En→De	Template	27.47	25.95	88%	56%	17.89	83.39
	Cascade	**22.44**†	**21.99**	87%	**56**%	22.36	78.47
	E2E-small	15.71	14.25	**93**%	55%	25.39	75.59
	E2E	22.00†	20.92	90%	**56**%	**21.25**	**79.24**

Table 5.1: Results for translation quality ($BLEU_{br}$, $BLEU_{nb}$), for conformity to the subtitling constraints (CPL, CPS) and for subtitle segmentation (TER_{br}) for the four systems. Differences in the results marked with † are not statistically significant. Best score among ST outputs in bold.

for more details on the *Sigma* metric). We additionally report TER_{br}[6], which is computed by masking all words except for the break symbols and calculating Translation Edit Rate (Snover et al., 2006) only for the breaks against the reference translation. This will allow us to determine the effort required by a human subtitler to manually correct the segmentation.

5.2.2 Results

The results are shown Table 5.1. As far as translation quality is concerned, the best performance is reached, as expected, in the *Template* setting, where the MT system is provided with 'perfect' source language captions. On the MuST-Cinema test set, this leads to $BLEU_{br}$ scores of 32.15 and 27.47 respectively for French and German. The *Cascade* setting follows with a $BLEU_{br}$ score reduction of 6 points for French and 5 points for German, largely because of error propagation from the ASR to the MT.

The *E2E-small* model (trained solely on MuST-Cinema) achieves 22.52

[6] `TER|#:1|c:lc|t:tercom|nr:no|pn:yes|as:no|v:2.0.0`

and 15.71 BLEU_{br} points for French and German respectively, which is a relatively low performance compared to the rest of the systems. This 'low-resource' setting is a didactic experiment aimed at exploring how far the data-hungry neural approach can go with the limited amount of data typically available in the domain of ST for subtitling (280K and 234K sentences). It should be noted that MuST-Cinema is the only ST corpus of the subtitle genre, respecting the subtitling constraints and containing break symbols. Consequently, the interference of other data may hurt the conformity to the subtitling constraints, despite improving the translation performance. On the other hand, pre-training has evolved into a standard procedure for coping with the data-demanding nature of NMT. Therefore, pre-training also in the case of end-to-end speech translation offers a comparable setting with the template and cascade experiments. Indeed, after fine-tuning the pre-trained model on MuST-Cinema, *E2E* reaches 26.35 and 22.00 BLEU_{br} points for French and German, obtaining almost the same scores as the *Cascade*. The Cascade seems to have an edge only for pure translation quality in the case of German, where it scores 1 BLEU_{nb} point above *E2E*. This shows that, when increasing the size of the training data for *E2E*, the gap between the cascade and the end-to-end approach is closed and that end-to-end approaches may have finally found a steady ground for flourishing in different applications.

Conformity to the subtitling constraints and subtitle segmentation When it comes to the conformity to the subtitling constraints, the results show a different picture (see CPS and CPL of Table 5.1). *E2E* generates subtitles having length of maximum 42 characters in 95% and 90% of cases. *E2E* achieves higher conformity even compared to *Template*, for which the segmentation is already provided to the system in the form of break symbols, while *Cascade* is behind by 3-4%. A similar tendency is observed with the

TER_{br}results measuring the proper placement of the break symbols. While *Template* benefits from the source language segmentation and therefore requires less edits to properly segment the subtitles, *Cascade* is more disadvantaged in guessing the correct position of the break symbols, as shown by a 23.22 and 22.36 TER_{br}score. For *E2E-small* $TER_{br}$is the highest, possibly due to the low translation quality. However, *E2E* achieves similar performance with *Cascade* for French and even outperforms the Cascade by 1 TER_{br}point for German, showing that less effort would be required to correct the segmentation. *Sigma* is in line with TER_{br}in ranking *Cascade* as having the best segmentation for French and *E2E* for German. However, for French the segmentation performance between the two E2E models is of similar quality, contrary to the ranking of $TER_{br}$and $BLEU_{br}$. As we will show in Section 6.2, both these metrics are to some degree affected by translation quality and cannot fully isolate the segmentation signal. For this reason, in the next sections we will only report $BLEU_{nb}$ as a metric of translation quality and *Sigma* for segmentation quality.

While the results on segmentation are not conclusive, the higher conformity and segmentation quality for *E2E* (at least for German) could be an indication that the direct system receives information compared to the cascade, which allows for better guessing the positions of the break symbols in the translation. This suggests that subtitle segmentation decisions are not solely determined by reaching a maximum length of 42 characters, but a combination of multiple (possibly intersemiotic) factors can offer a better answer to automatic subtitling.

Analysis The higher scores for CPL and TER_{br} for German in Table 5.1 suggest that the *E2E* system is better at modelling the subtitling constraints of length and proper segmentation. We shed more light into this aspect by analysing factors which might be determining the system's be-

haviour in relation to the insertion of the break symbols.

One question quickly arising is how the system can determine whether to insert a subtitle break symbol <eob>(which means that the next subtitle will follow on a new screen) or a line break symbol <eol>(which means that the next line of the subtitle will appear on the same screen). We have mentioned in Section 2.2.2 that the *in* and *out* times of a subtitle should follow the rhythm of speech. Therefore, we expect that the end of a subtitle block should correspond to the end of a speech act, a pause or a terminal juncture. On the other hand, while line breaks can still overlap with pauses or signal the change of speaker, their function is to split a long subtitle into two smaller parts in order to fit the screen. Consequently, the insertion of <eol> is determined more by the length and the syntactic properties of the subtitle and less by the natural rhythm of the speech. If this hypothesis holds, the choice of whether to insert an <eob> or an <eol> symbol is defined by prosodic properties and not solely by reaching the maximum length of 42 characters.

To test this hypothesis, we compute the duration of the pause after each word in the source side of the MuST-Cinema test set. To achieve this, we perform forced alignment of the transcript against the audio and subtract the end time of each word from the start time of the next word:

$$pause_{w1w2} = start_time_{w2} - end_time_{w1} \tag{5.1}$$

Then we separate the pauses in 3 groups: *i)* pauses corresponding to positions where <eob>is present, *ii)* pauses corresponding to positions where <eol>is present and *iii)* pauses after which there is no break symbol (*None*). In Table 5.2 we report average and standard deviation of the pause duration for each category.

Pauses corresponding to the positions where <eob> symbols are present are more than x10 longer than the pauses in positions without any break

Pause type	Avg	Stdev
None	0.039	0.022
<eob>	0.551	0.181
<eol>	0.074	0.027

Table 5.2: Average pause duration and standard deviation (in seconds) for the category without breaks (None), and for the categories with the two types of break symbols.

symbols (None). Even if we take the most extreme cases (based on standard deviation), any pause above 0.37 seconds would require the insertion of <eob>. Pauses corresponding to <eol> symbols are on average x2 longer, but there is an overlap between the possible durations of the *None* and the <eol> category. This confirms our hypothesis about the different roles of the two subtitle breaks. Therefore, prosodic information is an important factor which can help the ST system determine the subtitling segmentation according to the speech rhythm. This finding provides evidence towards a clear limitation of the cascade setting, where a raw textual transcription from the ASR does not provide any prosodic information to the MT system. The MT system in the cascade setting is disadvantaged by the inability to: *i)* recover from possible ASR errors, and *ii)* make decisions determined by factors other than text.

With this knowledge, we analyse the breaks in the results of the two systems. To control for differences in translation, we select all sentences with at least 2 breaks and with the same number of break symbols (regardless of whether <eob> or <eol>) between the reference, the output of the Cascade and of the E2E. The resulting sentences are 137 for French and 158 for German. We calculate the accuracy of the type of break symbols for the two systems. For French the accuracy is 89% for *Cascade* and 93% for *E2E*. For German the accuracy is 85% for *Cascade* and 88% for *E2E*. This difference in accuracy suggests that the direct system may be aided by the

acoustic information and specifically by the pause duration in determining the correct break symbol. Given the importance of acoustic information for subtitling, it would be fair to mention that the limitation of the cascade in the current setting could be possibly overcome with techniques encoding pause duration as additional information in the MT system.

EN	“Who do you report to?” <eob>“It depends”.
CAS	Wen melden Sie an? <eol>Es hängt davon ab.
E2E	Wen berichten Sie? <eob>Es kommt darauf an.
REF	“An wen schickst du deine Berichte?” <eob>“Das kommt darauf an”.
One executive at another company <eob>likes to explain how he used to be <eol>a master of milestone-tracking.	
Un cadre d’une autre entreprise aime expliquer <eob>comment il était jadis un maître <eol>du trek capital.	
Un dirigeant d’une autre entreprise <eob>aime expliquer comment il était <eol>un maître de traçage en pierre.	
Un cadre d’une autre entreprise <eob>aime raconter comme il était passé maître <eol>dans la surveillance des étapes.	
But you know how they say <eol>that information is a source of power?	
Vous savez comment dire que l’information <eol>est une source de pouvoir.	
Saviez-vous comment dire <eol>que l’information est une source de pouvoir ?	
Mais vous savez qu’on dit que <eol>l’information est source de pouvoir ?	

Table 5.3: Examples of translations by the cascade (CAS) and the end-to-end model (E2E) compared to the source sentence (EN) and the reference (REF). The sentence-final <eob>has been removed.

Table 5.3 presents some examples, evaluated also against the video. In the first example, the decision of which type of break to insert can only be determined by the duration of the pause that comes between them. Indeed, the speaker in the video asks the question and then leaves some time to the audience before giving the answer. The pause between the two sentences is about 2 seconds. In this case, the first sentence should be in one subtitle block, then disappear, and the second sentence should come in the next subtitle block in order not to reveal the answer before it was spoken by the speaker. This information is only available to the E2E system.

In the second example, although both systems output the right type of break, there are differences in the actual break positions. *E2E* inserts the first break symbol in the same position as in source and reference (*entreprise*), while *Cascade* inserts it at a later position (*expliquer*), resulting in a subtitle of 46 characters, which is above the 42-character length limit. *Cascade* correctly inserts the second break (*maître*), as in the reference. Here, *E2E* copies the break position from the source sentence, which is in a different position compared to the reference (after the word *be* instead of the word *master* as in the reference). *E2E* is faithful to the segmentation of the source language when it corresponds to the pauses of the speaker. The positions chosen by *Cascade* to insert breaks are before a conjunction (*comment*) and a preposition (*du*). Contrary to *E2E*, the segmentation decisions in *Cascade* are based more on syntactic patterns, learned from the existing human segmentation decisions in the training data.

The third example shows that prosody is also important for other factors related to the translation. *Cascade*, not receiving any punctuation, did not reproduce the question in the translation, while the intonation may have helped *E2E* to render the sentence as a question despite using the wrong tense (*saviez* instead of *savez*). All in all, these examples confirm our analysis and once again indicate the importance of the intersemiotic

nature of subtitling when developing automatic systems for this task.

To conclude, we have presented the first Speech Translation systems specifically tailored for subtitling. The first system is an ASR-MT cascade, while the second a direct, end-to-end ST system. These systems allow, for the first time, to create subtitles of high conformity to the subtitling constraints in the absence of a human transcription of the source speech (template). We have shown that while the two systems have similar translation quality, the E2E seems to be modelling the subtitle constraints better. We pose that this could be attributed to acoustic features, such as natural pauses, becoming available to the E2E system through the audio input. This leads to a segmentation closer to the speech rhythm, which is key to a pleasant user experience. Our work takes into account the inter-semiotic nature of subtitling by avoiding conditioning the translation on the textual source language length, as in previous approaches to NMT for subtitling, arriving to the conclusion that 42 is not the answer to everything in automatic subtitling. Rather, key elements for good automatic subtitling are also prosodic elements such as intonation, speech tempo and natural pauses.

5.3 Joint generation of captions and subtitles

In the previous section, we introduced the first ST models for automatic subtitling, overcoming the need to manually generate a source language template as input to the MT system. However, obtaining the intralingual subtitles (hereafter 'captions') is necessary for a range of applications, while in some settings captions need to be displayed along with the target language subtitles. Such 'bilingual subtitles' are useful in multilingual online meetings, in countries with multiple official languages, or for language learners and audiences with different accessibility needs (see Section 2.2.1

for more details on intra-inter- and bi-lingual subtitles). In those cases, captions and subtitles should not only be consistent with the visual and acoustic dimension of the audiovisual material but also between each other, for example in the number of blocks they occupy, their length and segmentation. Consistency is vital for user experience, for instance to elicit the same reaction among multilingual audiences, and to ensure that no one is left behind. Moreover, consistency between captions-subtitles is important in the localisation industry, to facilitate the quality assurance process.

Previous work in MT and ST for subtitling has focused on generating interlingual subtitles (Matusov et al., 2019; Karakanta et al., 2020a; Koponen et al., 2020c), a) without considering the necessity of obtaining captions consistent with the target subtitles, and b) without examining whether the joint generation leads to improvements in quality. We hypothesise that knowledge sharing between the tasks of transcription and translation could lead to such improvements. Moreover, joint generation with a single system can avoid the maintenance of two different models, increase efficiency, and in turn speed up the localisation process. Lastly, if joint generation improves consistency, joint models could increase automation in subtitling applications where consistency is a desideratum.

In this section, we address these issues for the first time, by jointly generating both captions and subtitles from the same audio source. We experiment with the following models: 1) Shared Direct (Weiss et al., 2017), where the speech encoder is shared between the transcription and the translation decoder, 2) Two-Stage (Kano et al., 2017), where the transcript decoder states are passed to the translation decoder, and 3) Triangle (Anastasopoulos and Chiang, 2018b), which extends the two-stage by adding a second attention mechanism to the translation decoder which attends to encoded speech inputs. We compare these models with the approaches for automatic subtitling proposed in the previous section: an independent

direct ST model and a cascade (ASR+MT) model. Moreover, we extend the evaluation beyond the usual metrics used to assess transcription and translation quality (respectively WER and BLEU), by also evaluating the form and consistency of the generated subtitles.

Sperber et al. (2020) introduced several lexical and surface metrics to measure consistency between transcription and translation, but they were only applied to standard, non-subtitle, texts. Subtitles, however, are a particular type of text structured in blocks which accompany the action on screen. Therefore, we propose to measure their consistency by taking advantage of this structure and introduce metrics able to reward subtitles that share similar structure and content. Our contributions can be summarised as follows:

- We employ ST to directly generate both captions and subtitles without the need for human pre-processing (transcription, segmentation).
- We propose new, task-specific metrics to evaluate subtitling consistency, a challenging and understudied problem in subtitling evaluation.
- We show increased performance and consistency between the generated captions and subtitles compared to independent decoding, while preserving adequate conformity to subtitling constraints.

5.3.1 Models

To study the effectiveness of the different existing ST approaches in the subtitling scenario, we experiment with the following models (also shown in Figure 5.1):

The **Multitask Direct (DirMu)** model consists of a single audio encoder and two separate decoders (Weiss et al., 2017): one for generating the

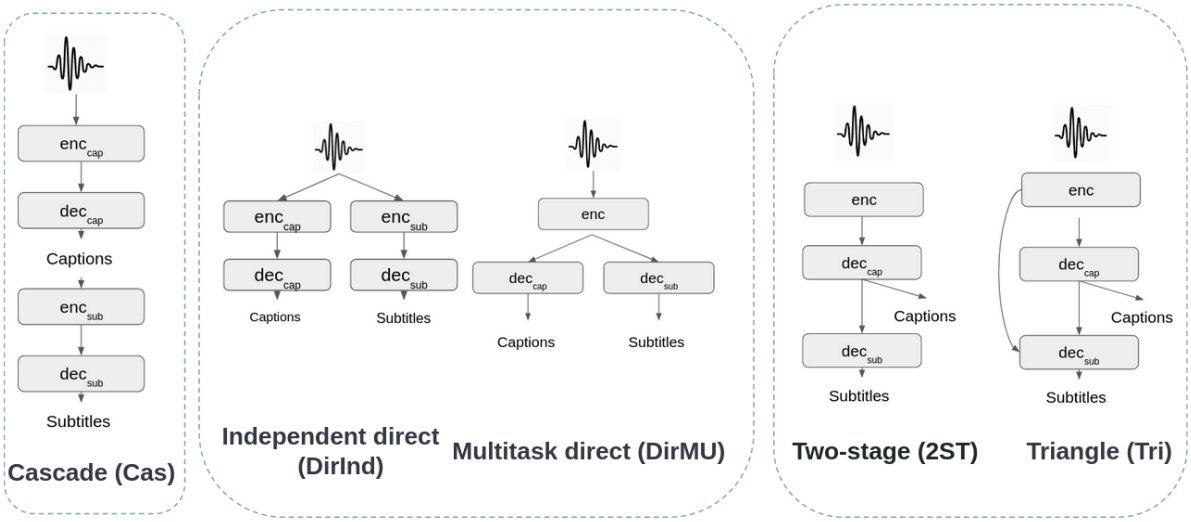


Figure 5.1: Schema of the models used in the experiments. Adapted from (Sperber et al., 2020).

source language captions, and the other for the target language subtitles. The weights of the encoder are shared. The model can exploit knowledge sharing between the two tasks, but allows for some degree of flexibility since inference for one task is not directly influenced by the other task.

The **Two-Stage (2ST)** model (Kano et al., 2017) also has two decoders. The difference here is that the translation decoder receives the transcription decoder states, instead of the encoder states. This is the only source of information for the translation decoder as it does not attend to the encoder output. Inference is not synchronous for each word position; the translation decoder 'waits' for the transcription decoder to finish decoding the transcription.

The **Triangle (Tri)** model (Anastasopoulos and Chiang, 2018b) is similar to the two-stage model, but with the addition of an attention mechanism to the translation decoder, which attends to the output embeddings of the encoder. Both `2ST` and `Tri` support coupled inference and joint training.

We compare these models with two popular ST solutions. The first is a **Cascade (Cas)** model, a combination of ASR + NMT, presented in

Section 5.2. The cascade features full dependence between transcription and translation, which will potentially lead to high consistency.

The **Direct Independent (DirInd)** system consists of two independent direct ST models, one for the transcription (as in the ASR component of the cascade) and one for the translation. It hence lies on the flexibility edge of flexibility-consistency spectrum compared to the models above.

5.3.2 Evaluation of subtitling consistency

While some metrics for evaluating transcription-translation consistency have been proposed in (Sperber et al., 2020), these do not capture the peculiarities of subtitling. The goal for consistent captions/subtitles is having the same structure (same number of subtitle blocks) and same content (each pair of caption-subtitle blocks has the same lexical content). Consider the example of a subtitle file with dual subtitles in Figure 5.2.

0:00:50,820, 00:00:53,820
To put the assumptions very clearly:
Enonçons clairement nos hypothèses : **le capitalisme,**

00:00:53,820, 00:00:57,820
capitalism, after 150 years, has become acceptable,
après 150 ans, est devenu acceptable, **au même titre**

00:00:58,820, 00:01:00,820
and so has democracy.
que la démocratie.

Figure 5.2: Source captions (En) and Target subtitles (Fr) for the same subtitle template.

Three blocks appear sequentially on the screen based on timestamp information, each of them containing one line of text in English (caption) and French (subtitle). Since the source utterance is split across the same number of blocks (3), the captions and subtitles have the same structure.

However, they do not have the same lexical content. The first block contains the French words *le capitalisme*, which appear in the second block for the English captions. Similarly, *au même titre* corresponds to the third block in relation to the captions. This is problematic because terms do not appear in the same blocks (e.g. capitalism), and also leads to sub-optimal segmentation, since the French subtitles are not complete semantic units (logical completion occurs after *hypothèses* and *acceptable*).

We hence define the consistency between captions and subtitles based on two aspects: the structural and the lexical consistency. Structural consistency refers to the way subtitles are distributed on a video. To be structurally consistent, captions and subtitles for each source utterance should be split across the same number of blocks. This is a prerequisite for bilingual subtitles, since each caption-subtitle pair has the same timestamps. In other words, the captions and subtitles should appear and disappear simultaneously. Therefore, we define **structural consistency** as the percentage of utterances having the same number of blocks between captions and subtitles.

The second aspect of subtitling consistency is lexical consistency. Lexical consistency means that each caption-subtitle pair has the same lexical content. It is particularly important for ensuring synchrony between the content displayed in the captions and subtitles. This facilitates language learning, when terms appear in similar positions, and quality assurance, as it is easier to spot errors in parallel text. We define **lexical consistency** as the percentage of words in each caption-subtitle pair that are aligned to words belonging in the same block. In our example, there are six tokens of the subtitles which are not aligned to captions of the same block: {*le capitalisme , au même titre*}. For obtaining this score, we compute the number of words in each caption aligned to the corresponding subtitle and vice versa. For each caption-subtitle pair, this process results in two lexi-

cal consistency scores: $\text{Lex}_{caption \rightarrow subtitle}$ and $\text{Lex}_{subtitle \rightarrow caption}$, where, in the former, the number of aligned words is normalised by the number of words in the caption, while, in the latter, by the number of words in the subtitle. These two quantities are then averaged into a single value (Lex_{pair}). The corpus-level lexical consistency is obtained by averaging the Lex_{pair} of all caption-subtitle pairs in the test set.

To illustrate this metric with the example in Figure 5.2, structural consistency is 100% because we have $\#blocks_{cap} = \#blocks_{sub}$. For lexical consistency, $\text{Lex}_{caption \rightarrow subtitle} = 81\%$ (words in captions present in the corresponding subtitles are 7/7, 8/10, 3/5) and $\text{Lex}_{subtitle \rightarrow caption} = 75\%$ (words in subtitles present in the corresponding captions are 5/8, 8/11, 5/5). The total lexical consistency is the average of the two, resulting in Lex = 78%.

5.3.3 Experimental setting

Data For our experiments we use MuST-Cinema (Karakanta et al., 2020b). For a sound comparison with Karakanta et al. (2020a) in Section 5.2, we conduct the experiments on 2 language pairs, English→French and English→German. We expect the captions and subtitles of TED Talks to have high consistency, since the captions serve as the basis for translating the speech in target subtitles.

The text data is segmented into subwords with Sentencepiece (Kudo and Richardson, 2018) with the unigram setting. In line with recent works in ST, we found that a small vocabulary size is beneficial for the performance of ST models. Therefore, we set a shared vocabulary of 1024 for all models except the MT component of the cascade, where it is set to 24k.

For the audio input, we use 40-dimensional log MFCC speech features. The ASR encoder was pretrained on the IWSLT 2020 data, i.e. Europarl-ST (Iranzo-Sánchez et al., 2020), Librispeech (Panayotov et al., 2015),

How2 (Sanabria et al., 2018), Mozilla Common-Voice,[7] MuST-C (Di Gangi et al., 2019a), and the ST TED corpus.[8]

Model training The ASR and ST models are trained using the same settings. The architecture used is S-Transformer, (Di Gangi et al., 2019b), with the same architecture and training settings as described in Section 5.2. The MT component is based on Transformer (big) (Vaswani et al., 2017). However, in this section, the ASR component is an end-to-end ASR which also outputs punctuation and segmentation breaks. For this reason, no other pre-processing (except for BPE) is applied to the training data. Since the goal of this section is to compare the architectures, to ensure a fair comparison with the direct and joint models, the MT component here is trained only on MuST-Cinema data. All experiments are run with fairseq (Ott et al., 2019). Training is performed on two K80 GPUs with 11 GB memory and models converged in about five days.[9]

Evaluation We evaluate three aspects of the automatically generated captions and subtitles: 1) quality, 2) conformity, and 3) consistency. For the **quality** of transcription we compute WER on unpunctuated, lowercased output, while for translation we use SacreBLEU (Post, 2018).[10] Since in Section 5.2 we observed a high correlation between BLEU_{br} and BLEU_{nb}, we report scores on the output sentences without subtitle breaks (regular WER and BLEU_{nb}). For evaluating the **conformity**, we use CPL for length conformity, CPS for reading speed and *Sigma* for the plausibility of segmentation. Lastly, we evaluate structural and lexical **consistency** between the generated captions and corresponding subtitles, as described in

[7] `https://voice.mozilla.org/`

[8] `http://iwslt.org/doku.php?id=offline_speech_translation`

[9] Our implementation of the `DirMu`, `Tri` and `2ST` models is publicly available at: `https://github.com/mgaido91/FBK-fairseq-ST/tree/acl_2021`.

[10] `BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.3`

Section 5.3.2. Word alignments are obtained using fast_align (Dyer et al., 2013) on the concatenation of MuST-Cinema training data and the system outputs. Text is tokenised using Moses tokeniser and the consistency percentage is computed on the tokenised text.

5.3.4 Results

En→Fr	WER	BLEU	CPL	CPS	*Sigma*	Struc.	Lex.
Cas	19.69	**25.99**	93 / 91	85 / 70	86.2 / 82.9	**.98**	**.99**
DirInd	19.69	22.13	93 / **94**	85 / **73**	86.2 / 83.3	.75	.86
DirMu	**17.73**	23.83	**95** / 92	85 / **73**	**87.0** / 82.1	.77	.87
2ST	19.05	23.79	94 / **94**	85 / 71	86.1 / 84.2	.83	.84
Tri	18.93	23.61	91 / 90	85 / 72	85.8 / **84.4**	.82	.92
En→De							
Cas	18.52	***18.84***	93 / 90	62 / 58	71.9 / 78.2	**.95**	**.96**
DirInd	18.52	16.79	93 / 88	62 / 59	71.9 / 76.9	.73	.86
DirMu	**16.95**	17.65	**94** / **91**	62 / 59	**73.9** / 76.0	.75	.82
2ST	18.93	*18.22*	**94** / 90	62 / **60**	70.9 / 77.6	.82	.81
Tri	19.10	*18.24*	92 / 90	62 / **60**	70.5 / **78.7**	.78	.91

Table 5.4: Results for quality (WER and BLEU), subtitling conformity (CLP for length, CPS for reading speed and *Sigma* for segmentation), and subtitling consistency (Structural and Lexical) for French and German model outputs. Conformity scores are reported for captions / subtitles. **Bold** denotes the best score. Results that are not statistically significant – according to pairwise bootstrap resampling (Koehn, 2004), $p<0.05$ – than the best score are reported in *italics*.

Transcription/Translation quality The first two columns of Table 5.4 show the WER and BLEU score for the examined models. In terms of **transcription quality**, `DirMu` (Multitask Direct) obtains the lowest WER for both languages with 17.73 for French and 16.95 for German. As far as the other models are concerned, there is a different tendency for French and

German. `Tri` (Triangle) and `2ST` (Two-Stage) perform equally better than `Cas` and `DirInd` for French, while for German `Cas` and `DirInd` have higher transcription quality than `Tri` and `2ST`. An explanation for this incongruity is that these two models perform coupled inference, therefore the benefit of the joint decoding for the transcription can be related to similarities in terms of vocabulary between the two languages. Since French has a higher vocabulary similarity to English, with many words in TED Talks being cognates (e.g. specialised terminology), it is possible that joint decoding favours the transcription for French but not for German.

When it comes to **translation quality**, `Cas` outperforms all other models for French with 26.9 BLEU points, while the differences are not statistically significant among `DirMu`, `2ST` and `Tri`. For German, however, `Cas`, `2ST` and `Tri` perform on par. The MT component trained only on MuST-Cinema data may be considered quite a weak model. The performance of the cascade would certainly increase with a stronger MT component. However, the goal of this work is to ensure a fair comparison of the architectures under equal data conditions. In addition, using additional unsegmented data (not containing subtitle breaks) may hamper with the performance of the cascade in terms of conformity and consistency. The model obtaining the lowest scores is `DirInd`. This finding confirms our hypothesis that joint decoding, despite being more complex, improves translation quality thanks to the knowledge shared between the two tasks at decoding time.

In comparison to previous work, our transcription results are contrary to Sperber et al. (2020), who obtained the lowest WER with the cascade and direct independent models. However, for translation quality our best models are `Cas`, `2ST` and `Tri`, as in previous work. Moreover, in line with Anastasopoulos and Chiang (2018b), the gains for `Tri` are higher for translation than for transcription. Comparing the BLEU score of our `DirInd` models to the $\text{BLEU}_{nb}$of the direct models in Section 5.2, we found

that our models have lower performance with 22.13 BLEU compared to 25.46 for French and 16.79 compared to 20.92 for German. It should be kept in mind that, to provide a fair comparison among models in this section, no additional data have been used for pre-training (except for ASR pre-training). Indeed, if we compare the models in this section with the E2E-small, which has not received any pre-training, the performance is higher by ∼2 BLEU points.

All in all, we found that coupled-inference, supported by `Cas`, `2ST` and `Tri`, improves translation but not transcription quality. On the contrary, multi-tasking as in `DirMu` is beneficial for transcription, possibly because of a reinforcement of the speech encoder. However, could this improvement come at the expense of conformity to the subtitling constraints?

Subtitling conformity Columns 3-5 of Table 5.4 show the percentage of captions/subtitles conforming to the length (CPL), reading speed (CPS) and segmentation (*Sigma*) constraints discussed in Section 2.2.2. We observe that joint decoding does not lead to significant losses in conformity. Specifically, the captions generated by `DirMu` have the highest conformity in terms of length (95% and 94%), reading speed (85% and 62%) and segmentation quality (87 and 74). Moreover, the high conformity score for `DirMu` correlates with the low WER, showing that quality goes hand-in-hand with conformity.

For the conformity of the target language subtitles, instead, the picture is different. The differences in conformity seem to be language-specific, therefore a clear winner among the models cannot be drawn. Even though the differences are not large, `Cas` has lower conformity to length (91% and 90%) and reading speed (70% and 58%), while it has a low segmentation score for French, but not for German. `DirMu` has by far the worst segmentation accuracy. The highest segmentation accuracy is consistently achieved

by `Tri`, which has however the lowest accuracy in captioning segmentation. There seems to be a compromise between captioning performance and subtitling performance. It is possible that the structure of the captions constrains the models (`DirMu` in that case) to segment the subtitles in positions which may not be optimal for the target language norms, and vice versa (in the case of `Tri`). We can conclude that the quality improvements of coupled inference and multi-tasking come with a slight compromise of subtitling conformity, as a result of loss of flexibility in decoding.

Subtitling consistency The last two columns of Table 5.4 present the results for the subtitling consistency. In terms of **Structural consistency** (Struc.), the model achieving the highest scores is `Cas`, with 98% and 95% of the utterances being distributed along the same number of blocks. As expected, the lowest structural consistency is achieved by `DirInd` (75% and 73%), which determines independently the break positions. Among the joint models, `Tri` outputs captions and subtitles with higher consistency than `DirMu`, but both are outperformed by `2ST` (83% and 82%). Our hypothesis is that by attending only the caption decoder, `2ST` behaves similarly to the cascade, and the translation decoder better replicates the block structure. We noted that the reference captions and subtitles have lower consistency (92% both for French and German) than the cascade. This shows that the cascade copies the same <`eob`>tokens and achieves extreme structural consistency, which is a desideratum for our study case but may be harmful in other scenarios, since it leads to lower conformity (see results on conformity above). Indeed, in scenarios where consistency is not a key, subtitlers should have the flexibility to adjust segmentation, time codes, as well as delete, merge and split the subtitles to suit the needs of their target languages (Oziemblewska and Szarkowska, 2020).

The **Lexical consistency** (Lex.) results show that `Cas` is the model

with the highest content overlap in parallel caption-subtitle blocks with 99% and 96% of the words being aligned to the same block. As with the structural consistency, the lexical consistency of the cascade is higher than the references (95% for French, 86% for German). The direct model with the highest lexical consistency is `Tri` (92% and 91%). Interestingly, despite its high structural consistency, `2ST` does not distribute the content consistently in the parallel blocks, achieving the lowest conformity (81%). The `DirMu` also achieves lower consistency than `DirInd` for German (82% compared to 86%) but not for French (87% compared to 86%). It is worth noting that lexical consistency is generally lower for German than for French. Indeed, a 100% lexical consistency between subtitles in languages with different word order is not always feasible or even appropriate. For example, the main verb in an English subordinate clause appears in the second position while in German at the end of the sentence. In order to adhere to grammatical rules, words in subtitles of different languages often have inter-block reordering. Therefore, the balance between flexibility and consistency is manifested here as a compromise between grammaticallity and preservation of the same lexical content in each pair of subtitles.

To sum up, the results of structural consistency show that the models are able to preserve the block structure between captions and subtitles in more than 75% of the utterances. In addition, the high lexical consistency shows that the block symbols are not inserted randomly, but placed in a way that preserves the same lexical content in the parallel blocks.

All in all, our results show that the evaluation of captions and subtitles is a multifaceted process that needs to be addressed from multiple aspects: quality, conformity and consistency. Missing one of the three can lead to wrong conclusions. For instance, only considering quality and consistency could lead one to disregard the importance of conformity and consider independent solutions an obsolete technology. Secondly, among

the Direct architectures, the use of techniques that allow linking the generation process of captions and subtitles helps to achieve overall better quality and consistency than independent decoding, with a slight discount in conformity. Between the `DirMu`, `2ST` and `Tri`, there is not a model that outperforms all the others in all the metrics, so the choice mainly depends on the application scenario. Lastly, comparing the Cascade and the Direct, the Cascade seems to be the best choice for generating high-quality captions and subtitles with maximum consistency, also given the (to date) limitation of the direct systems of not generating timestamps. Still, the recent advancements in Direct approaches result in competitive solutions with increased efficiency of maintaining one model for both tasks.

5.4 Live subtitling

The last application scenario to which we attempt to bring the new paradigm of automatic subtitling is live subtitling. Live subtitling, whether intralingual or interlingual, allows for obtaining subtitles in real time for achieving barrierless interaction between audiences with diverse linguistic and accessibility needs. As mentioned in Section 2.2.1, live subtitles were initially obtained by means of keyboards or stenotyping, but Automatic Speech Recognition (ASR) gave rise to respeaking, a technique where the subtitler respeaks the original sound into an ASR system, which produces the subtitles on the screen (Romero-Fresco, 2011). Although originally employed for intralingual subtitles, respeaking is gradually extending to produce interlingual subtitles, a task bearing close resemblance to simultaneous interpreting. Still, the tediousness of the task and the scarcity of highly-skilled professionals for live subtitling call for a more pronounced role of technology for providing real-time access to information.

The growing needs for access to multilingual spoken content have moti-

vated researchers to develop fully automatic solutions for real-time spoken language translation (Grissom II et al., 2014; Gu et al., 2017; Alinejad et al., 2018; Arivazhagan et al., 2019b; Ma et al., 2019). The new possibilities opened up by neural machine translation have led to improvements in automatic simultaneous speech-to-text translation (SimulST). In SimulST (Niehues et al., 2018; Ma et al., 2020; Ren et al., 2020), the generation of the translation starts before the entire audio input is received, which is an indispensable characteristic for achieving low latency (translation delay) between speech and text in live events. As mentioned in Section 3.2.4, the translation becomes available at consecutive steps, usually one word at a time. However, a display mode based on the word-for-word rate of generation of SimulST systems may not be optimal for displaying readable subtitles. Studies in intralingual subtitling have shown that a word-for-word display increases the number of saccadic crossovers between text and scene (Rajendran et al., 2013) and leaves viewers less time to look at the images (Romero-Fresco, 2010). For this reason, regulators, such as the UK Office of Communications, recommended displaying subtitles in blocks (Ofcom, 2015). Although leading to more readable subtitles, this display mode is not ideal for live events since waiting until the block is filled before displaying the subtitle would extremely increase latency at the risk of completely losing synchronisation with the speaker. Despite the existence of some applications of SimulST, so far no work has explored its potential for live subtitling and how the delay in generation impacts subtitle readability.

Given the burgeoning demand in live subtitles and previous studies on the readability of live subtitles, in this section we pose the following research questions: 1) Can automatic simultaneous translation be a viable method for producing live interlingual subtitles? 2) What are the challenges of the generation mode of SimulST systems for the readability of the subtitles? Our contributions can be summarised as follows:

- We explore the performance of a direct SimulST system on three language pairs (En→It, De, Fr) in terms of translation quality, and its ability to generate readable subtitles in terms of subtitling constraints.
- We investigate two methods for displaying live subtitles, i.e. i) word-for-word and ii) blocks, and how the display mode affects their readability (reading speed and delay).
- We propose scrolling lines, a mixed display method which takes advantage of the ability of our system to define proper line breaks and show that it leads to a more comfortable reading speed at an acceptable delay. Lastly, we discuss challenges and recommendations for applying simultaneous translation for live subtitling.

5.4.1 SimulST for live subtitling

Simultaneous ST models The SimulST systems we used are based on direct ST models (Bérard et al., 2016; Weiss et al., 2017). These systems have been shown to have lower latency (Ren et al., 2020) – an important factor in simultaneous systems – compared to cascade systems, which perform two generation steps. Moreover, in Section 5.2, we suggested that direct ST systems, having access to the audio source, make better subtitle segmentation decisions by taking advantage of the pauses in the audio.

To adapt SimulST systems for the task of live interlingual subtitling, we force the system to learn from human subtitle segmentation decisions in the training data of MuST-Cinema. These break symbols, if positioned properly, are the key element which allows us to experiment with different display modes for live subtitles.

Our direct SimulST models combine the efficiency of the wait-k strategy (Ma et al., 2019), and our findings in Section 5.2 for obtaining readable subtitles with direct ST systems. This decision policy was also chosen

step	Word-for-word	Blocks	Scrolling lines
In	In		
alcuni	In alcuni		
casi	In alcuni casi		
è	In alcuni casi è		
perché	In alcuni casi è perché		
non	In alcuni casi è perché non		
sono	In alcuni casi è perché non sono		
adatti	In alcuni casi è perché non sono adatti		
<eob>	In alcuni casi è perché non sono adatti	In alcuni casi è perché non sono adatti	In alcuni casi è perché non sono adatti
o	In alcuni casi è perché non sono adatti o	In alcuni casi è perché non sono adatti	In alcuni casi è perché non sono adatti
non	In alcuni casi è perché non sono adatti o non	In alcuni casi è perché non sono adatti	In alcuni casi è perché non sono adatti
hanno	In alcuni casi è perché non sono adatti o non hanno	In alcuni casi è perché non sono adatti	In alcuni casi è perché non sono adatti
etica,	In alcuni casi è perché non sono adatti o non hanno etica,	In alcuni casi è perché non sono adatti	In alcuni casi è perché non sono adatti
<eob>	In alcuni casi è perché non sono adatti o non hanno etica,	o non hanno etica,	In alcuni casi è perché non sono adatti o non hanno etica,
ma	In alcuni casi è perché non sono adatti o non hanno etica, ma	o non hanno etica,	In alcuni casi è perché non sono adatti o non hanno etica,
spesso,	In alcuni casi è perché non sono adatti o non hanno etica, ma spesso	o non hanno etica,	In alcuni casi è perché non sono adatti o non hanno etica,
ci	In alcuni casi è perché non sono adatti o non hanno etica, ma spesso ci	o non hanno etica,	In alcuni casi è perché non sono adatti o non hanno etica,
hanno	In alcuni casi è perché non sono adatti o non hanno etica, ma spesso ci hanno	o non hanno etica,	In alcuni casi è perché non sono adatti o non hanno etica,
condotto	condotto	o non hanno etica,	In alcuni casi è perché non sono adatti o non hanno etica,
<eol>	condotto	o non hanno etica,	o non hanno etica, ma spesso, ci hanno condotto
a	condotto a	o non hanno etica,	o non hanno etica, ma spesso, ci hanno condotto
obiettivi	condotto a obiettivi	o non hanno etica,	o non hanno etica, ma spesso, ci hanno condotto
sbagliati.	condotto a obiettivi sbagliati.	o non hanno etica,	o non hanno etica, ma spesso, ci hanno condotto
<eob>	condotto a obiettivi sbagliati.	ma spesso, ci hanno condotto a obiettivi sbagliati.	ma spesso, ci hanno condotto a obiettivi sbagliati.

Figure 5.3: Example sentence displayed in the three different modes. Words on the left column correspond to the time steps.

because it allows us to control the latency of our systems. In this way, we can study the effect of latency both on the conformity of the subtitles and on the subtitle display modes.

Display modes We experiment with the following three display modes: 1) word-for-word, 2) blocks, and 3) scrolling lines. Figure 5.3 shows an example of a subtitle displayed in the different modes.

Word-for-word: In the word-for-word display, words appear sequentially on the screen as soon as they are generated by the system. One line at the bottom of the screen is filled from left to right until no more space is available on the right. Then, the line disappears and a new line is filled again. As already mentioned, this display mode is used in most simultaneous applications and follows the generation process of the SimulST system.

Naturally, the length of the subtitle depends on the size of the screen. In our case, we selected a maximum of 84 characters, which matches the max. length of a full subtitle block of the block method (see below) and approximates the length observed in the STACL demo.[11]

Blocks: In this mode, the subtitles are displayed only when a full subtitle block is completed. This means that the system continues generating words which are only displayed when the block delimiter <eob> is generated. The subtitle block remains on screen until the next <eob> symbol is generated. In this way, the first subtitle is substituted by the next one and so on. Display in blocks is used primarily in offline subtitling, where subtitles are prepared beforehand.

Scrolling lines: Instead of waiting for the full block, we propose a mode in which each line is displayed as soon as a break is predicted (either <eob> or <eol>). Whenever a new break is predicted, the previous line moves to the upper row of the block and the new line occupies the lower row. Since the allowed number of lines in a block is two, each line moves from the lower to the upper row before disappearing. This mode combines the benefits of the two previous methods. It reduces the dynamicity of the text compared to the word-for-word display, it makes content available earlier than the block display, it allows for access to extended context and it reduces long-distance eye movements since the previous line appears above. This mode is similar to the most popular mode employed for broadcasting, with the difference that there is no word-for-word display inside the lines.

The last two display modes are possible because of the ability of our SimulST system to predict subtitle breaks, which was not considered before in SimulST.

[11] https://simultrans-demo.github.io/

5.4.2 Evaluation of display modes

Our evaluation of the generated subtitles follows the Ofcom recommendations (Ofcom, 2015) and focuses on three key dimensions: quality, reading speed and delay between utterance and subtitle. For the display modes the quality is fixed, since they are applied to the same output. We thus focus on the speed and delay. In general, the reading speed is computed in characters per second (CPS), as the number of characters over the total time of display. A low but also constant reading speed is pivotal for user experience, considering that it represents how fast or slow a user has to read. Since each of the display methods described in Section 5.4.1 has a different granularity, the computation of the reading speed has to take into account these visualisation differences. The computations of reading speed and delay are described in detail for each visualisation mode below.[12]

Reading Speed. In the **Word-for-word** mode (see Figure 5.3) a word appears on screen as soon as it is generated and remains on screen until the block changes, i.e. when the 84-character limit is reached. Thus, the amount of time available for a user to read this word and all the following words of the block – hereinafter display time – is the interval elapsed from the generation of the word to the generation of the first word of the successive block. Consequently, the reading speed can be computed at each generation step as the length (in characters) of the generated word and of the successive words of the block divided by the display time. After computing the reading speed for each word of a block, the block-level reading speed is obtained by taking the maximum value since this represents the speed required for a user to avoid losing any part of the text.

Formally, each block corresponds to a group of words of maximum 84 characters or to the last group of words before the end of sentence symbol

[12]The code is available at: `https://github.com/sarapapi/reading_speed`

(<eos>) is emitted, i.e. the end of the audio segment. Thus, a block is composed of a set of W words $w_1, ..., w_W$ emitted at times $t_1, ..., t_W$ (in seconds). At time t_{W+1}, the successive block starts or <eos> is emitted.

With this notation, we can compute the reading speed as the ratio of the number of characters of the generated text and the time elapsed until its generation, as follows:

$$rs = \max_{i=1,...,W} \frac{len(text_i)}{elapsed_i} \tag{5.2}$$

where:

$$elapsed_i = \begin{cases} \texttt{DELAY_K}, & w_i = <\texttt{eos}> \\ t_{i+1} - t_i, & i = W \\ t_{i+1} - t_i + elapsed_{i+1}, & otherwise \end{cases} \tag{5.3}$$

$$text_i = \begin{cases} w_i & i = W \\ w_i + \texttt{SPACE} + text_{i+1} & otherwise \end{cases} \tag{5.4}$$

and SPACE represents a blank space added between the two texts. Practically, the time elapsed for each word is the time interval between its emission and the emission of the next word. The total time elapsed is the sum of the single time intervals. If the block is the last block of the audio segment, we do not know the emission time of the next word. For this reason, we use a fictitious delay:

$$\texttt{DELAY_K} = 0.280s \cdot k$$

where k corresponds to the k of the wait-k policy. This amount of time is a lower bound for the generation time of the first token of the next segment. As the wait-k policy reads for k steps (each of them lasting 280ms) and then generates the first word, the actual elapsed time will always be higher since it includes the time required for generating the word. Thus, DELAY_K

represents a conservative estimation of the time available to read the last word.

In the **Blocks** mode, a block stays on screen until the next block is filled. The reading speed is computed at block level, dividing its length by the time elapsed between the display time of the current block and that of the successive one. In an audio segment with B blocks, each block b is composed by W_b words $w_1, .., w_{W_b}$, where w_{W_b} is the <eob>. Each word w_i is emitted at time t_i, thus each block b is emitted at time t_{W_b}, hereinafter t_b. Using this notation, Equations 5.3 and 5.4 become:

$$elapsed_b = \begin{cases} \texttt{DELAY_K}, & b = B \\ t_{b+1} - t_b, & otherwise \end{cases} \tag{5.5}$$

$$text_b = \sum_{i=1}^{W_b - 1} \begin{cases} \texttt{BLANK}, & w_i = \texttt{<eol>} \\ w_i, & i = W_b - 1 \\ w_i + \texttt{SPACE}, & otherwise \end{cases} \tag{5.6}$$

where BLANK corresponds to the empty string and, as before, DELAY_K conservatively accounts for the last unknown block time. $text_b$ is the sum of the characters of the individual words w_i. The symbols <eol> and <eob> (the W_b-th word) are not counted as characters in the reading speed computation since they are only used for formatting the subtitles and are not read by the user. Consequently, the reading speed of a block is:

$$rs = \frac{len(text_b)}{elapsed_b} \tag{5.7}$$

Our proposed display mode, **Scrolling Lines**, also exploits the subtitle structure considering both <eob> and <eol> as a unique <eol> delimiter. Since each line scrolls up when another is generated and two lines stay together on the screen, each line is displayed until the next two lines are generated. As a consequence, the reading speed is computed at line

level, dividing the length of a line by the time needed to generate the two successive lines.

If we denote L as the number of lines present in an audio segment, then each line l is composed by W_l words $w_1, .., w_{W_l}$ emitted at times $t_1, ..., t_{W_l}$, where w_{W_l} is the <eol> and $t_{W_l} = t_l$ is its emission time. In this case, the reading speed is calculated at line-level instead of block-level, considering that each line is displayed until the next two lines (since a block can be composed by two lines) are produced. Thus, Equations 5.5, 5.6 and 5.7 are modified as follows:

$$rs = \frac{len(text_l)}{elapsed_l} \tag{5.8}$$

$$elapsed_l = \begin{cases} \texttt{DELAY_K}, & l = L \\ (t_{l+2} - t_{l+1}) + (t_{l+1} - t_l), & otherwise \end{cases} \tag{5.9}$$

$$text_l = \sum_{i=1}^{W_l - 1} \begin{cases} w_i, & i = W_l - 1 \\ w_i + \texttt{SPACE}, & otherwise \end{cases} \tag{5.10}$$

where, in this case, DELAY_K conservatively accounts for the last unknown line time.

Delay. The delay is estimated as the time between speech and subtitling. While in intralingual subtitling the correspondence between audio and subtitle is easier to establish, the interlingual setting poses the challenge of finding the correspondences between source audio and target text. Since a word-aligner would capture semantic correspondences, we opt for a temporal-based correspondence, based on system lagging. Therefore delay is calculated as:

$$delay = \sum_{i=1}^{w_i} (t_display_w - t_received_w - \texttt{DELAY_K}) \tag{5.11}$$

where $t_display_w$ is the time the word was displayed on screen and $t_received_w$ the utterance time corresponding to the displayed token, minus the k-wait delay. For word-for-word, the time of display corresponds to the system elapsed time, therefore the delay equals the system lagging. For block and scrolling lines, the display time is the time elapsed for the generation of the break (<eob> for blocks or any break for scrolling lines).

5.4.3 Experimental setting

Data For our experiments we use MuST-Cinema (Karakanta et al., 2020b). This corpus is ideal for exploring display modes other than word-for-word because it contains subtitle breaks as special symbols. We conduct experiments on three language pairs, English→Italian,German,French. The text data were tokenised using SentencePiece (Kudo and Richardson, 2018) with the unigram setting (Kudo, 2018), trained on the training data with a 10k-token vocabulary size. The source audio was pre-processed with SpecAugment (Park et al., 2019), then the speech features (80 log Mel-filter banks) were extracted and Cepstral Mean and Variance Normalization was applied at global level. The configuration parameters are the default ones set by (Ma et al., 2020).

Training settings Our SimulST systems are Transformer-based models (Vaswani et al., 2017), composed of 12 encoder layers, 6 decoder layers, 256 features for the attention layers and 2,048 hidden units in the feed-forward layers. All models are based on a custom version of (Wang et al., 2021), having two initial 1D convolutional layers with *gelu* activation functions (Hendrycks and Gimpel, 2020), but adapted to the simultaneous scenario as per (Ma et al., 2020). Moreover, the encoder self-attentions are biased using a logarithmic distance penalty (Di Gangi et al., 2019b), leveraging the local context. Training was performed with cross entropy

loss, Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1e-4 with an inverse square-root scheduler and 4,000 warm-up updates. We set mini-batches of 5,000 max tokens and update the gradients every 16 mini-batches. The best checkpoint was selected based on the lowest cross entropy value on the MuST-Cinema dev set. READ/WRITE actions of the wait-k policy are decided by means of a pre-decision module at BPE (token) level (Ma et al., 2020). The adopted pre-decision module is fixed, which triggers the decision process at every pre-defined number of frames. Since a frame covers 10ms of the audio, an encoder state covers 40ms due to a 4x subsampling by the initial convolutional layers. Since the average length of a word in MuST-Cinema is 270ms, we consider 7 encoder states for a READ/WRITE action, which is the default parameter used by (Ma et al., 2020), leading to a window size of 280ms. To explore the quality vs latency compromise and to study the effect of the system latency on the delay of the subtitles, we experimented with two values of k, resulting in wait-3 and wait-5 models. For comparison, we also trained one offline ST system per language.

Evaluation The evaluation focuses on two different aspects: 1) system performance and 2) display modes. For performance, quality is evaluated with BLEU_{br}using SacreBLEU (Post, 2018). The latency of the system is evaluated with Average Lagging (AL) (Ma et al., 2019) adapted to the ST scenario by (Ma et al., 2020). To test the ability of the systems to generate length-compliant subtitles, we report CPL. The display modes are evaluated in terms of reading speed and delay, as described in Section 5.4.2.

5.4.4 Results

Quality, Latency and Conformity As far as quality is concerned (Table 5.5), the wait-3 strategy achieves low BLEU scores but there is significant im-

Model	en-it				en-de				en-fr			
	BLEU	AL	CPL	*S*	BLEU	AL	CPL	*S*	BLEU	AL	CPL	*S*
offline	17.57	-	95%	81	13.97	-	96%	76	16.70	-	97%	80
wait-3	10.90	1755	93%	77	6.60	1422	96%	76	12.49	1570	94%	78
wait-5	13.52	1936	92%	80	9.75	2050	94%	72	16.40	2035	95%	79

Table 5.5: BLEU, Average Lagging (AL) in ms, conformity to the length constraint (CPL) and segmentation (*Sigma*) on three language pairs of MuST-Cinema.

provement for wait-5, even reaching the performance of the offline system for French. These scores are in line with those reported in SimulST settings, while in our setting the difficulty is exacerbated by the requirement to correctly place the subtitling breaks. In fact, the offline system, despite not being optimised with the best settings and hyper-parameters specifically for the offline mode, still performs comparatively to our offline systems in Section 5.2 for quality (21.03 BLEU for French and 14.25 for German) and segmentation accuracy (83 for French and 79 for German), while their length conformity is higher by ∼2%. Segmentation is higher for the offline models, while for the simultaneous there is no clear winner. As for latency (AL), we observe an increase between 0.2-0.6 seconds from wait-3 to wait-5, which lags behind by 2 seconds. Still, these spans are lower than the average Ear-to-Voice Span (EVS) threshold reported for intralingual respeaking (2.1 seconds) and way below the EVS for interlingual respeaking (4.1 secs) (Chmiel et al., 2017). This shows that, despite the poor quality, SimulST could have the potential of reducing the delay in interlingual live subtitling. In terms of conformity, our systems properly insert the break symbols, despite receiving partial input, since more than 90% of the generated subtitles conform to the length constraint. This ability of our SimulST systems is indispensable for taking advantage of the predicted subtitle structure to experiment with display modes in blocks and lines.

Display mode and reading speed When comparing the reading speed (*rs*) of the three modes (Table 5.6), the word-for-word and block mode have the highest *rs* for the wait-5 and wait-3 strategy respectively. The standard deviation is much higher for the word-for-word mode, which indicates a large variation in *rs*. This could be attributed to the generation rate of SimulST systems. Because of the k-wait policy, these systems wait at the start of the utterance but, when the end of the input is reached, they perform greedy search and emit all remaining words at once. This rate leads to a jerky display, where some words remain on screen for a long time and others flush before the viewer manages to read them. However, the block mode has the lowest percentage of subtitles achieving a reading speed of max 21 CPS. The problem of this mode is that each block remains on screen until the next <eob> is generated, which corresponds to the duration of the following block. For example, if a block of two lines with 40 characters each is followed by a block of one line of 25 characters, the first block has a short time to be displayed, resulting in a high *rs*, and vice versa. One future direction would be to adjust the time of each block to better accommodate its reading speed, but, in initial experiments we found that this approach led to excessively high delay. Scrolling lines, our proposed method, achieves by far the lowest mean *rs*, with all models scoring below the 21 CPS threshold. The same tendency is shown for the percentage of conforming subtitles, where conformity to reading speed reaches ∼80%. It is worth noting that *rs* increases from wait-3 to wait-5 for the block and line modes in en→de, contrary to the other languages. This correlates with the lower length conformity (94% for wait-3 to 91% for wait-5) and shows the importance of correctly predicting the break positions for the success of the display methods relying on these breaks.

As for delay, the word-for-word mode has the lowest delay, which corresponds to the system lagging. The block mode has the highest delay,

while our proposed method reduces the delay by 0.6 seconds on average compared to the block display, remaining close to a 4-second EVS. Our results are validated by the inversely proportional relationship between *rs* and delay. Scrolling lines seems to achieve a fair compromise between a comfortable reading speed and an acceptable delay, while combining the benefits of the presence of the word on the right, less dynamic text and the preservation of the block structure which is familiar to most viewers.

	display	*wait-3*			*wait-5*		
	mode	rs	$\leq$ 21 CPS	delay	rs	$\leq$ 21 CPS	delay
en-it	word	53.5 ± 9.9	61%	1755	40.1 ± 8.0	70%	1936
	block	53.4 ± 9.1	38%	4690	36.5 ± 7.0	62%	5004
	line	17.6 ± 5.2	79%	4092	14.4 ± 4.1	85%	4461
	display	*wait-3*			*wait-5*		
	mode	rs	$\leq$ 21 CPS	delay	rs	$\leq$ 21 CPS	delay
en-de	word	29.1 ± 7.2	70%	1422	58.4 ± 10.5	63%	2050
	block	33.3 ± 6.2	37%	4772	52.6 ± 9.2	56%	4503
	line	12.4 ± 3.7	85%	4090	19.9 ± 5.4	78%	3894
	display	*wait-3*			*wait-5*		
	mode	rs	$\leq$ 21 CPS	delay	rs	$\leq$ 21 CPS	delay
en-fr	word	39.7 ± 7.9	55%	1570	53.4 ± 9.0	57%	2035
	block	43.8 ± 7.6	37%	4872	46.1 ± 8.0	56%	5273
	line	15.4 ± 4.2	78%	4217	18.4 ± 4.8	78%	4708

Table 5.6: Reading speed (*rs*) mean and standard deviation in characters per second (CPS), percentage of subtitles with a *rs* of max. 21 CPS and display delay (in ms) on three language pairs of MuST-Cinema.

To sum up, in this section we adapted SimulST systems for the task of live subtitling, by forcing the systems to generate subtitle breaks. We showed that SimulST systems are able to generate properly-formed subtitles. Given this finding, we moved on to explore display strategies alternative to the word-for-word display, the established display mode in SimulST. Word-for-word display is sub-optimal for readability and com-

prehension (Romero-Fresco, 2010). For automatically generated live subtitles, we found that it leads to an extremely variable reading speed, with some words lagging on the screen while the words towards the end of the utterance flushing through the screen. On the other hand, the display in blocks, which is the traditional mode for displaying offline subtitles, leads to a large delay and improves the reading speed only for SimulST systems with a higher latency. Our proposed display method, scrolling lines, is the only one achieving a comfortable mean reading speed below 21 CPS, with around 80% of the subtitles having acceptable reading speed, while the delay remains around the 4-second threshold. Our findings are only suggestive and are based on the existing literature on subtitle reading. They should be complemented with actual user studies, possibly using behavioural techniques, to determine which display mode leads to a more pleasant user experience. Unfortunately, the low translation quality of current systems will very likely hamper the possibility of observing a real effect of the display mode.

As for the feasibility of SimulST for live subtitling, there is still a long way to go in several directions. From a technical point of view, the principal issue is still the poor translation quality, which could benefit from advancements in tailored architectures. Improvements in the evaluation of reading speed and delay are also required, with metrics correlating with viewers' reading patterns. Lastly, scholarly work is needed around user perception studies in automatic subtitling and live interlingual subtitling. Technology is moving faster than research and user studies are key to ensure our implementational efforts are moving in the right direction. We hope the work presented in this section will set the ball rolling for further research in automatising live interlingual subtitling.

5.5 Conclusions

In this chapter we have presented direct ST models for three different subtitling scenarios. For offline subtitling (Section 5.2), our results showed that ST could be used as an alternative to template creation, compromising translation quality by ∼6 BLEU points in exchange for independence from a human-created template. Between cascade and direct subtitling approaches, there seems to be no clear winner since translation quality is on par for both ST systems. Direct systems generate subtitles with higher length conformity but the quality of segmentation seems to depend on the target language. Still, direct systems seem to have an edge on selecting the correct type of subtitle break, possibly because of access to the source audio. In Section 5.3, we explored the feasibility of jointly generating captions and subtitles for the same audio input. We tested the models in terms of quality, conformity and consistency. We found that models linking the caption and subtitle generation process, either by receiving the captions or the decoder states of the captioning decoder, have higher translation performance and enhanced consistency compared to independent models. On the contrary, direct models, including the ones using multi-tasking techniques, have the highest captioning performance and conformity but to the expense of subtitling quality and consistency. Lastly, in Section 5.4, we experimented with simultaneous ST for producing live subtitles. We showed that simultaneous models produce readable subtitles in terms of length and segmentation, with a considerably lower delay compared to human respeakers and interpreters. This allowed us to experiment with display modes other than the default word-for-word display used in simultaneous translation applications. This display mode has been found to be sub-optimal for comfortable reading and comprehension. When displaying the generated subtitles in a scrolling lines fashion, we obtained a lower

and more uniform reading speed, with a reasonable increase in delay. Despite these encouraging results, the low translation quality is the major impediment to applying direct simultaneous ST systems for obtaining live subtitles and for evaluating the display modes through user studies. The focus of the next chapter is the evaluation of subtitling-specific aspects.

Chapter 6

Evaluation of Subtitle-specific Aspects

Progress in a field can be guaranteed only when this progress can be measured in reliable and robust ways. The third, but not less crucial, component of addressing subtitling from a computational perspective is evaluation. We have argued that quality in subtitling depends on multiple variables, both related to the multimodal nature of the source and the formal aspects of the target. For this reason, MT evaluation paradigms may be inadequate for capturing all aspects of quality in subtitling. In this chapter, we address the need for metrics and methodologies to evaluate subtitling-specific aspects. In Section 6.1 we present a short review of previous experimental designs in evaluating MT for subtitling and outline the additional challenges for evaluation under the new paradigm of automatic subtitling. Section 6.2 is dedicated to the evaluation of subtitle segmentation. We present *Sigma*, a subtitle segmentation score derived from an upper bound of BLEU, which allows for evaluating the segmentation for end-to-end subtitling systems. We then move to an analysis of autospotting and its effect on the subtitle post-editing process (Section 6.3). To this aim, we collect high-quality post-edits from professional subtitlers and release a corpus containing product-, process- and participant-based

subtitling data. Finally, in Section 6.4 we conduct a survey of subtitlers' perspectives when post-editing automatically-generated subtitles. Based on the collected feedback, we draft a series of recommendations for improving automatic subtitling tools in the future.

6.1 Is BLEU all you need?

In the previous chapter, we showed the feasibility of producing quality subtitles for a variety of subtitling scenarios with direct ST models, which do not rely on a human template. In addition, several tools are being deployed featuring fully automatic generation of subtitles, either for gisting purposes (e.g. Youtube automatic captioning) or for supporting subtitlers and content creators (e.g. Matesub, OOONA[1]). This transition from partial to extended automation in subtitling leads to the question: Are experimental designs and evaluation metrics developed for the MT paradigm adequate for evaluating and investigating automatic subtitling as an extrinsic task?

To answer this question, we reviewed the experimental designs from previous work in MT for subtitling (see Section 3.4). Experimental designs for the evaluation of MT for subtitling have focused on translation quality and productivity gains. Quality has been evaluated automatically using string-matching metrics against reference translations, with BLEU (Papineni et al., 2002) being the most prominent. Several studies used multiple metrics and reference translations (when available) (Volk et al., 2010; Etchegoyhen et al., 2014; Matusov et al., 2019; Koponen et al., 2020c), recognising that string-based metrics are not sufficient for representing the functional equivalence required by creative and multimodal texts such as subtitles. However, no metrics have been proposed for evaluating subtitle-specific aspects such as spotting and the placement of line breaks in interlin-

[1]https://ooona.net/

gual subtitling. Productivity gains have been assessed through post-editing (PE), following the paradigm of PE-based evaluation in MT research. PE-based evaluations of translated subtitles (Popowich et al., 2000; de Sousa et al., 2011; Aziz et al., 2012) have been performed in a text-oriented fashion, often outside the interfaces used by subtitlers, without access to the video, ignoring the effort of correcting the timestamps or segmentation. These choices suggest that previous designs have considered a narrow range of subtitling scenarios (template translation) and have mainly ignored the dimensions of time and form, which impact on both the viewers' perception of quality and the subtitlers' effort to edit the output to its final acceptable format (Szarkowska et al., 2020). Moreover, user feedback, despite being indicative of the usefulness of a technology, has been collected in few studies and only in unstructured ways. From these elements, we concluded that current evaluation metrics and methodologies capture only partially the aspects of quality and effort related to the complex nature of subtitling.

The rising paradigm of automatic subtitling poses extra challenges for experimental designs and evaluation. The first one is the *variable number of segments*. MT quality metrics are computed based on a one-to-one correspondence between reference-output sentences. Moreover, in MT Post-Editing (MTPE), rarely do translators merge or split sentences. For subtitling, however, adding, deleting, splitting and merging subtitles is the norm, since target subtitles may be segmented in different, but equally plausible ways. This leads to a variable number of segments both in relation to the source and the reference. As a result, the subtitles cannot be evaluated directly with automatic metrics without any previous alignment. Second, *autospotting and segmentation come into play*. Even though they are important quality parameters (Section 2.2.3) and contribute to the temporal, technical and cognitive effort, evaluating their quality and effort required to bring them to optimal quality is still an open question.

Lastly, *disentangling the variables* contributing to quality and effort is tedious. The technical requirements determine the degree of text condensation required and thus have a direct impact on the translation. Vice versa, changes in the wording may require adjustments of the automatically predicted timestamps or line breaks. As a result, assessing quality becomes a problem composed of multiple variables, some of which may be weighted more than others. For example, is a perfect translation with implausible line breaks and a fast reading speed better than an erroneous translation with proper breaks and comfortable reading speed? Determining proper weights for each of these variables requires insights from MT evaluation and experimental studies in AVT perception and production. For PE, estimating how translation, spotting and segmentation edits contribute to the effort is equally intricate. Time is a sensitive measure difficult to interpret, since it is affected by many factors. The total time spent post-editing a subtitle contains both the time to change the wording and the time of adjusting auto-spotting, among other cognitive processes. Understanding the impact of each element on the PE process is therefore cumbersome.

From the above it becomes clear that evaluation of automatic subtitling systems is still at an embryonic but burgeoning phase. To promote automatic subtitling, evaluation should move beyond BLEU and comprise all dimensions of subtitling. Other steps towards reproducible and reliable experimental research in automatic subtitling include improved, task-tailored interfaces for PE, more detailed reporting standards[2] and high-quality, public test data and benchmarks. In the following sections, we address several of the challenges in evaluating automatic subtitling, by proposing a new metric for segmentation, exploring the effect of auto-spotting on the PE process and conducting a comprehensive survey of subtitlers' experi-

[2]In an attempt to promote better reporting in research in automatic subtitling, we publish a reproducibility checklist, which can be found in Appendix B.

ence when post-editing automatic subtitles.

6.2 Evaluating subtitle segmentation

In this section, we study ways to evaluate the quality of the output segmentations delivered by end-to-end subtitling systems. Automatising the generation of subtitles has been a long-standing issue, and is nowadays more and more often performed with neural models trained end-to-end (Lakew et al., 2019; Liu et al., 2020a). Contrary to cascaded systems, which typically contain an independent segmentation module that can be evaluated as a standalone component by simply measuring its ability to reproduce a reference segmentation of a reference text (*perfect text*), end-to-end systems directly output a segmented text, which may not correspond to the reference subtitles (*imperfect text*). Separating text from segmentation errors thus becomes an issue. The textual changes do not allow for computing standard segmentation metrics (e.g. precision, recall, F1 (Álvarez et al., 2016), WindowDiff (Pevzner and Hearst, 2002), Boundary Similarity (Fournier, 2013)) for judging the quality of segmentation.[3] Figure 6.1 shows an example of a **perfect text** which can be evaluated with standard metrics because the textual content is identical to the reference text, and of an **imperfect text**, where computing them would be impeded by the differences in the textual content compared to the reference. In Chapter 5, we attempted to address this problem with metrics such as BLEU$_{br}$ and TER$_{br}$, which include segmentation tags in the computation of the overall output quality. However, their ability to single out segmentation errors was not empirically demonstrated. For this reason, in this section we perform a systematic assessment of subtitle segmentation metrics, with the aim to

[3] A comprehensive analysis of the robustness/sensitivity of standard and subtitle segmentation metrics for perfect texts can be found in (Karakanta et al., 2022c)

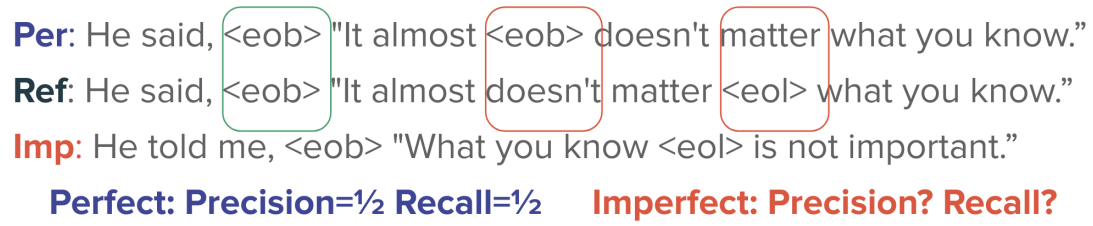


Figure 6.1: Reference-based evaluation for perfect/imperfect texts

better understand their behaviour in relation to textual errors.

Our contributions can be summarised as follows:

- An exploration of the relationship between BLEU_{br} and BLEU_{nb} and the usefulness of BLEU_{br} for evaluating segmentation for *perfect* and *imperfect* texts.
- A new score *Sigma*, derived from an estimation of an upper bound of BLEU_{br}, which isolates the segmentation signal for *imperfect* texts.
- EvalSub: A tool for computing reference-based segmentation scores for automatic subtitles.[4]

6.2.1 BLEU_{br} vs. BLEU_{nb}

In spite of its flaws, BLEU remains the go-to metric for MT research and studies on automatic subtitling. BLEU without (BLEU_{nb}) and with (BLEU_{br}) boundaries have sometimes been reported together in studies of automatic subtitling (e.g. in our offline subtitling experiment in Section 5.2 and in (Buet and Yvon, 2021)), the motivation being that BLEU_{nb} should evaluate the content, and BLEU_{br} the segmentation. Yet, the relationship between these two scores suggests that this interpretation may be oversimplistic, motivating deeper analyses.

[4]The toolkit and code to replicate the experiments is available at `https://github.com/fyvo/EvalSubtitle`.

REF	the car has just left Paris <eol> for its destination London <eob> where it will arrive next Sunday <eol> if all goes well . <eob>
HYP	the car has just left Paris <eol> for his destination : London <eob> where he arrive <eol> next Sunday , if <eol> all goes well . <eob>

Figure 6.2: Comparing n-grams counted by BLEU_{br} and/or BLEU_{nb}.

Because of considering break symbols, BLEU_{br} is in general computed on longer sequences than BLEU_{nb}, which means more n-grams to match. Since predicting the right number and type of subtitle break is generally easier than predicting the actual words, BLEU_{br} usually has a higher unigram precision. For example, for the imperfect text on Figure 6.1, unigram precision for BLEU_{nb} is 61% (8/13), while for BLEU_{br} 66% (10/15). The higher-order precision scores are in turn also impacted. This suggests that the absolute or relative difference between the two scores cannot be a proper signal of segmentation quality alone: interpreting $\text{BLEU}_{br} > \text{BLEU}_{nb}$ as a sign of good segmentation may be correct, but the intensity of this signal cannot be realistically assessed from these two measures alone.

Differences between BLEU_{br} and BLEU_{nb} result from matches obtained for the n-grams containing a segmentation tag (henceforth *n-tagrams*). An example is shown in Figure 6.2. The bigram 'left Paris' and trigram 'all goes well' are counted both by BLEU_{nb} and BLEU_{br}; the bigram 'London <eob>' and the trigram 'Paris <eol> for' are counted by BLEU_{br} but not by BLEU_{nb}; conversely the bigram 'if all' and trigram 'arrive next Sunday' are counted by BLEU_{nb}, but not by BLEU_{br}.

A first possible way to disentangle the effects of the segmentation would thus be to separately compute two scores: BLEU_{nb} and BLEU_{em},[5] where BLEU_{em} only measures precision scores with respect to n-tagrams. Unigram precision only counts segmentation tags, bigram precision counts

[5]This approach is investigated by Élise Michon (thus *em*), personal communication with the authors.

2-tagrams such as 'w <tag>' or '<tag> w' etc. However, BLEU_{em} remains highly correlated with BLEU_{nb}. This is because n-tagram matches for BLEU_{em} directly depend on the precision of $(n-1)$-grams for BLEU_{nb}. For instance, 'left Paris <eol>' is correct only because 'left Paris' is also a match. This implies that the n-gram scores define an upper bound on the possible $n+1$-tagram matches. We therefore discard BLEU_{em} and use the upper bound of the n-gram scores for computing *Sigma*, our proposed subtitle segmentation score.

6.2.2 *Sigma*: a new Subtitle Segmentation Score

In the previous section we set out the reasons why BLEU_{br} cannot disentangle the effect of translation quality from the segmentation signal for imperfect texts. To address this limitation, we propose *Sigma*, a new subtitle segmentation score, which signals the segmentation accuracy, irrespective of the textual changes of the output compared to the reference. We define *Sigma* (S) as the ratio of BLEU_{br} to its upper bound, denoted $\text{BLEU}_{br}{}^{+}$:

$$S = \frac{\text{BLEU}_{br}}{\text{BLEU}_{br}{}^{+}} * 100 \tag{6.1}$$

Values close to 100 should signal a good segmentation, while values close to 0 a bad segmentation, *irrespective of the value of* BLEU_{nb}.

$\text{BLEU}_{br}{}^{+}$, the upper bound of BLEU_{br}, can be understood as the best BLEU_{br} we would obtain for a given imperfect text if all subtitle boundaries were correctly placed. In other words, $\text{BLEU}_{br}{}^{+}$ is a proxy to the best achievable BLEU_{br} score for a given BLEU_{nb}. It can be computed from BLEU_{nb}, by approximating how much the n-gram precisions would increase from the presence of correctly placed segmentation tags. More specifically, we denote:

- p_1, p_2, p_3, p_4 respectively as the unigram, bigram, trigram and fourgram-

modified precisions computed from BLEU_{nb},

- α the ratio of the number of segmentation boundaries to the number of word tokens l, and
- $p'_1, \ldots, p'_4$ the corresponding n-gram precisions for BLEU_{br}

Under the assumption that boundaries are correct, the expected number of correct unigrams in a text of l words augmented with boundaries is just $p_1 \times l + \alpha \times l$, yielding:

$$p'_1 = \frac{p_1 + \alpha}{1 + \alpha} \tag{6.2}$$

For higher order n-grams, the exact computation is more involved, but a simple upper bound is the following:

$$p'_n \leq \frac{(1 - (n-1)\alpha) \times p_n + n\alpha \times p_{n-1}}{1 + \alpha}. \tag{6.3}$$

This holds because we assume that: (a) each boundary is part of n n-tagrams, within which it is surrounded by regular tokens;[6] and (b) the text is sufficiently long so that we can make the approximation $\frac{l}{l+1} \approx 1$. When *Sigma* is computed at corpus level, the second assumption should always hold.

6.2.3 Experimental setting

To explore the behaviour of BLEU_{br} for evaluating segmentation and the added value of *Sigma* in case of imperfect texts, we perform changes to the reference text and/or segmentation by controlling the degree of noise inserted. This artificial setting allows us to control the quality of the text and segmentation in a given output. We conduct three experiments:

[6] In theory there might be several breaks in one n-tagram, but this would imply very short lines and rarely appears in our references.

Ref	He said, <eob>"It almost doesn't matter <eol> what you know."	
Ex1	He said, <eob>"It almost doesn't matter what you <eol> know."	shift.2.50
Ex2	He **X**, <eob>"It _ doesn't matter <eol> what **X** you know."	mixed.20
Ex3	He **X**, <eob>"It _ doesn't matter what **X** you <eol> know."	mixed.20.50

Table 6.1: Examples of the noise inserted in each of the three experiments.

In **Experiment 1**, we focus on evaluating the segmentation for *perfect texts* with BLEU$_{br}$. To investigate the sensitivity/robustness of BLEU$_{br}$ to segmentation errors, we perform changes in the reference segmentation in a controlled way. Specifically, we apply the following operations randomly on the reference segmentation: 1) shift, where a boundary is shifted 1, 2 or 3 positions to its left/right, 2) addition, where a new boundary is added to the segment, 3) deletion, where a boundary is deleted from the segment, and 4) replacement, where a boundary is substituted with the other boundary type, e.g. <eol> substituted with <eob>. For each operation type, we gradually increase the percentage of boundaries affected by the operation (20%, 40%, 60% 80% and 100%). For example, shift.1.20 corresponds to 20% of the reference boundaries shifted by one position, while delete.80 means that 80% of the reference boundaries are deleted. Additions are made with respect to the number of boundaries and not to the number of possible insertion positions (spaces); that is, add.100 doubles the number of boundaries. Finally, BLEU$_{br}$ is computed between the modified test set and the true reference.

In **Experiment 2**, we move to BLEU$_{br}$ for evaluating segmentation for *imperfect texts*. We first explore the relation between BLEU$_{br}$ and BLEU$_{nb}$ for imperfect system outputs when the segmentation tags remain untouched. To simulate these outputs, we insert noise in the reference text, *without affecting* the type or position of boundaries. The noising process consists of applying a mix of edit operations (insertions, deletions,

substitutions, in equal shares) corresponding to a certain percentage of the number of tokens (from 0 to 90, with a step of 10).

In **Experiment 3**, we move to the case when textual errors in imperfect texts are combined with segmentation errors. We apply segmentation changes to the references generated in Experiment 2, where the text was changed. We perform a mix of operations on the boundaries, following the same procedure as for the words above. We then compare the behaviour of *Sigma* compared to BLEU_{br} for different values of BLEU_{nb}.

Lastly, to overcome the fact that standard metrics cannot be computed on imperfect texts, we apply a boundary projection method based on reference-hypothesis alignment. The reference subtitles are aligned to the hypothesis subtitles using the MWER algorithm (Matusov et al., 2005a) and the boundaries of the hypothesis are copied in the aligned reference. Projecting the boundaries from the hypothesis to the reference allows us to compute standard segmentation metrics between the aligned reference and the true reference, as in Experiment 1. We apply the boundary projection on the outputs of the four En→Fr offline subtitling systems (Sec. 5.2).

The subtitling data used in the experiments is the English side of the English-French test set of MuST-Cinema (Karakanta et al., 2020b). BLEU is computed with SacreBLEU[7] (Post, 2018).

6.2.4 Results

In Section 4.3, we extensively used BLEU_{br} for evaluating the segmentation accuracy of segmenter models. Our first question here is whether BLEU_{br} is a good metric for evaluating segmentation in the case of perfect texts, as the ones output by our segmenters. The behaviour of BLEU_{br} to segmentation degradations (Experiment 1) is shown in Figure 6.3. BLEU_{br} globally remains within the 45–100 range, since it considers textual content. When

[7] `BLEU|#:1|c:mixed|e:no|tok:13a|s:exp|v:2.0.0`

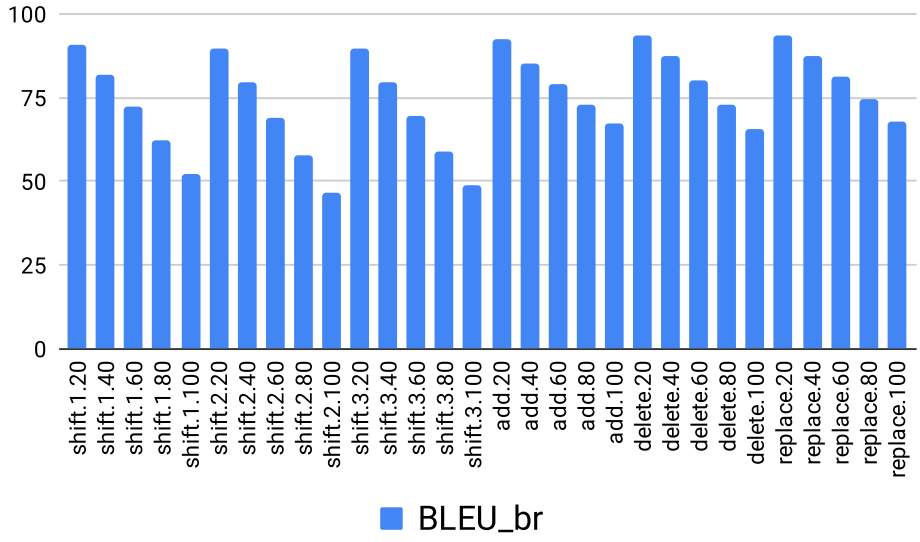


Figure 6.3: Behaviour of $BLEU_{br}$ when gradually transforming the reference segmentation with shift.1, shift.2, shift.3, add, delete, and replace operations applied on the boundaries.

looking at the first three series from the left of the graph, we observe that $BLEU_{br}$ is hardly sensitive to shift size; shifting 1, 2 or 3 positions yields almost the same scores. This shows that $BLEU_{br}$ penalises equally near and complete misses, contrary to metrics such as WindowDiff or Segmentation Similarity, which would penalise a shift of 1 position less that a shift of 3 positions. This property may be useful for other segmentation tasks (e.g. sentence segmentation), but for subtitle segmentation a shift of 1 is likely to break the semantic/syntactic unit, due to the frequent alternation between function/content words in the languages we explore. In this respect, $BLEU_{br}$ behaves similarly to F1, for which a missed position is a clear error, and justifies the high correlation between these two metrics observed in Section 4.3. However, when comparing shifts to additions, deletions and replacements (three series from the right), shifts are more penalised than the other types of noise. Despite being a precision-based metric, $BLEU_{br}$ here is actually robust to error type, since it equally penalises deletions,

additions, as well as replacements. It is worth considering that one wrongly omitted boundary will affect the n-gram precision (for $n > 1$), although not as much as one wrongly added boundary. Therefore, this balance between deletion and addition penalisation is the effect of brevity penalty, which decreases the score of segments with missing boundaries. Our findings in this experiment show that, for perfect output text ($BLEU_{nb}$=100), $BLEU_{br}$ indeed ranks segmentation from good to bad. $BLEU_{br}$ has the added benefit of equally penalising over-segmentation and under-segmentation (due to the balance between additions and deletions), as well as the prediction of the wrong type of break (as shown by the replacements), which is the main limitation of metrics not considering multiple boundaries (e.g. F1, pK, WindowDiff).

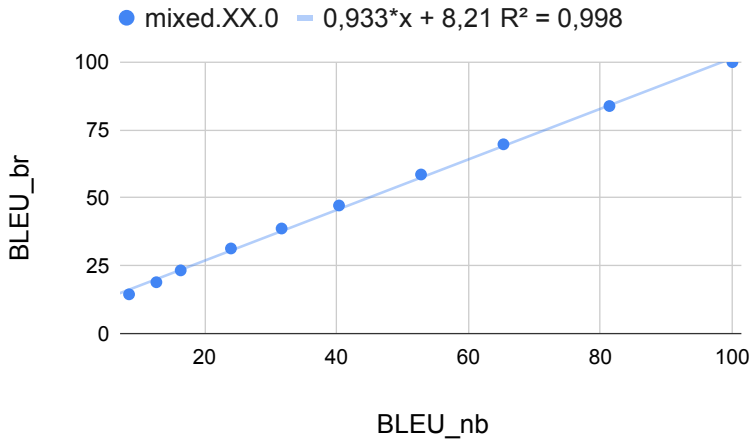
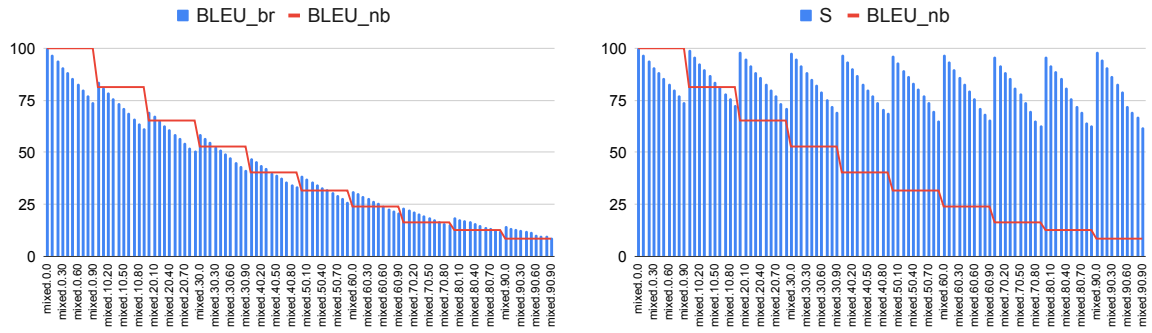


Figure 6.4: Linearity of $BLEU_{br}$ wrt $BLEU_{nb}$, for instances where only text was submitted to noise (Experiment 2). Linear regression gives a coefficient of determination of 0.998, and a standard error of 1.2.

The question we now turn to is the usefulness of $BLEU_{br}$ *for imperfect texts*. We would like to independently evaluate quality in content prediction (with $BLEU_{nb}$) and accuracy in segmentation prediction. Can $BLEU_{br}$, as used in previous work, or *Sigma* (equation 6.1), play that role? Firstly, for 'perfect' segmentation but imperfect text (Experiment 2), Figure 6.4 shows



(a) BLEU_{br} (b) *Sigma*

Figure 6.5: Values of BLEU_{br} (a) and *Sigma* (b) after applying segmentation noise at different values of BLEU_{nb}.

that the relationship between BLEU_{br} and BLEU_{nb} is linear. This indeed confirms our hypothesis that the two metrics are so correlated that their difference cannot be a strong signal of segmentation quality. It also shows that BLEU_{br} cannot exceed an upper bound which is strongly related to BLEU_{nb}, since in this setting the segmentation has not been affected by noise. Given this dependency between BLEU_{br} and BLEU_{nb}, reporting both scores is uninformative and the segmentation signal should be sought in a different relationship.

Finally, we move to the scenario of imperfect text and imperfect segmentation (Experiment 3). Adding noise to the textual content of the references, we first obtain 10 BLEU_{nb} values from 100 down to 8.5, corresponding to mixed.0-mixed.90 in Figure 6.5. Then, for each BLEU_{nb} value obtained, we subject the text to 10 levels of segmentation noise, corresponding to mixed.X.0-mixed.X.90. Figure 6.5a shows the behaviour of BLEU_{br} (blue bars) to the controlled noising process in relation to the translation quality of BLEU_{nb} (red line). We observe that BLEU_{br} decreases as translation quality drops. This does not allow comparing the segmentation accuracy of systems with a different translation quality, e.g. a system scoring 40 BLEU_{nb} (mixed.40) with that of a system with a

BLEU_{nb} of 30 (mixed.50), since BLEU_{br} will always be higher for the first system. In addition, the more BLEU_{nb} drops, the more often BLEU_{br} exceeds BLEU_{nb}. For the same level of segmentation noise (e.g. mixed.X.50), the difference $\text{BLEU}_{br} - \text{BLEU}_{nb}$ would be positive for mixed.40.50, while 0 for mixed.50.50. This empirically confirms our hypothesis that the mere difference between the two cannot be a clear segmentation signal.

On the contrary, *Sigma* demonstrates a different behaviour, as shown in Figure 6.5b where we plot *Sigma* for various values of BLEU_{nb}. *Sigma* remains stable irrespective of the value of BLEU_{nb}, in a similar range, and responds linearly to the amount of segmentation noise. Because of these properties, *Sigma* allows for evaluating the segmentation between systems of different translation quality. We observe a minor drift of the range of values when BLEU_{nb} decreases (from [74.3–100] for mixed.0 to [63.2-95.8] for mixed.90). However, in a realistic scenario, BLEU_{nb} would typically be constrained to an interval between 25 and 55 (corresponding to the 4 central series on Figure 6.5b). As a result, when comparing the segmentation accuracy of two systems close in BLEU_{nb} value, the impact of that drift would be all the more limited. For example, for the same level of segmentation noise, the difference in *Sigma* between two systems 10 BLEU_{nb} points apart, (mixed.40.50 and mixed.50.50) is about 0.4 points. This shows that *Sigma* is a good approximate value for capturing segmentation quality, irrespective of the quality of the generated content.

Table 6.2 shows the scores for the segmentation metrics previously used in this thesis. Firstly, all metrics agree that the NMT system has the highest segmentation quality. When comparing the scores for the three Speech Translation systems (`Cas`, $\texttt{e2e}_{base}$ and $\texttt{e2e}_{pt}$), *Sigma* agrees with most metrics on a better segmentation for the cascade. It is worth noticing that BLEU_{br} computed between the system output and the original reference (Column 7) does not agree with the rest of the metrics, confirming that it

System	After boundary projection					No projection		
	Prec	Rec	F1	$BLEU_{br}$	TER_{br}	$BLEU_{br}$	TER_{br}	*S*
NMT	.711	.735	.723	83.18	6.87	32.16	19.38	89.2
Cas	<u>.639</u>	.667	<u>.653</u>	<u>76.14</u>	<u>8.91</u>	26.34	<u>23.23</u>	<u>83.1</u>
$e2e_{base}$	.601	.667	.632	75.00	9.29	22.53	24.48	81.8
$e2e_{pt}$	.610	<u>.702</u>	<u>.653</u>	74.89	9.24	<u>26.36</u>	23.52	81.5

Table 6.2: Segmentation scores of *Sigma* and the examined metrics for the output of four systems after projecting their boundaries to the reference. The two columns to the left of *Sigma* are $BLEU_{br}$ and TER_{br} scores between the output and the reference without projection. Best score among the ST outputs is underlined.

cannot disentangle the effect of translation from segmentation. However, $BLEU_{br}$ computed between the aligned reference and the original reference (Column 5) shows again that $BLEU_{br}$ can be used as a segmentation metric which considers both break types in the case of perfect texts. *Sigma* has a good agreement with the standard segmentation metrics computed after alignment but does not require the additional alignment step.

To sum up, our analysis using artificial noise in segmentation has shown that for perfect texts, $BLEU_{br}$ satisfies our criteria for a good subtitle segmentation metric. However, when moving to imperfect texts, $BLEU_{br}$ correlates highly with regular BLEU, therefore the segmentation signal cannot be extracted from a simple difference between $BLEU_{br}$ and $BLEU_{nb}$. We thus introduce a new subtitle segmentation score *Sigma*, as the ratio of $BLEU_{br}$ to its approximated upper bound. *Sigma* does not suffer from the confound of text quality as $BLEU_{br}$ and TER_{br} do. Another benefit is that *Sigma* does not require any alignment between reference and hypothesis (e.g. as with SubER (Wilken et al., 2022)), which may introduce errors, especially for low quality texts. We believe that the final response on the most accurate method to evaluate subtitle segmentation can be given only after obtaining correlations with human judgements. The analysis

presented in this section has shed light into the critical aspects of evaluating subtitle segmentation, in order to better design user studies to collect these human judgements and to refine the approximation of the upper bound $\text{BLEU}_{br}{}^{+}$ for computing *Sigma* in subtitling tasks.

6.3 Auto-spotting in the spotlight

With the increasing technologisation of subtitling to respond to the rising demand for multilingual subtitles, subtitling tools are being tailored to post-editing. These solutions include not only automatic transcription and translation, but also automatic prediction of timestamps (auto-spotting). However, previous research in PE for subtitling has mainly focused on assessing quality and productivity when post-editing machine translated subtitles for which the timestamps are already provided (Volk et al., 2010; Etchegoyhen et al., 2014; Matusov et al., 2019). Therefore, the effect of auto-spotting on the process of post-editing automatic subtitles (PEAS) has remained unexplored. In this section, we explore auto-spotting through a product- and process-based analysis of post-editing automatic subtitles (PEAS). Our research questions are as follows: What is the relationship between translation quality, auto-spotting accuracy and the time spent editing a subtitle? What is the contribution of editing the auto-spotting on the temporal and technical effort in PEAS?

6.3.1 Post-editing corpus

To study the questions mentioned above, we collected subtitling post-editing data in a real use-case scenario where professional subtitlers edit automatically generated subtitles. The data collection was supported through an EAMT grant under the programme "2021 Sponsorship of Activities - Students' edition". Three professional subtitlers, two for English→Italian,

one for English→German, with experience in subtitling and MTPE post-edited the 9 TED talks from the MuST-Cinema test set. The PE task was performed in MateSub, a novel subtitling tool for post-editing fully automatic subtitles (see Section 2.2.4). The subtitlers were advanced users of the tool. We prepared *ad hoc* guidelines where we highlighted all the specific characteristics of the task.[8] The post-editing task was split in 4 sessions of a total duration of 3 hours, each one containing 3 sub-tasks. The average duration of each video for a sub-task is 4 minutes. In order to avoid fatigue effects, the subtitlers took a break of at least 15 minutes between each sub-task. Furthermore, to balance learning effects, the order of the talks was reversed for the second Italian subtitler. Original and post-edited subtitles and productivity data (time-to-edit) were collected in process logs implemented in MateSub. Keystrokes were logged with InputLog (Leijten and Van Waes, 2013). At the end of the task the subtitlers completed a questionnaire to collect 1) data related to their experience with MT and subtitling, 2) feedback on their user experience with automatic subtitling using the User Experience Questionnaire of (Koponen et al., 2020a), and 3) particular problems faced and general impressions on automatic subtitling.

The structure of the collected data is shown in Figure 6.6. For each of the language pairs (en→it/de), we provide the presegmented *.wav* files of the MuST-Cinema release. The *srt* directory contains the subtitle files (.srt) per talk for the system output (sys) and each post-editor (pe1, pe2). The *txt* directory contains aligned .txt files consisting of the source transcription (src), the target translation (ref) and the .yaml files from MuST-Cinema, the subtitling system output (sys) and the PE versions by each subtitler (sub1, sub2). The sentence alignment of the system subtitles and the PE versions with the MuST-Cinema reference was performed manually to ensure maximum quality. The project logs (*log* directory) and the

[8]The complete guidelines given to subtitlers are available at: `https://bit.ly/34XvHFi`

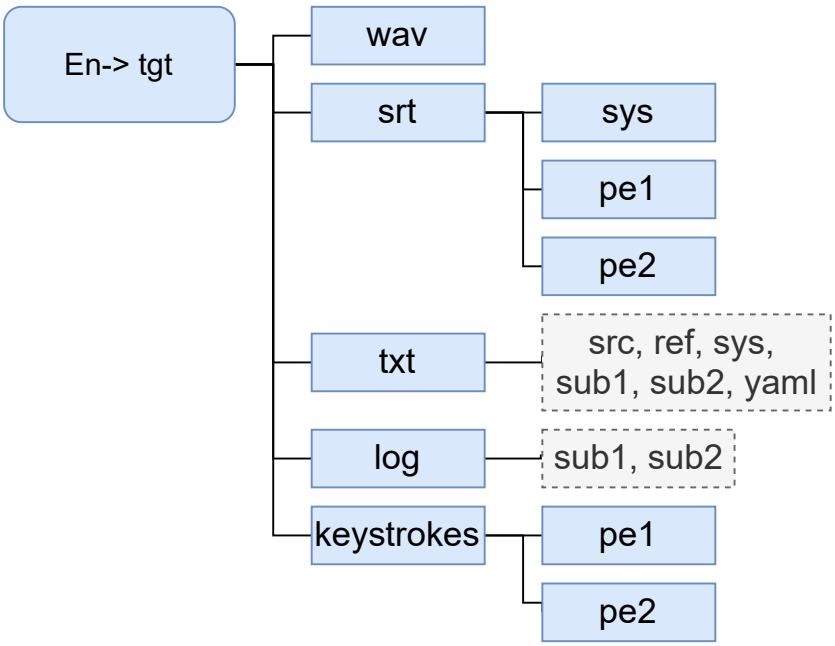


Figure 6.6: Structure of the collected data.

keystrokes are released as .csv files. For En→It, we collected 1,199 subtitles from subtitler1 and 1,208 subtitles from subtitler2, while for En→De 1,198 subtitles were obtained. The data will be publicly released along with MuST-Cinema v.2 to promote research in automatic subtitling.

6.3.2 Experimental setting

Methodology In the first research question, we explore relationships between auto-spotting accuracy, translation quality and time spent on each subtitle. Based on the process logs, we compute the correlation between:

- **Auto-spotting accuracy (dSpot)**: Total difference between the original timing of an automatic subtitle and the final timing of the post-edited subtitle (in seconds), computed as:

$$dSpot = abs(start_{orig} - start_{final}) + abs(end_{original} - end_{final}) \quad (6.4)$$

- **Edit distance (ED)**: Changes in the translation of a subtitle are

computed in terms of mean edit distance (Levenshtein) between the original and the post-edited subtitle text. Even though HTER (Snover et al., 2006) correlates better with human judgements at corpus level (Fischer and Läubli, 2020), our analysis requires edit estimation at subtitle level. HTER values are computed by dividing with the reference length, which in our subtitles can be very short or even zero when an automatic subtitle is deleted. For this reason we select ED.

- **Post-editing speed (PES)**: Speed in words per minute. Minutes are obtained from the Time-To-Edit (TTE) variable, defined as the cumulative time spent on each subtitle. TTE counts the total time a subtitler remains "active" on a subtitle regardless of whether they perform an operation (type, drag, play) or not.

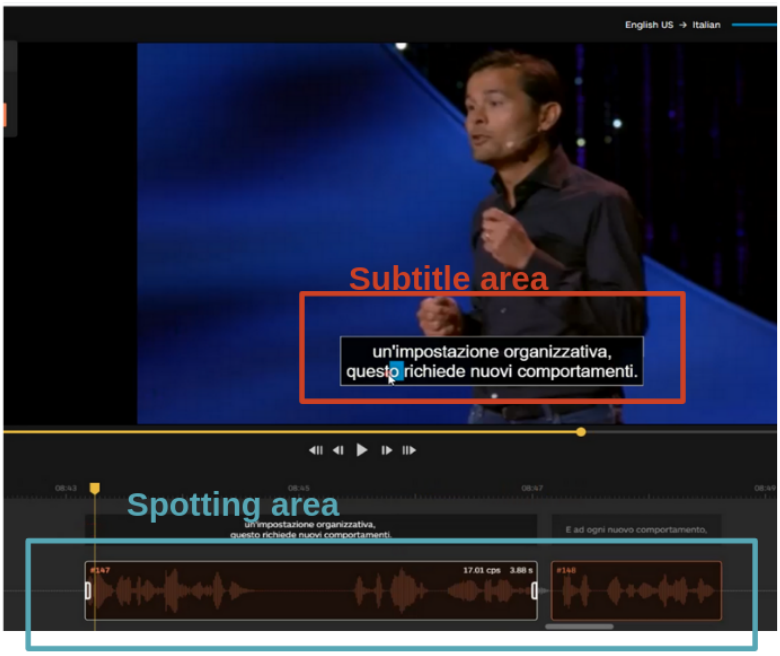


Figure 6.7: Areas of interest for selecting the number of mouse clicks performed in each.

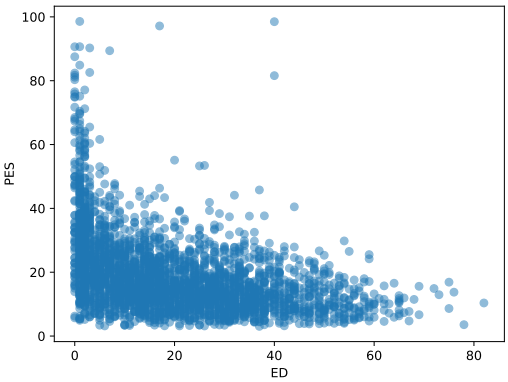
To compute the correlation we remove subtitles with a TTE>400,000 milliseconds (6.6 minutes). We noticed in the screen recordings that these

cases are mostly subtitles at the beginning and end of the task, when the Matesub window was left unintentionally open. Moreover, we had to leave out new subtitles (subtitles inserted by the subtitlers) since they do not have an original timing to compute dSpot. This process removed 176 subtitles for subtitler1, 217 for subtitler2 and 289 for the German subtitler.

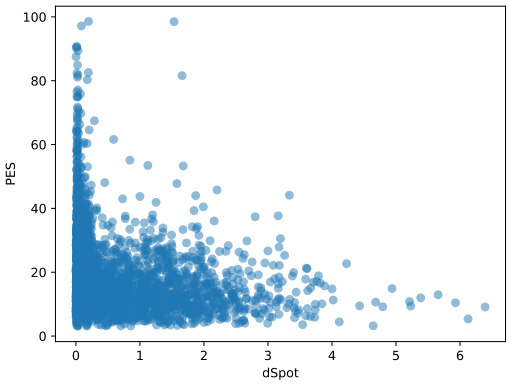
Since dSpot and ED are product-based metrics, we further conduct a process-based analysis of the keystrokes of auto-spotting/text edits. The subtitlers adjust the auto-spotting by clicking and dragging subtitle boxes in the spotting area at the bottom of the Matesub screen. To isolate the spotting edits from other operations, we select mouse clicks inside the spotting area, based on x,y coordinates. We consider as text edits all keyboard operations, and mouse clicks inside the subtitle area. Figure 6.7 shows the two areas. We then compare the total number of interaction events and total event duration between the spotting and text groups.

6.3.3 Results

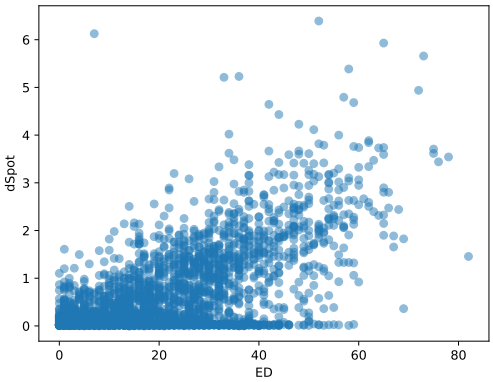
For our first research question, which aims at investigating the relationship between translation quality, auto-spotting accuracy and temporal effort, we plot the scatterplots for the selected variables (Figures 6.8). Figure 6.8a shows the relationship between PES and edit distance. As expected, PES decreases as the amount of translation editing increases. However, the amount of text editing is not (inversely) proportional to the speed. For "perfect" subtitles (ED=0) there is a large variability in PES, after which there is an exponential drop in speed. This is in line with the findings of Krings (2001) for MTPE, who found that the relationship between MT output quality and the temporal and cognitive PE effort is non-linear. Even more so, in PEAS, it is possible that PES is influenced by the multimodality of the source. The source text does not lie on the screen in front of the post-editor's eyes as in MTPE, but is "disguised" in the video; the subtitler



(a) Post-editing Speed and Edit Distance.



(b) Post-editing Speed and dSpot.



(c) Edit Distance and dSpot.

Figure 6.8: Scatterplots for the selected variables.

has to play the video for the duration of the utterance in order to access the source. This adds up to the TTE, affecting in turn the speed. For PES and dSpot (Figure 6.8b), we observe a similar exponential drop. The more the spotting changes compared to the predicted timestamps, the more time is needed to edit the subtitle. However, the effect of auto-spotting changes seems to flatten between $0.5 - 2$ seconds. For text vs spotting changes (Figure 6.8c), the relationship is more linear: the more the text changes compared to the system output, the more the spotting changes too. We expect that this is due to the change in length between the original

and the PE subtitle text; when the text expands, the duration of the subtitle needs to be increased accordingly to accommodate reading speed constraints, and vice versa. We also observe a solid horizontal line for ED when dSpot=0. This means that text edits can be performed without adjusting the spotting, for example when the translation is edited but its length remains equal.

Figure 6.9 shows the Pearson correlation matrix for Levenshtein ED, dSpot and PES. We observe a moderate negative correlation between PES-ED (-0.45) and a weak negative correlation between PES-dSpot (-0.26). As seen in the scatterplots, the weak correlation suggests that auto-spotting inaccuracies are less related to increased temporal effort, while translation quality seems to be a more decisive factor, at least for the product-based analysis. dSpot is a product metric which may be useful for evaluating the auto-spotting (in)accuracy but does not necessarily reflect the actual spotting effort. A strong positive correlation is found between ED-dSpot (0.64), confirming that changes in the translation will likely require adjustments of spotting, and vice versa.

We then move to the second research question to explore the contribution of spotting edits on the PE effort. We compare the technical and temporal effort of post-editing the text vs. the auto-spotting in terms of keystrokes for the three subtitlers (Table 6.3). Looking at the total number of interactions between spotting and text, interactions are between 4 to 10 times more for text edits for all subtitlers. However, when looking at the duration of interactions between spotting and text, the difference between spotting and text decreases. For it1, 134 minutes were spent on spotting edits and 199 on text edits, while the difference is minimal for it2 (5 minutes more for text edits). On the contrary, the German subtitler spent twice the time on text edits compared to spotting edits (177 vs 85 minutes). Despite this variability among subtitlers and language pairs, spotting edits take

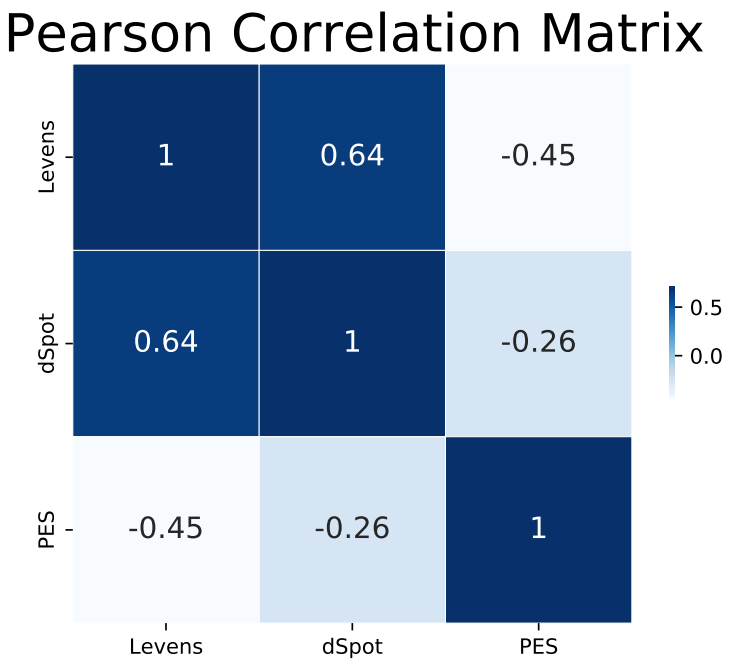


Figure 6.9: Correlation between HTER, TTE and dSpot.

up a significant time of the total PE process. Even though the number of interactions is smaller, spotting edits take on average longer, since they require holding the mouse clicks to adjust the timestamps. This effect of spotting edits on time is not reflected in the correlation analysis of PES-dSpot. Observing the PE process with the aid of the screen recordings, we noticed that subtitlers often expand the subtitle box to fit the text but then re-adjust the spotting closer to the initial start-end times after they have made additional changes to the translation. This will not lead to large changes in the final timestamps and thus to dSpot, but requires more mouse clicks. Moreover, adjusting the timing by a few milliseconds requires high precision and longer clicks, while larger, coarse adjustments could be performed with one short click. dSpot, as a product metric, does not capture the multiple attempts to adjust the timing in the way that the process-based keystroke analysis does. Therefore, the relation between

Subtitler	**area**	**Total interactions**	**Duration**
it1	Spotting	10459	134.77
	Text	44029	199.08
	subtitle	15855	153.21
	keyboard	28174	45.87
it2	Spotting	8516	116.93
	Text	54329	121.26
	subtitle	7269	57.13
	keyboard	47060	64.13
de	Spotting	5391	85.01
	Text	55595	177.55
	subtitle	11552	128.91
	keyboard	44043	48.64

Table 6.3: Total number of interaction events and total event duration in minutes for text and spotting edits. Text edits is the sum of *subtitle* and *keyboard*.

the system spotting accuracy and the actual effort in adjusting spotting should be further investigated.

We acknowledge the limitations of this analysis considering that the data has been collected from two post-editors in one language pair and one post-editor in another. Considerable variation between different individuals' process data has been observed in many studies. With such a low number of participants, care must therefore be taken when drawing conclusions from the results. Still, our analysis suggests that editing the auto-spotting contributes to the technical and temporal effort in PEAS and it is an important factor which should be considered both in evaluation of automatic subtitling and in subtitling process research. In addition, product-based analysis may be useful for system diagnostics, but the picture is incomplete without exploring the correlation of automatic metrics also through a process-based analysis.

6.4 User experience and subtitlers' views

In the last part of this chapter, we focus on participant-based methods for evaluating automatic subtitling. These methods are rarely employed in MT research since they are time-consuming and harder to design and analyse (Karakanta, 2022). However, they are an important tool for obtaining insights directly from experts and end users on aspects which cannot be captured by product- and process-based studies. Even though translators are fundamental for the advance of new technologies, their views are often not sufficiently considered (Guerberof-Arenas, 2013). The study of subtitlers' perceptions of the technology they interact with can be beneficial for all stakeholders in the AVT industry. The inclusion of subtitlers in the process of technological change can alleviate their resistance to adopting technologies (Cadwell et al., 2018). Developers can direct their implementation efforts in the right direction to provide user-friendly tools and interfaces (Moorkens and O'Brien, 2017), and AVT trainers can identify necessary skills for teaching and training (Bolaños-García-Escribano et al., 2021a). A better understanding of subtitlers' interaction with technology can help define the rising profession of the subtitler post-editor (Bywood et al., 2017), and establish metrics and standards to protect subtitlers against dropping rates and ensure fairness (Nikolic and Bywood, 2021).

In response to the challenges brought about by increasing technologisation, in this section we conduct a survey of subtitlers' perspectives on the developing paradigm of automatic subtitling. This survey is a timely and relevant contribution to take stock of this nascent technology and its implementation in the subtitling profession from the very beginning, while setting the stage for further developments. The survey focuses on the subtitlers' user experience when post-editing automatically-generated subtitles from and into different Western European languages (en→fr,es,pl,nl; it→fr;

es→en,it,de). It also aims at collecting feedback on the main issues and benefits of the technology, as well as on the impact of automatic subtitling on the subtitler's profession. Based on qualitative and quantitative analysis of a survey questionnaire, we provide a participant-based evaluation of automatic subtitling and a comprehensive view of subtitlers' attitudes towards this new paradigm. Our findings indicate that despite its current limitations mainly related to challenges in audio and speech processing, automatic subtitling has potential and its benefits are already recognised by the users. Based on the received criticisms, we provide a list of recommendations for future improvements in automatic subtitling tools, which we hope will serve as a guide for technology developers. We further release the questionnaire and responses to foster replication and reproducibility in automatic subtitling.[9]

6.4.1 Methodology

The survey described in this paper was conducted in December 2021 and consisted in respondents filling in a questionnaire after having taken part in testing sessions of an automatic subtitling tool.

The task In the PE task, subtitlers were required to post-edit the automatic subtitles of 8 video clips. The clips were self-contained excerpts from different TV series (drama), each around 3 minutes long, amounting to a total duration of 30 minutes. TV series were selected as the material to post-edit since they are representative examples of real subtitling tasks. In addition, series contain elements which are particularly challenging both for human subtitlers and automatic systems compared to TED talks, such as background noise, slang, overlapping speech and multi-speaker events. The change of subtitling material from TED talks to series in this section

[9]https://github.com/fatalinha/subtitlers-have-a-say

was determined by the fact that the task was part of an evaluation session for a commercial tool. The company kindly shared the questionnaire with the subtitlers, however we do not have any access to the output or post-edits. The original language of the series was English. Since all subtitlers edited the same clips but not all of them worked with English as source language, we used the dubbed version for subtitlers working from Spanish and Italian.

The task was performed over two consecutive days and the subtitlers took sufficient breaks between each video to avoid fatigue effects. The subtitlers worked from their personal office without any explicit time limit. The automatic subtitling tool selected for the task is Matesub (see more details in Section 2.2.4). The subtitles to be post-edited were created through the automatic subtitling system integrated in Matesub. Before starting the task, all participants, regardless of their previous experience with the subtitling tool, were asked to familiarise themselves with it by watching a video tutorial, in which the functionalities of the tool were explained. They were also asked to revise the Netflix timed-text guidelines.[10] This setting resulted in a homogeneous task for all participants, with a sufficient duration to develop reliable judgements and a clear view on their user experience.

Respondents The respondents were professional subtitlers who were recruited through a language service provider (Translated.com). Participation to the survey was voluntary and the responses were collected anonymously. Before starting the survey, participants were informed about the objective of the research, the purposes of the data collection and gave their consent. In total, 22 out of 24 subtitlers responded to the questionnaire

[10] https://partnerhelp.netflixstudios.com/hc/en-us/articles/360051554394-Timed-Text-Style-Guide-Subtitle-Timing-Guidelines

Language pair	Subtitlers
Spanish → English	2
Spanish → Italian	3
Spanish → German	3
Italian → French	3
English → French	2
English → Spanish	3
English → Polish	3
English → Dutch	3

Table 6.4: Respondents per language pair.

(91% response rate). The subtitlers worked in different language pairs. Table 6.4 shows the number of subtitlers for each language pair. Subtitlers worked in from-English, into-English, but also non-English language pairs, which are often disregarded in MT research (Fan et al., 2021). The focus of the survey is to obtain a broad overview of subtitlers' opinions on automatic subtitling, regardless of the language-specific performance of the technology. Therefore we opted for selecting respondents so as to cover a wide range of language pairs.

Survey and questionnaire The questionnaire was set up as an online form containing open and closed questions. It was delivered in English for all respondents and contained three parts. The first part collected factual information about the subtitlers, such as years of experience in subtitling, years of experience in MTPE and how often they use Matesub. Three questions focused on the working settings and the diffusion of MT in subtitling jobs. These questions asked how often their subtitling jobs involved using master templates, working directly from the video, and editing machine translated subtitles.

The second part of the questionnaire focused on the respondents' user

experience with the task of PE automatically generated subtitles. We used the User Experience Questionnaire (UEQ) by Koponen et al. (2020a), a version of the UEQ of Laugwitz et al. (2008) for end-user evaluation of software products, which has been adapted to post-editing experience. This selection of questionnaire facilitates comparison of PE in automatic subtitling with the PE experience based on a different system. By using an existing questionnaire, we respond to the need for standardisation in experimental research in AVT and MT. The questionnaire contained 13 pairs of adjectives related to the post editing experience, in the form *Post-editing was... (difficult/easy, unpleasant/pleasant, stressful/relaxed, labourious/effortless, slow/fast, inefficient/efficient, boring/exciting, tedious/fun, complicated/simple, annoying/enjoyable, limiting/creative, demotivating/motivating, impractical/practical*). Since the tool features auto-spotting and automatic segmentation, we included evaluations on the quality of spotting and segmentation and the perceived effort of editing them. The responses are provided on a scale of -3 to +3, with 0 representing a neutral mid-point. As in the UEQ, average scores between -0.8 and +0.8 are considered neutral evaluations, while scores below -0.8 correspond to negative evaluations and scores above 0.8 to positive evaluations.

The last part of the questionnaire contained open questions on the quality of MT, auto-spotting and automatic segmentation, as well as the subtitlers' opinion on the benefits of automatic subtitling, whether it helps the work of subtitlers and whether they see any dangers for the profession of subtitlers from using automatic subtitling. The answers to the open questions were analysed based on thematic analysis (Braun and Clarke, 2006), a method for analysing qualitative data using codes, with the Taguette[11] software. This analysis aimed at identifying main issues with the technologies implemented in the tool (automatic translation, auto-spotting,

[11] `https://www.taguette.org/`

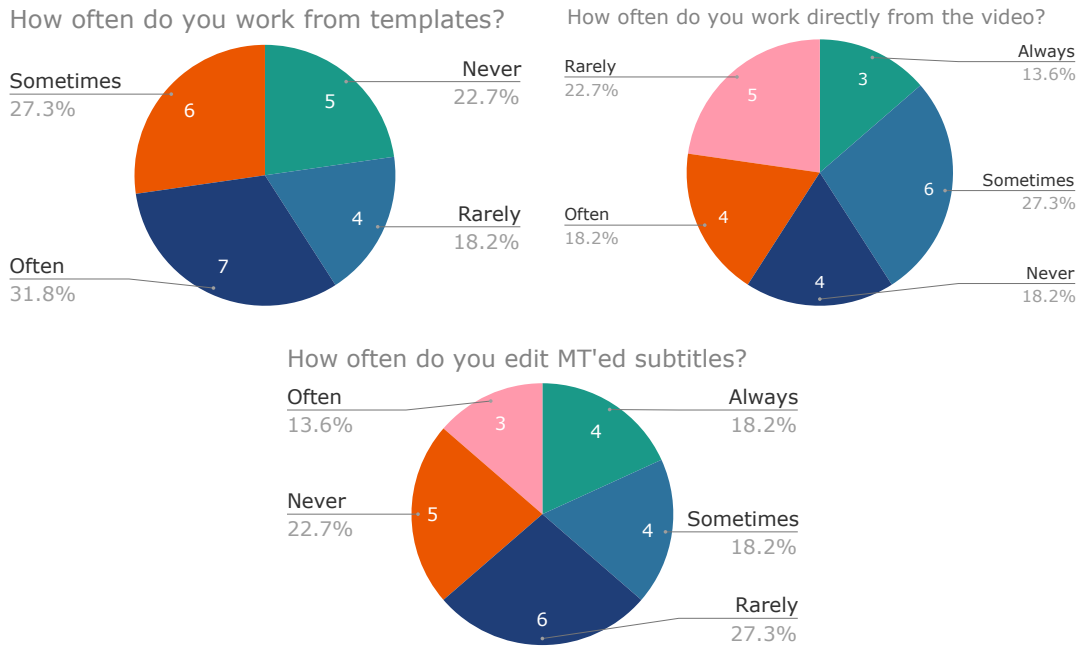


Figure 6.10: Respondents' answers to their working settings.

segmentation), as well as the main benefits from using automatic subtitling. The general opinion on usability is coded as positive, neutral/mixed or negative.

6.4.2 Results

Subtitlers' profiles and working settings The respondents had on average 2.3 years of experience as subtitlers (SD=1.5, range 1-5 years) and 2.6 years of experience with MTPE (SD=2.4, range 0-10 years). In terms of working settings, there is large variability in the way subtitling is performed. Figures 6.10 show the responses to the questions *How often do your subtitling jobs involve master templates?*, *How often do your subtitling jobs involve working directly from the video?* and *How often do your jobs involve editing machine-translated subtitles?*. Despite the variability in the professional conditions in subtitling when it comes to the use of tools, settings and requirements, the fact that 17/22 subtitlers edit automatically

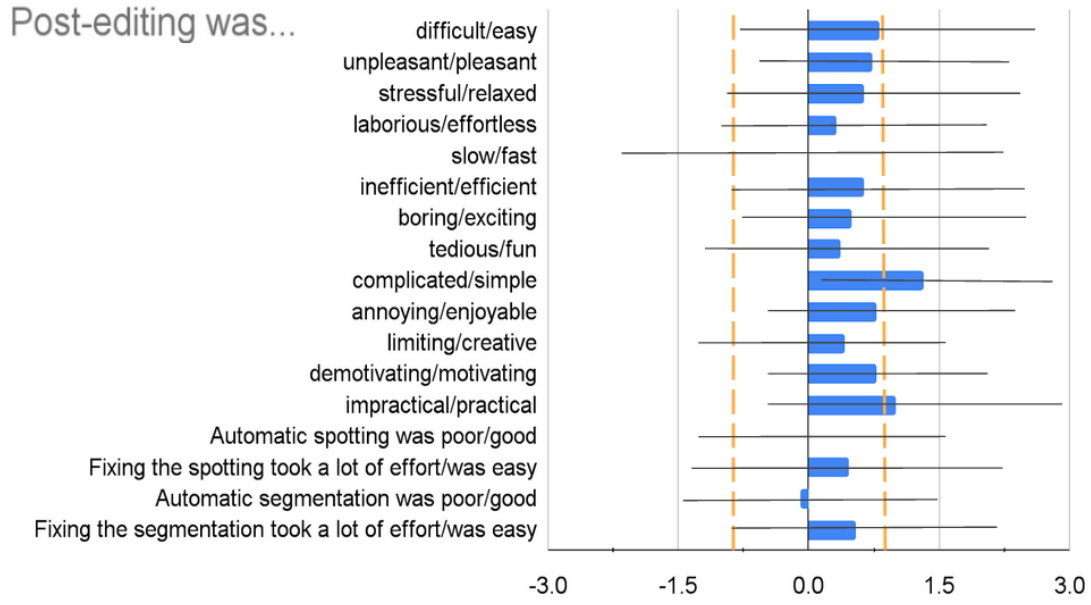


Figure 6.11: User experience (UX) scores. Interrupted vertical lines mark the -0.8/+0.8 threshold for neutral evaluations. Horizontal lines mark standard deviation.

translated subtitles shows that MT is a reality for subtitling. In addition, the responses confirm that our respondent sample covers different levels of expertise and a broad skill range.

User experience The mean scores for the user experience across subtitlers and language pairs are shown in Figure 6.11. Overall, the post-editing experience can be considered as neutral to positive, with all except one mean scores leaning on the positive side of the scale. The subtitlers found the post-editing process simple and practical. Even though still in the neutral range, the lowest scores were observed for the quality of autospotting and automatic segmentation, where mean scores are close to 0.

When comparing the UX scores with the ones obtained from the subtitlers post-editing the TED talks in Section 6.3, similar tendencies are observed (Figure 6.12). However, the scores are higher, with 10/17 judgements leaning towards the positive side. This indeed confirms that the choice of drama series is more challenging for the automatic subtitling sys-

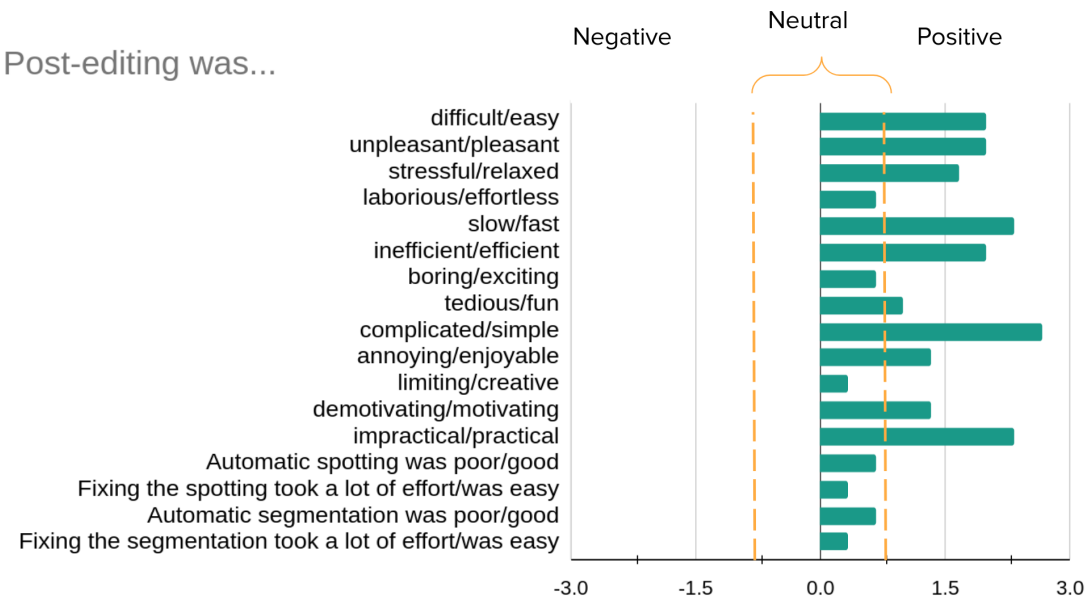


Figure 6.12: User experience scores for the subtitlers in Section 6.3 (TED talks).

tem compared to TED talks. It should however be kept in mind that the two groups have a different number of subtitlers and that the subtitlers in the previous section were much more acquainted with the tool. When comparing the scores with the study of Koponen et al. (2020a), our scores are more distributed towards the positive side, even though a direct quantitative comparison of the user experience between the two subtitling systems cannot be performed due to the differences in settings (e.g. language pairs, tool etc). It should also be noted that our sample is larger (22 respondents instead of 12) and with a larger variety in language pairs (8 compared to 4). In (Koponen et al., 2020a), the lowest average scores were found for the adjectives *laborious/effortless* and *limiting/creative*. This adjective pairs received low scores in our study too, however with *slow/fast* having the lowest score and a very large deviation. Similarly, the quality of autospotting and segmentation had lower scores than the effort to fix them. All in all, the user experience scores show that PE in automatic subtitling is found acceptable by the subtitlers and pointed out particular limitations,

Automatic translation		Autospotting		Segmentation	
Speech/audio recognition errors	10	Inaccurate (starting to early, too late)	10	Oversegmentation (too may short subtitles)	6
Lexical, punctuation, case	7	False negatives (no subtitle when speech)	5	No respect of syntactic/semantic units	5
Missing context, inconsistencies	5	False positives (subtitle when no speech)	3	No respect of constraints and guidelines	4
		Not respecting visual elements (shot changes)	2	Undersegmentation (too long subtitles)	3
Worked well	3	Worked well	6	Worked well	5

Table 6.5: Main issues related to automatic translation, autospotting and segmentation, and number of statements.

mainly related to the technical aspects of spotting and segmentation.

Subtitlers' feedback

Main issues with automatic subtitling Table 6.5 shows the main issues for automatic translation, auto-spotting and segmentation, as identified based on the thematic analysis of the subtitlers' responses to the open questions. For automatic translation, speech recognition errors seem to be the most common reason for errors in the translation (10 statements). Subtitlers mentioned that translation quality was highly influenced by the speaker's accent, audio quality and the speed of speech. For example, they mentioned that *muffled or fast speech, music and background noises can often confuse the AI.* Speech recognition errors have indeed been identified as the main issue for speech translation systems, regardless of whether they are direct or cascaded architectures (Bentivogli et al., 2021). The second group contained lexical errors, such as the translation of slang, idioms, colloquial expressions, figurative language and named entities, and in some cases, casing and punctuation (7 statements), with subtitlers reporting that

automatic translation still tends to be a bit too literal. Translations out of context or words translated individually or inconsistently across the video were also mentioned as common issues (5 statements). A subtitler noted that inconsistent translation suggestions by the system *may lead the human translator to lose consistency as well*. Three subtitlers thought translation worked well.

For autospotting, lack of accuracy was the main reported issue (10 statements), since subtitlers thought that subtitles often started too early or too late and were not properly synchronised with the speaker. False negatives (no subtitle created when there is speech) and false positives (subtitles created when there is no speech) were also reported in 5 and 3 statements respectively. All these factors are related to common speech recognition issues, for example when speech is not recognised due to bad audio quality or when background noise is recognised as speech. Some subtitlers (2 statements) mentioned that automatically-spotted subtitles did not respect shot changes and other visual elements. Six subtitlers reported that autospotting worked pretty well or did not report any issues.

For automatic segmentation, oversegmentation (unnecessarily segmenting subtitles into small pieces) and undersegmentation (failing to segment too long subtitles) were mentioned in 6 and 3 statements respectively. Other issues were that the segmentation did not respect the norms of the target language because of splitting semantic/syntactic units (5 statements), and that segmentation resulted in subtitles not respecting the guidelines and length/reading speed constraints. Five subtitlers affirmed that automatic segmentation worked well.

Main benefits of automatic subtitling When asked about the main benefits of automatic subtitling, *speed* was considered the main benefit by almost all subtitlers (18/22). Surprisingly, this is in contrast with the low mean

score for slow/fast in the UX questionnaire. When looking into the benefits reported by subtitlers who rated the PE experience as slow (negative values for *slow/fast*), all of them mentioned that it saves time, but only on the creation of subtitle boxes and setting the timestamps. This shows the importance of not relying only on quantitative scores in participant-based studies, but complementing the judgements with qualitative explanations. Additionally, *efficiency* was noted as a benefit in 10 statements and *reduction of effort* related to technical aspects in 6 statements. Specifically, subtitlers reported that automatic subtitling *saves a lot of tedious work*, *creates a guideline of what needs to be translated instead of watching the whole video* and *serves as a starting template*, which, as a result, *allows focusing more on the translation* rather than having to spend time on technical aspects. The provision of *useful suggestions* was mentioned in 2 statements, related to subtitling solutions that the subtitler had not considered or to terminology and vocabulary.

General impressions for the subtitling profession To the question whether they think that automatic subtitling helps the work of subtitlers, 14 subtitlers responded positively, 5 gave neutral/mixed statements and 3 claimed that in most cases automatic subtitling does not help. The subtitlers who responded neutrally mentioned as concerns that the quality depends on the language, audio quality, and that it may be useful only for some applications (e.g. *template* creation, *other audiovisual products, such as online conferences or courses, documentaries*).

When asked whether they see any possible danger to the profession because of automatic subtitling, 5 subtitlers mentioned they see no dangers at all, 8 subtitlers saw no dangers for the time being, given the current state of the technology and its low diffusion, while 9 subtitlers identified some type of danger. Possible dangers were the loss in the quality of the

final subtitles (4), dropping rates (2) and having less or no work if clients select cheaper, automatic options (5). Another danger identified was the improper application of the technology (3 statements), where subtitlers considered that the profession is not at risk only as long as a human is involved in the final phase.

6.4.3 Discussion

This study focused on subtitlers' user experience and perspectives on the task of post-editing automatically generated subtitles. Our findings suggest a neutral to positive experience. Even though there are those who still see no benefits from this new technology, automatic subtitling was welcomed with enthusiasm by many subtitlers, as an aid to save time and effort. As with studies on MTPE experience (Guerberof-Arenas, 2013; Bundgaard, 2017), subtitlers have expressed disfavour towards automatic subtitling with respect to technological flaws, but also acknowledged its positive aspects and expected technology to shape their profession in the near future. As for the dangers to the profession, most criticisms were not rooted in the fear of being outperformed by automatic systems, but rather in the effect of technology on the final product and market consequences (Vieira, 2020). The positive aspects of technology can only be appreciated when combined with respectful and ethical professional and market practices.

Previous work reporting feedback of subtitlers focused on a setting where MT was applied to human-generated subtitles. The views of the subtitlers involved did not lead to auspicious conclusions in favour of the use of MT in subtitling. In spite of encouraging automatic evaluation scores, subtitlers were cautious in reporting productivity gains in (Volk et al., 2010), while in (Etchegoyhen et al., 2014) PE experience was rated as rather negative (2.37 on a 1-5 scale), with MT being useful only for sim-

ple and short sentences. An increase in productivity for simple sentences was reported in (Matusov et al., 2019), where the two subtitlers rated their experience as fair. In (Koponen et al., 2020a) the participants did not find PE particularly difficult but characterised it as negative or limiting and did not think MTPE increased productivity. Similar criticisms were reported for MT quality in our study, with MT described as too literal, unable to properly translate spoken and figurative language. However, most subtitlers acknowledged that automatic subtitling makes their work faster and more efficient, especially when compared to *old-style subtitling*. The difference of our study compared to studies of MT for subtitling is the automatisation not only of the translation, but also of the technical aspects of spotting and segmentation. Integrating and evaluating these aspects is vital for the establishment of the new paradigm of automatic subtitling, as we showed in Section 6.3. Subtitlers recognised the importance of automatising these aspects, which are often characterised as tiresome and dull. By not focusing only on the translation but the automation of the technical aspects, automatic subtitling allows subtitlers to spare time and effort on the tedious part of the work (spotting and segmentation) and unleash their creativity in adjusting the final text.

Our study aimed to provide a broad view of subtitlers' perspectives, by complementing quantitative scores with open questions, attempting to cover several language pairs and a range of subtitler profiles. However, we acknowledge that the findings should be interpreted with some caution. Questionnaire-based studies have a context-bound nature and may be affected by factors such as the system (quality, language), the participants (age, familiarity with technology) and the setting (Tuominen, 2018). Some limitations should thus be considered when drawing conclusions.

Firstly, responses and user experience scores may have been affected by the language pair, due to differences in the subtitling quality depend-

ing on the ASR and MT performance, despite keeping all other settings (videos, instructions) equal. Still, we opted for not reporting results separately for each language pair, since the sample size per pair (2-3) would be too small to draw robust and generalizable conclusions on a per-language basis. Second, even though we attempted to include a broad range of professional subtitler profiles, the group is not necessarily representative of the subtitlers' general population. For example, the respondents' age, a variable not collected in our survey, may affect their technological acceptance. Moreover, their experience in subtitling, template translation and MTPE varies. We found in statistical tests that the only variable affecting the user experience is MTPE experience. Subtitlers with less experience (<= 2 years) had significantly higher user experience scores than the more experienced ones.[12] It is possible that experts, already being used to a certain level of MT output quality and to their preferred interfaces, are less willing to change tasks and tools, while novices, having less consolidated working practices, are more open and less critical against new interfaces and workflows. Accepting to take part in a task involving automatic subtitling already means the subtitlers were willing, curious or even familiar with the technology, and therefore may have been positively inclined towards automatisation in subtitling, contrary to many AVT professionals (Audiovisual Translators Europe, 2021).

Lastly, the interface used in PE has a great influence on user experience. We selected Matesub as a typical instance of an automatic subtitling tool. However, the generalisability to other tools is not guaranteed. In an attempt to test whether previous experience with Matesub had an effect on user experience, we separated the respondents in two groups based on their responses to the question *How often do you use Matesub in your subtitling*

[12]Novices (N=14, M=1.0, SD=0.7) vs Experts (N=8, M=−0.4, SD=1.1). Based on an equal-variance independent samples t-test: $(t(20) = 3.82, p = .001)$

jobs: regular users (often, sometimes) and occasional (never, rarely). We found that familiarity with the tool did not have an effect on the average user experience scores.[13] This shows that the tool is user-friendly, with a steep learning curve, and does not require extensive training. Less user-friendly tools may negatively affect the post-editing experience. Despite these limitations, this study presents a screenshot of the current state of the quickly evolving technology, necessary to drive implementation efforts in the right direction.

6.4.4 Recommendations for improvement

Our findings have identified some limitations of current automatic subtitling systems. Based on the subtitlers' feedback, we present a list of suggestions for improving automatic subtitling tools in a direction that benefits the user experience. The suggestions are listed in order of priority.

• **Improving autospotting and segmentation**. The main benefit of automatic subtitling according to the subtitlers was eliminating tedious work and leaving more space for creativity. Given that many criticisms were addressed to the quality of autospotting and segmentation, improvements in the automation of technical aspects are a priority. Except for improving the accuracy of auto-spotting through enhanced audio processing and a more syntactically-informed segmentation, interaction with these elements could become more user-friendly. For example, it could be useful to implement interactive features such as automatic adjustment of subtitle boxes to match length and reading speed constraints after subtitlers translate or finish editing one subtitle.

• **Improved audio pre-processing**. Most problems in the translation, autospotting and segmentation stemmed from the segmentation of the audio. This is an open problem in speech processing (Gaido et al.,

[13]Regular (N=14, M=0.6, SD=1.2) vs Occasional (N=8, M=0.4, SD=0.9). ($t(20) = 0.42, p = .679$)

2021b; Tsiamas et al., 2022); audio segmentation is typically approached by breaking the audio on speaker silences, considered as a proxy of clause boundaries, and not on syntactic information. A syntax-unaware segmentation is responsible for translations out of context and the issues in segmentation (over-undersegmentation, no respect of syntactic units). In addition, the reported cases of false positives/false negatives in autospotting (see Table 6.5) indicate that voice activity detection technologies should be improved to properly distinguish speech from noise.

• **Improving in-video consistency**. Consistency of MT suggestions is important for easily spotting errors and for avoiding repetitive corrections. Consistency can be improved through adaptive MT (Biçici and Yuret, 2011; Denkowski et al., 2014; Turchi et al., 2017) or document-level MT (Lopes et al., 2020).[14] Another direction could be the integration of external resources, such as termbases and translation memories. These aids have passed the test of time and are the first requirement of users before overshooting with MT solutions (Audiovisual Translators Europe, 2021).

• **User experience vs. automatic metrics**. Punctuation and casing was reported as an issue for automatic translation. However, WER, the metric used to evaluate ASR systems, is normally computed in a case/punctuation insensitive way. Casing and punctuation cannot be derived directly from the audio and therefore these errors are traditionally considered as less relevant by the scientific community. On the contrary, in the context of automatic subtitling they must be weighed appropriately. This points out the need for task-specific evaluation metrics, which take into account elements that shape user experience.

• **Incorporation of elements from the visual modality**. Since subtitling is multimodal and intersemiotic, ignoring elements from the visual

[14]However, it should be noted that (Koponen et al., 2020b) found no preference for document-level MT compared to sentence-level MT in subtitling.

modality can result in errors. Some features from the visual modality are already integrated in many (non-MT) tools, e.g. marking of shot changes. Another useful feature could be the recognition of on-screen text.

In this section we presented findings on subtitlers' user experience and perspectives when post-editing automatically generated subtitles, based on a survey questionnaire. Subtitlers' experience was marked as neutral to positive. Thematic analysis of the open questions showed that the main issues of automatic subtitling stem from failures in speech recognition and pre-processing, which result in error propagation, translations out of context, inaccuracies in auto-spotting and suboptimal segmentation. However, subtitlers acknowledge the positive sides of the technology, which are speed and effort reduction, especially related to the technical aspects, as well as the provision of useful suggestions. We conclude that, despite current limitations, automatic subtitling tools can be beneficial for subtitlers, as long as improvements consider subtitlers' opinions, and ethical and professional standards are respected. We expect that as automatic subtitling tools mushroom, larger studies will be needed to explore different variables and monitor the progress in automatic subtitling.

6.5 Conclusions

This chapter has dealt with one of the most challenging issue in automatic subtitling: evaluation. I have argued that the main challenges coming from the new paradigm of automatic subtitling are the variability in the number of segments, the need to evaluate the additional requirements of spotting and segmentation and the complexity of the subtitling process. For segmentation, standard segmentation metrics (e.g. precision, recall, F1) cannot be applied in the case of imperfect texts, when the textual content does not match that of the reference. To address this lacuna, we proposed

Sigma, a new subtitle segmentation score, which evaluates segmentation irrespective of text quality. We have shown that *Sigma* is not affected by text quality, as in the case of BLEU_{br} and TER_{br}, but is also not susceptible to alignment errors (Section 6.2). Then, we turned our attention to auto-spotting and its effect on the post-editing process (Section 6.3). We collected product and process-based subtitle post-editing data from professional subtitlers, which allowed us to analyse the relationship between translation quality, auto-spotting accuracy and time spent to edit a subtitle. We found that PE speed decreases more for translation edits than for auto-spotting edits. Still, spotting edits are bound to the text edits, especially when the text expands or shrinks. Our analysis of keystrokes showed that, even though most edits are related to text, spotting edits take up a significant part of the total PE process. We recommend complementing product-based analyses of PE output with process-based explorations of effort when devising new metrics for automatic subtitling. Finally, we gave voice to the subtitlers, the actual users of automatic subtitling, as a tool to improve productivity (Section 6.4). Our questionnaire survey involving 22 subtitlers in 8 language pairs aimed at exploring their user experience and at collecting detailed feedback on the technology and market-related aspects. A quantitative and qualitative analysis of the responses showed that automatic subtitling is perceived in a neutral to positive way by subtitlers, who found speed and effort reduction, especially when it comes to the technical aspects to be the main benefit of the technology. The main issues identified stemmed from audio (pre-)processing errors which had repercussions on the quality of translation, spotting and segmentation. Based on the analysis of the feedback, we drafted a list of recommendations for improving future automatic subtitling tools. We recognise that the solutions proposed in this chapter may not have led to the ultimate automatic metric or experimental design for evaluating automatic subtitling, but we hope

that they have shed some light in the darkness of evaluation uncertainty and will set the ball rolling for future developments in evaluating automatic subtitling as an extrinsic task.

Chapter 7

Conclusion

This chapter concludes the thesis, by placing it in the context of the current audiovisual translation, machine translation and especially speech translation research. After summarising the results, I discuss the new directions and challenges opened by the thesis, before closing with my view on the future of automation in AVT.

7.1 A new paradigm

When I started my PhD research in early 2019, a lot of the technological and societal developments occurred until the time of writing this thesis could not have been foreseen. Audiovisual Translation was already a vibrant and rapidly expanding field, dealing with the challenges posed by the boom in audiovisual production. Machine translation had well transitioned from statistical to neural, with NMT being enthroned both in academia and industry. NMT approaches overcame the initial instabilities and had by then become robust and more efficient. Speech translation had seen equally significant improvements, even though still dominated by cascades. Only in 2019 did direct systems start to catch up with cascades (Niehues et al., 2019), with the difference between the two approaches reduced to a couple of BLEU points.

When it comes to subtitling, research in MT for subtitling had been stagnant. With the exception of a few small-scale studies, the only projects attempting to automatise subtitling in the 10's were the Stockholm MT System for Film Subtitles (Volk et al., 2010) and the EU-funded SUbtitling by MAchine Translation (SUMAT) (Etchegoyhen et al., 2014) projects, stemming from the SMT era. They both reported productivity gains for some languages but recognised that the ground was not ready for the establishment of MT in the subtitling world. Unfortunately, the data used to train the systems in both projects were to a large degree proprietary, so they were not released to the research community, which impeded further studies on MT for subtitling.

Except for the notably behemothian OpenSubtitles project, there existed only few subtitling corpora. Still, even the existing, publicly available parallel corpora of subtitles were compiled with text translation in mind. For this reason, audio (let alone video) was not contained in any of these corpora. In addition, timestamps and line breaks were lost in the compilation process. As for evaluation, studies on MT for subtitling had focused only on coarse productivity (time) gains, while quality was measured with the usual MT evaluation metrics: BLEU and TER. There was not one single metric for evaluating subtitle-specific aspects.

Inspired by the breakthroughs in ST and the increased performance and efficiency of NMT, I envisioned a new paradigm, that of automatic subtitling. This paradigm would move away from MT for subtitling to integrated approaches which decrease the human labour involved in creating a timed transcript of the source speech, while at the same time respecting the form and specific constraints of subtitling. Intuitively, automatic subtitling, incorporating the speech modality through ST, is in line with the multimodal nature of subtitling. End-to-end modelling of the formal requirements of the target medium is an attractive prospect both for ex-

perimenting with new NMT methods and for offering integrated solutions. This thesis is a collection of my contribution towards the establishment of this new paradigm in three main areas: data, models and evaluation. The results can be summarised as follows:

- Existing subtitling corpora suffered from two main limitations for building automatic subtitling systems: the lack of the source audio and the loss of line breaks in the target. The MuST-Cinema corpus we have released is the first speech-to-subtitles corpus overcoming the limitations of previous subtitling corpora. MuST-Cinema allowed for experimenting with various subtitling tasks, from audio-informed segmentation to ST for subtitling. Future attempts to create corpora for automatic subtitling should follow a similar path to promote further advances.

- Quality is more important than quantity in automatic subtitle segmentation. We have shown that even a small amount of high-quality human segmentation decisions can lead to systems which successfully segment sentences into readable subtitles. Confirming our intuition on the significance of multimodality, audio-informed segmentation approaches led to further gains in segmentation accuracy even for languages unseen during training, in a zero-shot fashion.

- Speech translation is a viable alternative to the subtitling template. We found that the two dominant approaches, cascade and direct, perform relatively on par when it comes to quality. As we have seen in the case of segmentation, multimodality through direct ST models seems to benefit the modelling of subtitling constraints also in translation. In addition, models that allow linking the caption-subtitle generation process lead to quality gains without compromising significantly the conformity to the subtitling constraints. For live subtitling, Simultaneous ST systems may still perform poorly in terms of translation quality, but have the benefit

of shorter delay compared to human subtitlers. This can be exploited for experimenting with alternative visualisation modes, leading to a more comfortable reading speed for the users.

- The evaluation of subtitling is a complex process comprising multiple elements, one of them being form-related parameters. We introduced simple metrics for formal aspects such as length and reading speed conformity and a new score for evaluating the quality of segmentation, applicable to end-to-end subtitling systems. Even though their success needs to be verified against human judgements, these scores have revealed that there are multiple layers of quality in subtitling which should be considered in system evaluation along with translation quality.

- As argued in the background section on AVT, automatic subtitling comprises *practices, processes and products* and thus its study requires employing all categories of AVT models. To this aim, we collected product and process data from subtitlers post-editing automatic subtitles. We provided a rich corpus for exploring and testing automatic subtitling metrics, additionally extending the MuST-Cinema test set with multiple references. From the study of the post-editing data, we found that spotting has an effect both on quality and productivity and it is closely bound to the quality and edits of the textual subtitle content. As for the study of practices, we have given voice to the users of automatic subtitling tools. Our survey showed that subtitlers have a relatively positive attitude towards automatic subtitling as long as professional and ethical matters are respected. Future improvements of automatic subtitling tools should focus on improving speech (pre-)processing and on automatising the technical aspects of subtitling as well as on the interaction of subtitlers with the technical features, leaving them the freedom to focus on the most creative part of their work.

- Our review of previous approaches to automatising subtitling highlighted that the complexity of subtitling can lead to a methodological labyrinth when it comes to experimental research. From participants' profiles to training settings and the granularity of the translation unit (subtitle, sentence, document), reproducibility and replicability are not guaranteed. We have drafted a reproducibility checklist for research in automatic subtitling, which we hope will help the community design and reproduce experiments. We strongly encourage participant-based evaluations and their publication, through releasing the data, surveys and responses.

We are now witnessing the dawn of this new paradigm. At the moment of concluding this thesis, research in automatic subtitling is burgeoning, in AVT as well in ST. On the one hand, more researchers in AVT are realising that automation is a reality and are conducting user studies to test its applicability. On the other hand, the MT-ST community is becoming more aware of the particularities of subtitling compared to a verbatim translation of a speech transcript. This year, IWSLT hosted a new shared task on isometric translation, relevant for subtitling and dubbing, and is considering a shared task on subtitling for its 2023 edition. I am glad that our contributions presented in this thesis have helped raise this bilateral awareness and hope that these findings have set the stone rolling for future developments.

7.2 Open challenges

The contributions mentioned in this thesis have set the floor for the rising paradigm of automatic subtitling. Now that the 'salad days' of automatic subtitling have passed, it is time to consider how our research efforts can lead to a mature field. As with every new research direction, automatic subtitling has unleashed a hydra of new possibilities and challenges. Although

not exhaustive, I here outline a number of research directions opened up to further improve this paradigm.

How can we improve data collection? As discussed previously, representative data for the task of automatic subtitling can be scarce, since subtitling data often come from proprietary sources and are protected by copyright. Even though the compilation of training data is of equal importance, a realistic priority when it comes to subtitling data is obtaining reliable tests and benchmarks. While MuST-Cinema is a significant contribution towards data for automatic subtitling, the test set is small, it covers only the domain of TED talks and does not contain multi-speaker events. Similar to ASR and NMT where the current tendency is to move from controlled environments "into the wild" (Arivazhagan et al., 2019a; Rio et al., 2021), there is more often the need to test automatic subtitling systems in different domains, transitioning from subtitling lectures to meetings, entertainment productions etc. A publicly available, multimodal, diverse test set covering a range of subtitling phenomena could be a starting point for more reproducible research in automatic subtitling both from the point of view of MT and AVT studies. Its creation, validation and management would create collaboration opportunities among MT researchers, professional subtitlers and AVT scholars.

How do the findings generalise to other AV genres and domains? Following directly from the question on subtitling data collection, another question arising from this thesis is the generalisability of the findings to other AV genres and types of content. Limiting the scope of this thesis to TED material and guidelines was a one-way street because of the license, availability and readiness of TED data. However, it should be noted that audiovisual content is not a single uniform genre and can in fact take many different forms and genres, each their own specificities (Burchardt et al., 2016). Other genres e.g. series, news broadcasts, political debates,

instructional videos will likely have different requirements both in terms of translation, style and formal constraints, and may pose different challenges for the automatic subtitling systems. For example, automatically subtitling multi-speaker videos will require speaker diarisation while text condensation is relevant for some subtitling genres. This does not diminish the value of the methods and findings presented in this thesis, however they should serve as a starting point for experimenting with automatic subtitling techniques under several AV content types and domains.

How can we get to fully end-to-end approaches? The models proposed in this thesis generated text segmented in subtitles directly from the audio. In this way, they tackled the segmentation requirement but not the spotting. This is already implemented in some cascade solutions, where timestamps are retrieved from the ASR component. Automatically predicting the timestamps using direct approaches is still an open issue. In addition, the extension of multimodality to include information from the visual mode could be another breakthrough in automatic subtitling, leading to fully comprehensive multimodal approaches that take the most from the video. Hopefully, drawing from this thesis, these research directions will be actively explored in future work on automatic subtitling.

How can we increase customisation? There is not one correct format for subtitles. Rather, subtitling guidelines differ based on the type of AV content, user group and preferences. This requires systems to adapt their output based on particular variables, such as maximum length, genre, degree of condensation, simplification (Buet and Yvon, 2021; Liu et al., 2020a; Lakew et al., 2019; Varis and Bojar, 2021; Lakew et al., 2022). Incorporating such constraints in an efficient way, possibly within the same model, is a research direction that will definitely flourish in the future.

How do we best evaluate automatic subtitling systems? Given the complexity of quality in subtitling, there could be two directions to evaluate

automatic subtitling: one is to separately evaluate how well each constraint/aspect is matched, then possibly derive an aggregate score; the other is to try to reproduce human references. In this thesis we followed the first approach, by separately evaluating each aspect, which also allowed for a more in-depth analysis of the characteristics of the proposed models. However, this approach is not a desideratum in practical applications were system goodness should be judged with one score. Recently, SubER (Wilken et al., 2022) was proposed, which considers translation quality and (indirectly) timing and segmentation with a single score. This metric showed relatively higher correlations with human judgements compared to other proposed subtitle metrics. Except for working on metrics which obtain higher correlations, another line of research is improving the design of protocols for collecting human judgements. PE-based evaluation can be used to consider productivity gains but perception judgements from non-translators may show a different picture (Szarkowska et al., 2020). Even the design of evaluations to collect human judgements is an involved process, given the plethora of variables present. For this reason, studies in automatic subtitling can shed light into the critical aspects, in order to better design user studies to collect these human judgements.

7.3 The future of AVT

The needs for AVT are exploding day by day. To keep up with the booming needs for accessibility, (partially) automatising AVT is inevitable. Subtitling is the leading AVT type in the process of automatisation, with projects ranging from platforms providing fully automatic subtitles to tools to support users and professionals in creating better subtitles, faster. With these technological advancements, machines are steadily entering among the agents creating subtitles, along with professionals, amateurs, volun-

teers and fansubbers. This inevitably leads to the question: Will the future of AVT be more human- or machine-translated? And if automatically-translated content prevails in AVT, how do we ensure that the repercussions for viewers, researchers, developers and subtitlers will be positive?

There is a number of steps that can be undertaken to accompany automation with responsible and fair application. Firstly, to ensure that automation really benefits viewers, it is important to increase MT literacy and to inform them about the quality they can expect and the risks related to automatic subtitling. User-centred approaches can help understand the effect of the disruptive technology on their perceptions and needs. This understanding can then drive developers towards solutions really tailored to the users' needs and not only towards 'fancy' solutions. For the researchers involved in automatic subtitling, creating bridges between the AVT, MT and ST communities can greatly help steer research questions in the right direction. As for subtitlers, automatic subtitling should not substitute but complement. For example, automatic subtitles could be aimed to different products (e.g. user-generated content) and different audiences. 'Augmented' subtitling with the support of user-friendly tools can empower subtitlers and increase quality along with productivity. All these steps can ensure that the benefits of automatic subtitling can reach all involved agents. Communication and access to information is already audiovisual. Let our technological approaches make communication and information more inclusive, fair and accessible, no matter the linguistic or sensory needs.

Bibliography

Pablo Daniel Aguero, Jordi Adell, and Antonio Bonafonte. 2006. Prosody generation in the speech-to-speech translation framework. In *Acoustics, Speech and Signal Processing, ICASSP 2006*, volume 1. IEEE International Conference.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Ma-*

chine Translation, pages 1–88, Online. Association for Computational Linguistics.

Ashkan Alinejad, Maryam Siahbani, and Anoop Sarkar. 2018. Prediction improves simultaneous neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3022–3027, Brussels, Belgium. Association for Computational Linguistics.

Tanel Alumäe and Mikko Kurimo. 2010. Efficient estimation of maximum entropy language models with n-gram features: an SRILM extension. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, pages 1820–1823, Makuhari, Chiba, Japan.

Aitor Álvarez, 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.

Aitor Álvarez, Marina Balenciaga, Arantza del Pozo, Haritz Arzelus, Anna Matamala, and Carlos-D. Martínez-Hinarejos. 2016. Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3049–3053, Portorož, Slovenia. European Language Resources Association (ELRA).

Aitor Álvarez, 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.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Antonios Anastasopoulos and David Chiang. 2018a. Tied Multitask Learning for Neural Speech Translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 82–91, New Orleans, US-LA.

Antonios Anastasopoulos and David Chiang. 2018b. Tied multitask learning for neural speech 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 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

Ebrahim Ansari, 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. Association for Computational Linguistics.

Rachele Antonini and Delia Chiaro. 2005. The perception of subtitled humour in italy: An empirical study. *Cross-Cultural Encounters: Linguistic Perspectives*, pages 33–44.

Rachelle Antonini and Delia Chiaro. 2009. The perception of dubbing by italian audiences. In Jorge Díaz Cintas and G. Anderman, editors, *Audiovisual Translation: Language Transfer on Screen*, pages 97–114. Basingstoke: Palgrave Macmillan.

Gopala Krishna Anumanchipalli, Luís C. Oliveira, and Alan W. Black. 2012. Intent transfer in speech-to-speech machine translation. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 153–158.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu.

2019a. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019b. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. Association for Computational Linguistics.

Stephen Armstrong, Colm Caffrey, and Marian Flanagan. 2006. Translating DVD Subtitles from English-German and English-Japanese Using Example-Based Machine Translation. In *MuTra 2006—Audiovisual Translation Scenarios: Conference Proceedings*.

Association des traducteurs-adaptateurs de l'audiovisuel ATAA. n.d. Subtitling and Dubbing for Film and Television - A guide to foreign-language adaptation in France. https://beta.ataa.fr/documents/Guide_EN.pdf.

Audiovisual Translators Europe. 2021. AVTE Machine Translation Manifesto. https://avteurope.eu/wp-content/uploads/2021/09/Machine-Translation-Manifesto_ENG.pdf. Last accessed: 06/04/2022.

Wilker Aziz, Sheila Castilho Monteiro de Sousa, and Lucia Specia. 2012. Cross-lingual Sentence Compression for Subtitles. In *16th Annual Conference of the European Association for Machine Translation*, EAMT, pages 103–110, Trento, Italy.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019a. A Comparative Study on End-to-end Speech to Text Translation. In *Proceedings of the International Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 792–799, Sentosa, Singapore.

Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian Herold. 2020. Start-before-end and end-to-end: Neural speech translation by AppTek and RWTH Aachen University. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 44–54, Online. Association for Computational Linguistics.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019b. On Using SpecAugment for End-to-End Speech Translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, US.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Chikiu 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.

Eduard Bartoll. 2004. Parameters for the classi[a]cation of subtitles. In Pilar Orero, editor, *Topics in Audiovisual Translation*, Benjamins Translation Library. John Benjamins, Amsterdam/Philadelphia.

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.

Rocío Baños-Piñero and Frederic Chaume. 2009. Prefabricated orality a challenge in audiovisual translation. In G. Nadiani M. Giorgio Marrano and C. Rundle, editors, *inTRAlinea Special Issue: The Translation of Dialects in Multimedia*. inTRAlinea. Last accessed: June 2020.

BBC. 2021. Subtitle guidelines. Last accessed: 28 April 2022.

Daniel Beck, Trevor Cohn, and Gholamreza Haffari. 2019. Neural speech translation using lattice transformations and graph networks. In *Proceedings of the EMNLP Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 26–31, Hong Kong.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus di-

rect speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-End Automatic Speech Translation of Audiobooks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228, Calgary, Canada.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *Proceedings of the NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain.

Nicola Bertoldi and Marcello Federico. 2005. A new decoder for spoken language translation based on confusion networks. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 86–91, San Juan, Puerto Rico. IEEE.

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 272–283, Edinburgh, Scotland. Association for Computational Linguistics.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92—100.

Łukasz Bogucki. 2013. *Areas and methods of audiovisual translation research.* Peter Lang, Frankfurt am Main.

Łukasz Bogucki and Jorge Díaz-Cintas. 2020. An excursus on audiovisual translation. In Łukasz Bogucki and Mikołaj Deckert, editors, *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, pages 11–32. Palgrave Studies in Translating and Interpreting.

Łukasz Bogucki. 2009. Amateur subtitling on the internet. In Jorge Díaz Cintas and Gunilla Anderman G, editors, *Audiovisual Translation*, pages 49–57. Palgrave Macmillan, London.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Dominik Macháček, Sangeet Sagar, Otakar Smrž, Jonáš Kratochvíl, Peter Polák, Ebrahim Ansari, Mohammad Mahmoudi, Rishu Kumar, Dario Franceschini, Chiara Canton, Ivan Simonini, Thai-Son Nguyen, Felix Schneider, Sebastian Stüker, Alex Waibel, Barry Haddow, Rico Sennrich, and Philip Williams. 2021. ELITR multilingual live subtitling: Demo and strategy. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 271–277, Online. Association for Computational Linguistics.

Alejandro Bolaños-García-Escribano and Jorge Díaz-Cintas. 2020. The cloud turn in audiovisual translation. In Łukasz Bogucki and Mikołaj Deckert, editors, *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, Palgrave Studies in Translating and Interpreting, pages 519–544. Cham: Palgrave Macmillan.

Alejandro Bolaños-García-Escribano, Jorge Díaz-Cintas, and Serenella Massidda. 2021a. Latest advancements in audiovisual translation education. *The Interpreter and Translator Trainer*, 15(1):1–12.

Alejandro Bolaños-García-Escribano, Jorge Díaz-Cintas, and Serenella Massidda. 2021b. Subtitlers on the cloud: The use of professional web-based systems in subtitling practice and training. *Revista Tradumàtica. Tecnologies de la Traducció*, 19:1–21.

Petar Božović, 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.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

François Buet and François Yvon. 2021. Vers la production automatique de sous-titres adaptés à l'affichage (towards automatic adapted monolingual captioning). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 91–104, Lille, France. ATALA.

Kristine Bundgaard. 2017. Translator Attitudes towards Translator-Computer Interaction - Findings from a Workplace Study. *HERMES - Journal of Language and Communication in Business*, 56:125–144.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Aljoscha Burchardt, Arle Lommel, Lindsay Bywood, Kim Harris, and Maja

Popović. 2016. Machine translation quality in an audiovisual context. *Target*, 28(2):206–221.

Lindsay Bywood, Panayota Georgakopoulou, and Thierry Etchegoyhen. 2017. Embracing the threat: machine translation as a solution for subtitling. *Perspectives*, 25(3):492–508.

Patrick Cadwell, Sharon O'Brien, and Carlos S. C. Teixeira. 2018. Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, 26(3):301–321.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine*

Translation, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Cristian Cardellino. 2019. Spanish Billion Words Corpus and Embeddings. Last accessed: June 2020.

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

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

Wallace Chafe. 1985. Linguistic differences produced by differences between speaking and writing. In David Olson, Nancy Torrance, and Angela Hildyard, editors, *Literacy, Language, and Learning: The Nature and Consequences of Reading and Writing*, pages 105–22. Cambridge University Press.

Frederic Chaume. 2000. *La traducción audiovisual: estudio descriptivo y modelo de análisis de los textos audiovisuales para su traducción*. Ph.D. thesis, Universitat Jaume I de Castellón.

Frederic Chaume. 2012. *Audiovisual Translation: Dubbing*. Translation practices explained. St. Jerome Pub.

Frederic Chaume-Varela. 2006. Dubbing. In Keith Brown, editor, *Encyclopedia of Language & Linguistics*, Second edition, pages 6 – 9. Elsevier, Oxford.

Qiao Cheng, Meiyuan Fang, Yaqian Han, Jin Huang, and Yitao Duan. 2019. Breaking the data barrier: Towards robust speech translation

via adversarial stability training. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Colin Cherry, Naveen Arivazhagan, Dirk Padfield, and Maxim Krikun. 2021. Subtitle translation as markup translation. In *Proceedings of INTERSPEECH 2021*.

Agnieszka Chmiel, Agnieszka Szarkowska, Danijel Korzinek, Agnieszka Lijewska, Łukasz Dutka, Łukasz Brocki, and K. Marasek. 2017. Ear–voice span and pauses in intra- and interlingual respeaking: An exploratory study into temporal aspects of the respeaking process. *Applied Psycholinguistics*, 38:1201 – 1227.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *CoRR*, abs/1606.02012.

Kyunghyun Cho, B. van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation*, page 103.

Jorge Díaz Cintas and Aline Remael. 2021. *Subtitling: Concepts and Practices*. Translation practices explained. Routledge.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. Incremental decoding and training methods for simultaneous translation 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 2 (Short Papers)*, pages 493–499, New Orleans, Louisiana. Association for Computational Linguistics.

Hayley Dawson. 2019. Feasibility, quality and assessment of interlingual live subtitling: A pilot study. *Journal of Audiovisual Translation*, 2(2):36–56.

Sheila C. M. de Sousa, Wilker Aziz, and Lucia Specia. 2011. Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 97–103. Association for Computational Linguistics.

Dirk Delabastita. 1989. Translation and mass-communication: Film and tv translation as evidence of cultural dynamics. *Babel*, 35:193–218.

Michael Denkowski, Chris Dyer, and Alon Lavie. 2014. Learning from post-editing: Online model adaptation for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 395–404, Gothenburg, Sweden. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2010. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. 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 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019b. Adapting Transformer to End-to-end Spoken Language Translation. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria.

Mattia Antonino Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. 2019c. Robust neural machine translation for clean and noisy speech transcripts. In *International Workshop on Spoken Language Translation (IWSLT)*.

Jorge Diaz-Cintas. 2013. The technology turn in subtitling. In Marcel Thelen and Barbara Lewandowska-Tomaszczyk, editors, *Translation and Meaning – Part 9*, pages 119–132. Zuyd University of Applied Sciences.

Quoc Truong Do, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. 2016. Transferring emphasis in speech translation using hard-attentional neural network models. In *17th Annual Conference of the International Speech Communication Association (InterSpeech 2016)*, San Francisco, California, USA.

Quoc Truong Do, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2015. Improving translation of emphasis with pause

prediction in speech-to-speech translation systems. In *12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam.

Qian Dong, Yaoming Zhu, Mingxuan Wang, and Lei Li. 2022. Learning when to translate for streaming speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 680–694, Dublin, Ireland. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Jorge Díaz Cintas. 2001. Striving for quality in subtitling: The role of a good dialogue list. In Yves Gambier and Henrik Gottlieb, editors, *(Multi)media translation*, pages 199–-211. Benjamins, Amsterdam.

Jorge Díaz-Cintas. 2020. The Name and Nature of Subtitling. In Mikołaj Deckert Łukasz Bogucki, editor, *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, Palgrave Studies in Translating and Interpreting. Palgrave Macmillan.

Jorge Díaz-Cintas and Serenella Massidda. 2019. Technological advances in audiovisual translation. In Minako O'Hagan, editor, *The Routledge Handbook of Translation and Technology*. Routledge, London.

Jorge Díaz-Cintas and Pablo Muñoz Sánchez. 2006. Fansubs: Audiovisual translation in an amateur environment. *The Journal of Specialised Translation*, 6:37–52.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 Evaluation Campaign. In *International Workshop on Spoken Language Translation (IWSLT)*.

Johanes Effendi, Yogesh Virkar, Roberto Barra-Chicote, and Marcello Federico. 2022. Duration modeling of neural tts for automatic dubbing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8037–8041.

ESIST. European association for studies in screen translation.

Eden Estopace. 2017. Talent Crunch Hits Media Localization as Amazon, Netflix Accelerate Market Growth. https://slator.com/talent-crunch-hits-media-localization-as-amazon-netflix-accelerate-market-growth/.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine Translation for Subtitling: A Large-Scale Evaluation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 46–53.

Itamar Even-Zohar. 1979. Polysystem theory. *Poetics Today*, 1(1/2):287–310.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, pages 1–48.

Marcello Federico, Robert Enyedi, Roberto Barra-Chicote, Ritwik Giri, Umut Isik, Arvindh Krishnaswamy, and Hassan Sawaf. 2020. From speech-to-speech translation to automatic dubbing. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 257–264, Online. Association for Computational Linguistics.

María Jesús Fernández Fernández. 2009. The Translation of Swearing in the Dubbing of the Film South Park into Spanish. In Jorge Díaz Cintas, editor, *New Trends in Audiovisual Translation*, pages 210–225. Multilingual Matters, Bristol, UK.

José Luis Martí Ferriol. 2013. Subtitle reading speeds in different languages: the case of lethal weapon. *Quaderns. Revista de Traducció*, 20:201–210.

Lukas Fischer and Samuel Läubli. 2020. What's the difference between professional human and machine translation? a blind multi-language study on domain-specific MT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 215–224, Lisboa, Portugal. European Association for Machine Translation.

István Fodor. 1976. *Film Dubbing: Phonetic, Semiotic, Esthetic and Psychological Aspects*. Helmut Buske, Hamburg.

Marina Fomicheva and Lucia Specia. 2019. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558.

Chris Fournier. 2013. Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712, Sofia, Bulgaria. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021a. CTC-based compression for direct speech translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 690–696, Online. Association for Computational Linguistics.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2020a. Contextualized Translation of Automatically Segmented Speech. In *Proc. of Interspeech*, pages 1471–1475.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2020b. End-to-end speech-translation with knowledge distillation: FBK@IWSLT2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 80–88, Online. Association for Computational Linguistics.

Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2021b. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 55–62, Trento, Italy. Association for Computational Linguistics.

Yves Gambier. 2003. Introduction, screen transadaptation: Perception and reception. *The Translator*, 9:171–189.

Yves Gambier and Henrik Gottlieb. 2001. *(Multi)Media Translation: Concepts, Practices, and Research*. John Benjamins Publishing, Amsterdam.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, Online. Association for Computational Linguistics.

Boni García. 2017. Bilingual subtitles for second-language acquisition and application to engineering education as learning pills. *Computer Applications in Engineering Education*, 25(3):468–479.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

Panayota Georgakopoulou. 2019. Template files: The Holy Grail of subtitling. *Journal of Audiovisual Translation*, 2(2):137–160.

Henrik Gottlieb. 1998. Subtitling. In Mona Baker, editor, *Routledge Encyclopedia of Translation Studies*, pages 244–248. Routledge, London and New York.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and*

Interoperability with Discourse, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Alex Graves. 2013. Generating sequences with recurrent neural networks.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA.

Gian Maria Greco. 2019. Accessibility studies: Abuses, misuses and the method of poietic design. In *HCI International 2019 – Late Breaking Papers*, pages 15–27, Cham. Springer International Publishing.

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1342–1352, Doha, Qatar. Association for Computational Linguistics.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

Ana Guerberof-Arenas. 2013. What do professional translators think about post-editing? *JoSTrans - The journal of specialised translation*, 19.

Nicolas Guichon and Sinead McLornan. 2008. The effects of multimodality on l2 learners: Implications for call resource design. *System*, pages 85–93.

Steven Gutstein, Olac Fuentes, and Eric Freudenthal. 2008. Knowledge Transfer in Deep Convolutional Neural Nets. *International Journal on Artificial Intelligence Tools*, 17(03):555–567.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2020. Gaussian Error Linear Units (GELUs).

S. Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer and J. F. Kolen, editors, *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *Proc. of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. IEEE.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one

speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Sathish R. Indurthi, Houjeung Han, Nikhil K. Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2020. End-end Speech-to-Text Translation with Modality Agnostic Meta-Learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7904–7908, Barcelona, Spain.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Giménez. Adrià, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus For Speech Translation Of Parliamentary Debates. In *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233, Barcelona, Spain.

Roman Jakobson. 1959. On Linguistic Aspects of Translation. *On Translation.*

Henry Jenkins. 2006. *Convergence culture - Where Old and New Media Collide*. New York university press.

Carl Jensema. 1997. Final report for presentation rate and readability of closed captioned television. Technical report, Department of Education Technology Research CFDA 84.18OG.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184, Brighton, UK.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-based Curriculum Learning for End-to-end English-Japanese Speech Translation. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 2630–2634.

Takatomo Kano, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2013. Generalizing continuous-space translation of paralinguistic information. In *14th Annual Conference of the International Speech Communication Association (InterSpeech 2013)*, Lyon, France.

Dionysios Kapsaskis. 2011. Professional identity and training of translators in the context of globalisation: The example of subtitling. *Journal of Specialised Translation*.

Alina Karakanta. 2022. Experimental research in automatic subtitling: At the crossroads between machine translation and audiovisual translation. *Translation Spaces*.

Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022a. Post-editing in automatic subtitling: A subtitlers' perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 259–268, Ghent, Belgium. European Association for Machine Translation.

Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022b. Towards a methodology for evaluating automatic subtitling. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 333–334, Ghent, Belgium. European Association for Machine Translation.

Alina Karakanta, Franćois Buet, Mauro Cettolo, and Franćois Yvon. 2022c. Evaluating subtitle segmentation for end-to-end generation systems. In *Proceedings of LREC*, pages 3069–3078, Marseilles, France.

Alina Karakanta, Marco Gaido, Matteo Negri, and Marco Turchi. 2021a. Between flexibility and consistency: Joint generation of captions and subtitles. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 215–225, Bangkok, Thailand (online). Association for Computational Linguistics.

Alina Karakanta, 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. Association for Computational Linguistics.

Alina Karakanta, 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. European Language Resources Association.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020c. Point Break: Surfing Heterogeneous Data for Subtitle Segmentation. In *Seventh Italian Conference on Computational Linguistics, CLiC-It*.

Alina Karakanta and David Orrego-Carmona. in press. Subtitling in Transition: the case of TED talks. In *Translation in Transition: Bridging*

Human and Machine Intelligence, American Translators Association series. John Benjamins.

Alina Karakanta, Sara Papi, Matteo Negri, and Marco Turchi. 2021b. Simultaneous speech translation for live subtitling: from delay to display. In *Proceedings of the 1st Workshop on Automatic Spoken Language Translation in Real-World Settings (ASLTRW)*, pages 35–48, Virtual. Association for Machine Translation in the Americas.

Fotios Karamitroglou. 2001. *Towards a Methodology for the investigation of Norms in audiovisual translation. The choice between subtitling and Revoicing in children's programmes in Greece.* Rodopi, Amsterdam and Atlanta.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations.*

Carmen Klaus. 2015. *Translationsqualität und Crowdsourced Translation. Untertitelung und ihre Bewertung – am Beispiel des audiovisuellen Mediums TEDTalk.* Frank & Timme, Berlin.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio Augmentation for Speech Recognition. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3586–3589, Dresden, Germany.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86. Asia-Pacific Association for Machine Translation.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020a. MT for subtitling: Investigating professional translators' user experience and feedback. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 79–92, Virtual. AMTA.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020b. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124, Lisboa, Portugal. European Association for Machine Translation.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020c. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 2020)*, pages 115–124.

Izabela Krejtz, Agnieszka Szarkowska, and Krzysztof Krejtz. 2013. The effects of shot changes on eye movements in subtitling. *Journal of Eye Movement Research*, 6:1–12.

Hans P. Krings. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes.* The Kent State University Press, Kent, Ohio & London.

Helena Kruger. 2001. The creation of interlingual subtitles: Semiotics, equivalence and condensation. *Perspectives: Studies in Translatology*, 9:177–196.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Szu-Yu Kuo. 2014. *Quality in Subtitling. Theory and Professional Reality.* Ph.D. thesis, Translation Studies Unit, Imperial College London, London. Unpublished doctoral dissertation.

Alexander Künzli. 2020. From inconspicuousness to flow - the CIA model of subtitle quality. *Perspectives*, pages 326–338.

Surafel M. Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 54–61, Brussels. International Conference on Spoken Language Translation.

Surafel M. Lakew, Matteo Negri, and Marco Turchi. 2021. Zero-shot neural machine translation with self-learning cycle. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*

(LoResMT2021), pages 96–113, Virtual. Association for Machine Translation in the Americas.

Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. 2022. Isometric mt: Neural machine translation for automatic dubbing. In *Proceedings of ICASSP 2022*, Singapore.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the Output Length of Neural Machine Translation. In *Proceedings of the 16th International Workshop on Spoken Language Translation, (IWSLT)*.

Andrew Lambourne. 2006. Subtitle respeaking. *Intralinea, Special Issue on Respeaking*.

Hugo Larochelle and Geoffrey E Hinton. 2010. Learning to combine foveal glimpses with a third-order boltzmann machine. In *Advances in Neural Information Processing Systems 23*, pages 1243–1251. Curran Associates, Inc.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work*, pages 63–76, Berlin, Heidelberg. Springer.

Alon Lavie, Donna Gates, Marsal Gavalda, Laura Mayfield Tomokiyo, Alex Waibel, and Lori Levin. 1996. Multi-lingual translation of spontaneously spoken language in a limited domain. In *Proceedings of the International*

Conference on Computational Linguistics (COLING), Copenhagen, Denmark.

Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research: Using inputlog to analyze writing processes. *Written Communication*, 30:358–392.

Sixin Liao, Jan-Louis Kruger, and Stephen Doherty. 2020. The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension. *The Journal of Specialised Translation*.

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

Deborah L. Linebarger. 2001. Learning to read from television: The effects of using captions and narration. *Journal of Educational Psychology*, 93:288–298.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Danni Liu, Jan Niehues, and Gerasimos Spanakis. 2020a. Adapting end-to-end speech recognition for readable subtitles. In *Proceedings of the*

17th International Conference on Spoken Language Translation, pages 247–256, Online. Association for Computational Linguistics.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1128–1132, Graz, Austria.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020b. Bridging the Modality Gap for Speech-to-Text Translation. arxiv.org/pdf/2010.14920.pdf.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

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

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Georg-Michael Luyken, Thomas Herbst, J Langham-Brown, H. Reid, and H. Spinhof. 1991. *Overcoming Language Barriers in Television: Dubbing*

and Subtitling for the European Audience, volume 13 of *Media monograph (Manchester, England)*. The European Institute for the Media, Manchester.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Xutai Ma, Juan Pino, and Philipp Koehn. 2020. SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 582–587, Suzhou, China. Association for Computational Linguistics.

Alison Marsh. 2004. Simultaneous Interpreting and Respeaking: a Comparison. Master's thesis, University of Westminster, UK.

Giuseppe Martucci, Matteo Negri Mauro Cettolo and, and Marco Turchi. 2021. Lexical modeling of asr errors for robust speech translation. In *Proceedings of Interspeech 2021*, pages 2282–2287, Brno, Czechia.

Anna Matamala. 2021a. *Accessibilitat i traducció audiovisual*, 1 edition, volume 24 of *Biblioteca de traducció i interpretació*. Vic : Eumo Editorial.

Anna Matamala. 2021b. Audio description. In Yves Gambier and Luc van Doorslaer, editors, *Handbook of Translation Studies*, volume 5, pages 17–22. John Benjamins.

Anna Matamala. 2022. Audio subtitling. In Christopher Taylor and Elisa Perego, editors, *The Routledge Handbook of Audio Description*, page 13. Routledge, London.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005a. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Evgeny Matusov, Hermann Ney, and Ralph Schluter. 2005b. Phrase-based translation of speech recognizer word lattices using loglinear model combination. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 110–115, San Juan, Puerto Rico. IEEE.

Evgeny Matusov, 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. Association for Computational Linguistics.

Roberto Mayoral, Dorothy Kelly, and Natividad Gallardo. 1988. Concept of constrained translation. Non-Linguistic perspectives of translation. *Meta*, 33:356–367.

Maite Melero, Antoni Oliver, and Toni Badia. 2006. Automatic Multilingual Subtitling in the eTITLE Project. In *Proceedings of ASLIB Translating and the Computer 28*.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. In *Proceedings of the Second Conference on Machine Trans-*

lation, pages 99–107, Copenhagen, Denmark. Association for Computational Linguistics.

Giselle Spiteri Miggiani. 2019. *Dialogue Writing for Dubbing - An Insider's Perspective*. Palgrave Macmillan.

Joss Moorkens and Sharon O'Brien. 2017. Assessing User Interface Needs of Post-Editors of Machine Translation. *Human Issues in Translation Technology: The IATIS Yearbook*, pages 109–130.

Niko Moritz, Takaaki Hori, and Jonathan Le. 2020. Streaming automatic speech recognition with the transformer model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6074–6078. IEEE.

Mathias Müller and Martin Volk. 2013. Statistical machine translation of subtitles: From opensubtitles to ted. In *Language Processing and Knowledge in the Web*, pages 132–138, Berlin, Heidelberg. Springer Berlin Heidelberg.

Shravan Nayak, Timo Baumann, Supratik Bhattacharya, Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. See me Speaking? Differentiating on Whether Words are Spoken On Screen or Off to Optimize Machine Dubbing. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*, page 5. ACM, New York, NY, USA.

Netflix. 2021. Timed text style guides. Last accessed: 21 April 2022.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro,

Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Peter Newmark. 1988. *A Textbook of Translation*. Hemel Hempstead: Prentice Hall.

Ha Nguyen, Yannick Estève, and Laurent Besacier. 2021. An empirical study of end-to-end simultaneous speech translation decoding strategies. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7528–7532. IEEE.

Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. 2020. Improving Sequence-to-sequence Speech Recognition Training with On-the-fly Data Augmentation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain.

Jan Niehues. 2020. Machine translation with unsupervised length-constraints. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 21–35, Virtual. Association for Machine Translation in the Americas.

Jan Niehues, Roldano Cattoni, Sebastian Stucker, Matteo Negri, Marco Turchi, et al. 2019. The IWSLT 2019 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-latency neural speech translation. In *INTERSPEECH*.

Kristijan Nikolić. 2015. The pros and cons of using templates in subtitling. In Rocío Baños Piñero and Jorge Díaz Cintas, editors, *Audiovisual Translation in a Global Context: Mapping an Ever-changing Landscape*, pages 192–202. Palgrave Macmillan UK, London.

Kristijan Nikolic and Lindsay Bywood. 2021. Audiovisual Translation: The Road Ahead. *Journal of Audiovisual Translation*, 4(1):50–70.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Ofcom. 2005. Subtitling—an issue of speed? Technical report, Ofcom, London: Office of Communications.

Ofcom. 2014. Measuring live subtitling quality - results from the first sampling exercise. Technical report, Ofcom, London: Office of Communications.

Ofcom. 2015. Measuring live subtitling quality: Results from the fourth sampling exercise. Technical report, Ofcom, London: Office of Communications.

Alp Öktem, Mireia Farrús, and Antonio Bonafonte. 2018. Bilingual Prosodic Dataset Compilation for Spoken Language Translation. In *Proceedings of IberSPEECH 2018*, pages 20–24.

Alp Öktem, Mireia Farrús, and Antonio Bonafonte. 2019. Prosodic Phrase Alignment for Machine Dubbing. In *Proceedings of INTERSPEECH 2019*, pages 4215–4219.

Pilar Orero and Anna Matamala. 2007. Accessible opera: Overcoming linguistic and sensorial barriers. *Perspectives: Studies in Translatology*, pages 262–277.

David Orrego-Carmona. 2016. Internal structures and workflows in collaborative subtitling. In Rachele Antonini and Chiara Bucaria, editors, *Non-professional Interpreting and Translation in the Media*, pages 211–230. Peter Lang, Frankfurt.

David Orrego-Carmona and Yvonne Lee. 2017. Non-professional subtitling. In David Orrego-Carmona and Yvonne Lee, editors, *Non-Professional Subtitling*, pages 1–12. Peter Lang, Frankfurt.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Magdalena Oziemblewska and Agnieszka Szarkowska. 2020. The quality of templates in subtitling. A survey on current market practices and changing subtitler competences. *Perspectives*, 0(0):1–22.

Eithne O'Connell. 2007. *Chapter 8. Screen Translation*, pages 120–133. Multilingual Matters, Bristol, Blue Ridge Summit.

Minako O'Hagan. 2003. Can language technology respond to the subtitler's dilemma? - a preliminary study. In *Proceedings of the 25th International Conference on Translation and the Computer*.

Minako O'Hagan. 2019. Game localization - a critical overview and implications for audiovisual translation. In Luis Pérez-González, editor, *The Routledge Handbook of Audiovisual Translation*, 1st edition, pages 145–159. Routledge.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Sara Papi, Alina Karakanta, Matteo Negri, and Marco Turchi. 2022. Dodging the Data Bottleneck: Automatic Subtitling with Automatically Segmented ST Corpora. In *Proceedings of AACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A Simple

Data Augmentation Method for Automatic Speech Recognition. *arXiv preprint arXiv:1904.08779*.

Niki J. Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. In *International Conference on Machine Learning (ICML)*.

Maria Pavesi. 2008. Spoken language in film dubbing: Target language norms, interference and translational routines. In Delia Chiaro, Christine Heiss, and Chiara Bucaria, editors, *Between Text and Image: Updating research in screen translation*, 78, pages 79–99. Benjamins Translation Library.

Jan Pedersen. 2005. How is culture rendered in subtitles? In *Proceedings of the Marie Curie Euroconferences MuTra Challenges of Multidimensional Translation*, Saarbrücken.

Jan Pedersen. 2017. The FAR Model: Assessing Quality in Interlingual Subtitling. *The Journal of Specialised Translation*, 28:210–229.

Stephan Peitz, Simon Wiesler, Markus Nußbaum-Thom, and Hermann Ney. 2012. Spoken language translation using automatically transcribed text in training. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Elisa Perego. 2008. What would we read best? hypotheses and suggestions for the location of line breaks in films subtitles. *The Sign Language Translator and Interpreter*, 2:35–63.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Ngoc-Quan Pham, Thai-Son Nguyen, Thanh-Le Ha, Juan Hussain, Felix Schneider, Jan Niehues, Sebastian Stüker, and Alexander Waibel. 2019. The IWSLT 2019 KIT Speech Translation System. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Juan Pino, Liezl Puzon, Jiatao Gu, Xutai Ma, Arya D. McCarthy, and Deepak Gopinath. 2019. Harnessing Indirect Training Data for End-to-End Automatic Speech Translation: Tricks of the Trade. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Stelios Piperidis, Iason Demiros, and Prokopis Prokopidis. 2005. Infrastructure for a Multilingual Subtitle Generation System. In *9th International Symposium on Social Communication*, pages 24–28.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Fred Popowich, Paul McFetridge, Davide Turcato, and Janine Toole. 2000. Machine Translation of Closed Captions. *Machine Translation*, pages 311–341.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Tomasz Potapczyk and Pawel Przybysz. 2020. SRPOL's system for the IWSLT 2020 end-to-end speech translation task. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 89–94, Online. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, K. Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi Speech Recognition Toolkit. In *Proceedings of ASRU 2011*, pages 1–4, Big Island, Hawaii, USA.

Daniel Povey, Hossein Hadian, Pegah Ghahremani, Ke Li, and Sanjeev Khudanpur. 2018. A Time-Restricted Self-Attention Layer for ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878, Calgary, Canada.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In *Proceedings of Interspeech*, pages 2751–2755.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Luis Pérez-González. 2008. Audiovisual translation. In Mona Baker and Gabriela Saldanha, editors, *Encyclopedia of Translation Studies*, pages 13–20. Routledge, London; New York.

Luis Pérez-González. 2014. *Audiovisual Translation: Theories, methods, issues*. Routledge, London & New York.

Franz Pöchhacker. 2001. Quality assessment in conference and community interpreting. *Meta*, 46:410–425.

Franz Pöchhacker and Aline Remael. 2020. New efforts? a competence-oriented task analysis of interlingual live subtitling. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 18(0).

Peng Qi, 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*, pages 101–108, Online. Association for Computational Linguistics.

Dhevi J. Rajendran, 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.

Keith Rayner, Simon P. Liversedge, and Sarah J. White. 2006. Eye movements when reading disappearing text: The importance of the word to the right of fixation. *Vision Research*, pages 310—323.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Aline Remael and Nina Reviers. 2019. Media accessibility and accessible design. In Minako O' Hagan, editor, *The Routledge Handbook of Translation and Technology*, 1st edition, pages 482–497. Routledge, London.

Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. SimulSpeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796, Online. Association for Computational Linguistics.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Zelasko, and Miguel Jette. 2021. Earnings-21: A practical benchmark for ASR in the wild. In *Proceedings of INTERSPEECH*, pages 3465–3470, Brno, Czech Republic.

Allison Rittmayer. 2009. Translation and film: Slang, dialects, accents and multiple languages. *Comparative Humanities Review*, 3.

Isabelle S. Robert and Aline Remael. 2016. Quality control in the subtitling industry : An exploratory survey study. *Meta: Journal des traducteurs*, 61:578–-605.

Pablo Romero-Fresco. 2010. Standing on quicksand: hearing viewers' comprehension and reading patterns of respoken subtitles for the news. In Jorge Díaz Cintas, Anna Matamala, and Joselia Neves, editors, *New insights into audiovisual translation and media accessibility*, pages 175 – 194. Brill, Leiden, The Netherlands.

Pablo Romero-Fresco. 2011. *Subtitling through speech recognition: Respeaking*. Manchester: St. Jerome.

Nicholas Ruiz, Qin Gao, William Lewis, and Marcello Federico. 2015. Adapting machine translation models toward misrecognized speech with text-to-speech pronunciation rules and acoustic confusability. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany.

Ashutosh Saboo and Timo Baumann. 2019. Integration of dubbing constraints into machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 94–101, Florence, Italy. Association for Computational Linguistics.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset For Multimodal Language Understanding. In *Proceedings of Visually Grounded Interaction and Language (ViGIL)*, Montréal, Canada. Neural Information Processing Society (NeurIPS).

Klaus Schwab and Thierry Malleret. 2021. *COVID-19: The Great Reset*, 1.0 edition. World Economic Forum. Forum Publishing, Cologny/Geneva, Switzerland. Https://carterheavyindustries.files.wordpress.com/2020/12/covid-19_-the-great-reset-klaus-schwab.pdf.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Iris Serrat Roozen 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. European Language Resources Association.

Mayank Sharma, Yogesh Virkar, Marcello Federico, Roberto Barra-Chicote, and Robert Enyedi. 2021. Intra-sentential speaking rate con-

trol in neural text-to-speech for automatic dubbing. In *Proceedings of INTERSPEECH 2021*, pages 3151–3155, Brno, Czechia.

Selina Sharmin, Satu Mari-Anna Wiklund, and Liisa Tiittula. 2016. The reading process of dynamic text – a linguistic approach to an eye movement study. *SKY Journal of Linguistics*, pages 119–146.

Angela Sileo. 2020. Dubbing or Simil Sync? A Study on Reception in Italy. In Giulia Magazzù, Valentina Rossi, and Angela Sileo, editors, *Reception Studies and Adaptation: A focus on Italy*. Cambridge Scholars Publishing.

Ferrer Simó and María Rosario. 2005. Fansubs y scanlations: la influencia del aficionado en los criterios profesionales. *Puentes*, 6:27–43.

Slator. 2022. What to do about the talent crunch in media localization. https://slator.com/what-to-do-about-the-talent-crunch-in-media-localization/. Last accessed: 12/05/2022.

Mary Snell-Hornby. 1988. *Translation Studies: An integrated approach*. John Benjamins Publishing.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Hye-Jeong Song, 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). *Applied Sciences*, 9:1504.

Matthias Sperber, Graham Neubig, Ngoc-Quan Pham, and Alex Waibel. 2019. Self-attentional models for lattice inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1185–1197, Florence, Italy. Association for Computational Linguistics.

Matthias Sperber, Jan Niehues, Graham Neubig, Sebastian Stüker, and Alex Waibel. 2018. Self-Attentional Acoustic Models. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3723–3727, Hyderabad, India.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward Robust Neural Machine Translation for Noisy Input Sequences. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Matthias Sperber, Hendra Setiawan, Christian Gollan, Udhyakumar Nallasamy, and Matthias Paulik. 2020. Consistent transcription and translation of speech. *Transactions of the Association for Computational Linguistics*, 8:695–709.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Fred W.M. Stentiford and Martin G. Steer. 1988. Machine translation of speech. *British Telecom Technology Journal*, 6(2):116–122.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34(2):97–147.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Tetyana Sydorenko. 2010. Modality of input and vocabulary acquisition. *Lang Learn Technol*, pages 50–73.

A Szarkowska. 2016. Report on the results of an online survey on subtitle presentation times and line breaks in interlingual subtitling. part 1: Subtitlers. Technical report, London: University College London.

Agnieszka Szarkowska. 2020. Subtitling for the deaf and the hard of hearing. In Łukasz Bogucki and Mikołaj Deckert, editors, *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, Palgrave Studies in Translating and Interpreting, pages 249–268. Cham: Palgrave Macmillan.

Agnieszka Szarkowska, Díaz Cintas, Jorge, and Olivia Gerber-Morón. 2020. Quality is in the eye of the stakeholders: what do professional subtitlers and viewers think about subtitling? *Universal Access in the Information Society*, 20:661–675.

Agnieszka Szarkowska and Olivia Gerber-Morón. 2018. Viewers can keep up with fast subtitles: Evidence from eye movements. *PLOS ONE*, 13(6):1–30.

Agnieszka Szarkowska, Izabela Krejz, Olga Pilipuczuk, Łukasz Dutka, and Jan-Louis Kruger. 2016. The effects of text editing and subtitle presentation rate on the comprehension of reading patterns of interlingual and intralingual subtitles among deaf, hard of hearing and hearing viewers. *Across Languages and Cultures*, 17:183–204.

Sofia Sánchez Mompeán. 2020. *The Prosody of Dubbed Speech - Beyond the Character's Words*. Palgrave Studies in Translating and Interpreting. Palgrave.

Derek Tam, Surafel M. Lakew, Yogesh Virkar, Prashant Mathur, and Marcello Federico. 2021. Prosody-aware neural machine translation for dubbing.

TED. n.d. Subtitling tips. Last accessed: 21 April 2022.

Paul Thibault. 2000. The Multimodal Transcription of a Television Advertisement: Theory and Practice. *Baldry 2000*, pages 311–385.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2008. Synchronizing translated movie subtitles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources*

and Evaluation (LREC'12), 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 the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. *arXiv e-prints*, pages arXiv–2202.

Tiina Tuominen. 2018. Multi-method research - reception in context. In Elena Di Giovanni and Yves Gambier, editors, *Reception Studies and Audiovisual Translation*, volume 141, pages 69–90. BTL.

Marco Turchi, Matteo Negri, M. Amin Farajian, and Marcello Federico. 2017. Continuous learning from human post-edits for neural machine translation.

Dusan Varis and Ondřej Bojar. 2021. Sequence length is a domain: Length-based overfitting in transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8246–8257, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6000–6010.

Lawrence Venuti. 1995. *The Translator's Invisibility: A History of Translation*. Routledge, New York.

Lucas Nunes Vieira. 2020. Automation anxiety and translators. *Translation Studies*, 13(1):1–21.

Yogesh Virkar, Marcello Federico, Robert Enyedi, and Roberto Barra-Chicote. 2021. Improvements to prosodic alignment for automatic dubbing. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7543–7574.

Andrew Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13:260–269.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine Translation of TV Subtitles for Large Scale Production. In *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC'10)*, pages 53–62, Denver, CO.

Pavel Vondřička. 2014. Aligning parallel texts with Intertext. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1875–1879. European Language Resources Association (ELRA).

Alex Waibel, Ajay N Jain, Arthur E McNair, Hiroaki Saito, Alexander G Hauptmann, and Joe Tebelskis. 1991. Janus: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 793–796, Toronto, Canada. IEEE Computer Society.

Changhan Wang, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Ann Lee, Peng-Jen Chen, Jiatao Gu, and Juan Pino. 2021. fairseq sˆ2: A scalable and integrable speech synthesis toolkit. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 143–152, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Chengyi Wang, Yu Wu, Shujie Liu, Jinyu Li, Liang Lu, Guoli Ye, and Ming Zhou. 2020b. Low latency end-to-end streaming speech recognition with a scout network. *arXiv preprint arXiv:2003.10369*.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2625–2629, Stockholm, Sweden.

Candace Whitman-Linsen. 1992. *Through the Dubbing-Glass*. Peter Lang, Frankfurt am Main.

Patrick Wilken, Panayota Georgakopoulou, and Evgeny Matusov. 2022. SubER - a metric for automatic evaluation of subtitle quality. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 1–10, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Jie Yang and Yue Zhang. 2018. Ncrf++: An open-source neural sequence labeling toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Patrick Zabalbeascoa. 1996. Translating jokes for dubbed television situation comedies. *The Translator*, 2:252–257.

Patrick Zabalbeascoa. 2008. The nature of the audiovisual text and its parameters. *The Didactics of Audiovisual Translation*, pages 21–37.

Xingshan Zeng, Liangyou Li, and Qun Liu. 2021. RealTranS: End-to-end simultaneous speech translation with convolutional weighted-shrinking transformer. In *Findings of the Association for Computational Linguis-*

tics: ACL-IJCNLP 2021, pages 2461–2474, Online. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Ruiqing Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2022. Learning adaptive segmentation policy for end-to-end simultaneous translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7862–7874, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Baigong Zheng, Kaibo Liu, Renjie Zheng, Mingbo Ma, Hairong Liu, and Liang Huang. 2020. Simultaneous translation policies: From fixed to adaptive. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2847–2853, Online. Association for Computational Linguistics.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019a. Simpler and faster learning of adaptive policies for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1349–1354, Hong Kong, China. Association for Computational Linguistics.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019b. Simultaneous translation with flexible policy via restricted imitation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5816–5822, Florence, Italy. Association for Computational Linguistics.

Renjie Zheng, Mingbo Ma, Baigong Zheng, and Liang Huang. 2019c. Speculative beam search for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1395–1402, Hong Kong, China. Association for Computational Linguistics.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. 2020. Rethinking pre-training and self-training. In *Proceedings of NeurIPS*.

Appendix A

Machine Translation for Dubbing

The focus of this thesis is the automatisation of subtitling. Still, I could not completely ignore a topic I spent significant time of my PhD on. This is dubbing, the AVT sibling of subtitling. Even though this chapter is short and inconclusive, it deserves a special part in this thesis, as a reminder that patience and perseverance are needed to make beautiful ideas flourish when the preconditions for them will be in place.

In the previous chapters we approached subtitling from a computational perspective, from subtitling data compilation, direct ST architectures, to tailored evaluation methods. The time was ripe for such a contribution, which was built upon the existence of publicly available subtitling data and a growing interest from the ST community and industry in subtitling. Dubbing, another dominant AVT type, has not benefited from Machine Translation to the same extent, mainly because the prerequisites for developing and testing automatic dubbing systems are not yet in place. MT for dubbing still largely suffers from data scarcity, which has long impeded the deployment of NMT (let alone ST) solutions. However, the synchronisation in dubbing and its strong relationship with timing, prosody and phonetics would make dubbing an excellent test bed for experimenting with the limits of constrained NMT and direct ST.

This chapter takes inspiration from the particularities of the dubbing genre and the synchronisation constraints in dubbing and makes the first steps towards exploring automatic approaches to dubbing. The goal is to understand the language of dubbing based on computational methods and to prepare the ground for customised solutions to automatic dubbing. We first ask the question whether dubbing has two 'shades', determined by different synchronisation requirements between on-screen and off-screen dubbing (Section A.1). We investigate this dichotomy by performing a corpus-based analysis of the only publicly available dubbing corpus at different linguistic levels (Section A.2). Then, in Section A.3 we move to the application of Neural Machine Translation to the two shades of dubbing. In particular, we are interested in performance differences between on-screen and off-screen dubbing and on the effect of applying synchronisation constraints on each of the shades. This investigation is important to guide future decisions on constraining/relaxing the generation process depending on the degree of synchronisation required for a successful dubbing.

A.1 Does dubbing have two shades?

Dubbing is a form of audiovisual translation (AVT) which consists in replacing the original audio track of a film dialogue with another track containing the dialogue translated in a different language, which follows the timing and phrasing of the original (Luyken et al., 1991). The goal is to make it appear as if the film had been recorded in the target language all along. It is the preferred type of AVT in countries with large film and streaming markets, e.g. Germany, Italy, Spain and parts of Latin America (Bogucki and Díaz-Cintas, 2020). Dubbing is among the few translation areas where Machine Translation has found no steady ground yet. The first reason is the dearth of dubbing corpora, not only for training NMT

systems, but even for testing purposes. Lip-synchronised dubbing is employed for high-budget productions on proprietary material protected by copyright, such as films. Given the scarcity of public material containing dubbing, large-scale data collection is impracticable.

The second reason is the particularity of dubbing as a genre. Translation for dubbing is a text written to be spoken, therefore it should bear a close approximation to orality and reflect *oral unlaboured dialog* (Chaume-Varela, 2006). However, dubbing is not spontaneous, impromptu speech but rather a carefully-prepared imitation of spoken language. This *prefabricated orality* (Baños-Piñero and Chaume, 2009) is what confers dubbing particular linguistic characteristics. In addition, the most distinctive characteristic of dubbing is the need for synchronisation when an actor's mouth appears on screen. Isochrony requires that the duration of the source and target utterance is equal, in order for the translated dialogue to exactly fit the time during which the actor speaks. At a second level, lip-sync consists in adapting the translation to match the articulatory mouth movements of the actor, mainly matching open vowels and bilabials. In practice, this means matching an open mouth to an open vowel and a closed mouth to a bilabial.[1] Although isochrony is required whenever the face of the actor appears on screen, lip-sync is required only for close-ups and extreme close-ups.

The following example illustrates the importance of isochrony and lip-sync in translations for dubbing. The sentence "*I get strong off other people's fear*" could be translated as "*Me fortalezco con el miedo de otras personas*". However, this translation is too long (15 syllables) to match the duration of the source (8 syllables). Instead of this literal translation, the source phrase is translated as "*El miedo de los demás me da fuerzas*" (11

[1]A third type of synchronisation is *kinesic isochrony* (synchronisation with body movements) which would require a multimodal analysis of the actions of the characters on screen and is therefore outside the scope of this paper.

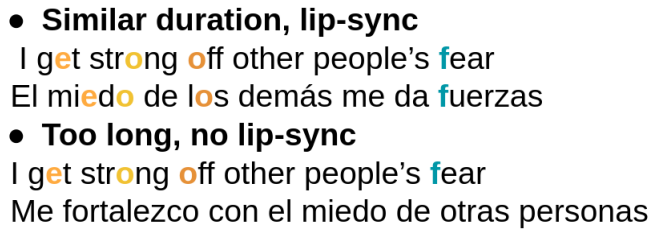


Figure A.1: Example of a translation done with lip sync and one not.

syllables), in order to better match the duration of the source utterance and the overlap of round vowels, e.g. *o* in *strong-miedo* and *f* between *fear-fuerzas*. See Figure A.1 on an alignment between source-target phonemes.

From the above, it becomes evident that there is a clear dichotomy in dubbing strategies when an actor's mouth is visible on screen or not (off screen). While on-screen dubbing is bound by synchronisation constraints in order to create the illusion of authenticity (Whitman-Linsen, 1992), off-screen dubbing should simply be representative of orality.

Previous work on automatic machine dubbing has focused mostly on matching the length between source and target utterances and has shown that, while translation quality degrades only moderately, introducing such matching constraints improves 'dubbability' (i. e., how well the translations can be spoken synchronously to the original speech). Among the first such approaches, Saboo and Baumann (2019) attempt to integrate the constraint of isochrony by selecting the translation which has a similar number of syllables to the source. Federico et al. (2020) adapt an NMT system using length-control mechanisms to generate translations of the same length as the source, although in terms of characters, which does not necessarily reflect duration of utterance. Later, Lakew et al. (2022) proposed self-learning to for the task of *Isometric NMT*, that is generating outputs ±10% longer or shorter than the input text, without significant degradation in the translation quality.

Another line of research focuses on improving the prosodic alignment

between the source and target segment. A good prosodic alignment means that the translation can optimally match the sequence of phrases and pauses of the source. Öktem et al. (2019) use the NMT attention mechanism to segment the translation into prosodic phrases in order to improve a Text-to-Speech system for dubbing. Tam et al. (2021) attempt to project the pauses from the source to the target with implicit and explicit prosody control methods. Virkar et al. (2021) improve the alignment by incorporating various features (language model, semantic matches, isochrony scores and speaking rate). However, they find that this approach often leads to unnatural speaking rates. These deficiencies can however be tamed with duration modelling (Effendi et al., 2022) and controlling the speaking rate (Sharma et al., 2021) in the Text-to-Speech component. Taking inspiration from the work on the on/off-screen dichotomy presented in this chapter, a subsequent study adapts the prosodic alignment to off-screen dubbing which requires less stringent synchronisation constraints (Virkar et al., 2021). However, none of these works has addressed or studied synchronisation constraint of lip sync.

In this chapter, we present work that addressed, for the first time, the dichotomy between on-screen and off-screen dubbing in NMT. As a first contribution, we manually annotate the only publicly available dubbing corpus (Öktem et al., 2018) by explicitly marking whether the actor's mouth appears on screen or not. This annotation is the first of its kind and provides an invaluable resource for the study of dubbing and further applications (see (Nayak et al., 2020)). Second, we look into distinctive features between on-screen and off-screen dubbing at the textual, prosodic and phonetic level, in an attempt to understand the effect of synchronisation constraints (if present) at different linguistic levels. Lastly, we tune a general NMT system to the dubbing domain and test its performance on on-screen and off-screen speech, investigating different tagging

strategies. We additionally apply isochrony constraints to the output and discuss the compromise between synchronisation and translation quality for the two modes. Our findings suggest that using a single NMT system for all "shades" of dubbing is not optimal, but viable solutions for dubbing-customised NMT should take into account the on/off dichotomy.

A.2 On/Off-screen dubbing

In this section, we present the on/off-screen annotation process and an analysis of the textual, prosodic and phonetic features between the classes.

A.2.1 Data and annotation

The Heroes corpus (Öktem et al., 2018) is the only freely available dubbing corpus and is based on the drama television series Heroes (seasons 2 and 3). It contains 7,000 single-speaker English utterance segments and their translations into Spanish, as well as time-alignments that allow for re-aligning the text with the video. Before starting the annotation, we verified that synchronisation is actually present in Heroes by watching the dubbed videos. We found that dubbing is convincing and that some degree of synchronisation is present, with a stronger emphasis on isochrony than lip sync. Despite this initial indication of a lack of perfect lip sync, Heroes is the only available dubbing corpus, and contains audio segments on the source and target side, which makes it promising for testing speech translation solutions for dubbing.[2]

A first annotator watched all videos and annotated the source side of the corpus on the word-level, specifying whether speech is on-screen (mouth visible on screen) or off. Consequently, each utterance can be fully on-screen, fully off-screen, or mixed (if the actor's mouth is visible only for

[2]The annotation is publicly available at `https://ict.fbk.eu/heroes-on-off/`

Class	Utt.s	Tokens-en	Tokens-es
On	4,163	29,338	25,159
Off	1,200	8,238	7,293
Mixed	1,614	18,792	16,086
Mixed-on		10,976	-
Mixed-off		7,816	-
All	6,977	56,368	48,538

Table A.1: Corpus statistics in utterances and source/target tokens for the three classes.

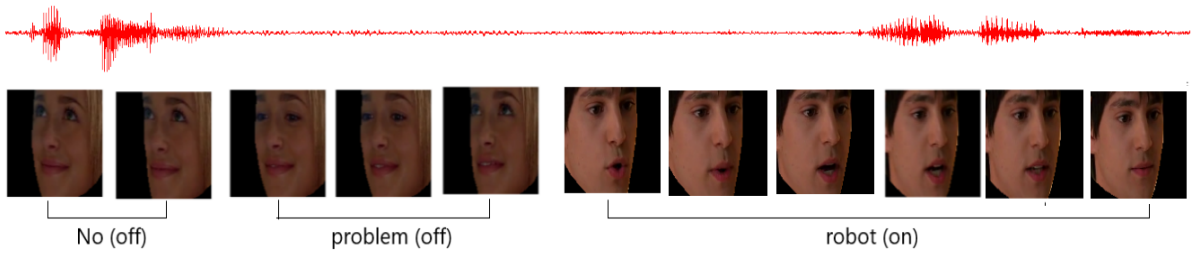


Figure A.2: An example: "No problem, robot", of the classification task along with the annotations (s2_1_250).

some words). An example of a mixed segment and it's annotation can be seen in Figure A.2. To validate the reliability of the annotation, a second annotator worked on 700 randomly selected utterances (10 % of the corpus). This yielded an inter-annotator agreement (Cohen's Kappa) of $\kappa = .73$ on the word level and .73 on the utterance level.[3] We conclude that annotation as on/off/mixed is possible with a substantial agreement across annotators. During annotation, we fixed several transcription errors and identified non-English source utterances, resulting in a total of 6,977 segments. Table A.1 presents the annotation statistics.[4] As can be seen, the majority of tokens appears on-screen (71 %; for utterances: 60 %). This

[3]Some deviation is expected, given that words can be short and audio-video alignment is never perfect.

[4]Since annotation is performed only on the source, it is not possible to report Spanish tokens for the mixed category.

imbalance is to be expected given the nature of drama television series, where the action focuses on dialogues among characters.

A.2.2 Analysis of textual, prosodic and phonetic features

We investigate characteristics that may differentiate on-screen from off-screen dubbing, focusing at the textual, prosodic and phonetic level. We first consider **textual** differences. We looked at relative length of translations, expressed by character and syllable ratios between source and target. The ratios are shown in Table A.2. We found very little differences between the mean of these ratios among the On/Off/Mixed classes. The same holds for the ratio between consonants and vowels, as well as for the duration of the utterances. However, a larger standard deviation of ratios for Off suggests that there is more freedom both in terms of length and word choice when no synchronisation is needed. The lack of differences indicates that textual features are probably not sufficient to distinguish between the ON and OFF category.

In order to identify possible latent characteristics (such as character sequence patterns), we also built a character-based neural sequence classifier that determines the screen class based on English, Spanish, or both texts. The classifier is a bidirectional LSTM, trained using the Dynet framework (Neubig et al., 2017). It is trained using 80% of the Heroes data per class, while 10% are left for development and testing respectively. We perform 10 cross-validation rounds. [5] Given that classes are unbalanced, we choose Area Under ROC curve as performance metric: an ideal classifier would cover an area of 1.0, random classification yields .5. Although we find some limited success in identifying mixed utterances, our overall classification results are rather poor, as shown in Table A.3. We there-

[5]Implementation details as well as the code can be found at `https://github.com/timobaumann/visualHeroesAnnotations/tree/master/text-classifier`

fore conclude that, in this data set, there are no clear distinctive text-level features (neither direct, nor latent features that could be used by a neural network classifier) that differentiate on-screen from off-screen dubbing, i.e., that possible traces of synchronisation are hardly discernible in the textual form.

Ratio	ON	OFF	MIXED
Character	1.10 (0.27)	1.10 (0.38)	1.13 (0.29)
Syllable	0.90 (0.26)	0.90 (0.36)	0.93 (0.27)
Vowel	0.92 (0.28)	0.93 (0.38)	0.94 (0.29)
Consonant	1.27 (0.41)	1.28 (0.56)	1.27 (0.39)
Duration	1.04 (0.31)	1.07 (0.36)	1.04 (0.24)

Table A.2: Mean and standard deviation on the source/target ratio of different units (characters, syllables, vowels vs. consonants and utterance duration in seconds).

classifier input	words	utt.s
English only	.549	.640
Spanish only		.641
English+Spanish		.634

Table A.3: Area under ROC curve results for classifying words as on/off (only for English input), or utterances as on/off/mixed (for all inputs).

Moving to the analysis of **prosodic** features, we focused on fundamental frequency (f0), intensity, and prosodic phrase length between the target and source language. Our hypothesis is that on-screen dialogue lines will present a higher similarity of prosodic features between source and target segments compared to off-screen dialogue lines. Since normalising frequency and intensity per segment level would flatten any particular prosodic characteristics, we directed the analysis at concurrent frequency and intonation peaks in paired source-target segments. Peaks were defined as values of 2 standard deviations above or below the absolute mean

f0/intensity values, computed by taking all the values in the data set. Peaks can thus be maximum or minimum points; a peak in intensity implies the actor speaks at a loud voice or whispers. To account for speaker differences, we used the normalised values contained in the corpus (expressed in a logarithmic semitone scale relative to the speaker normalised value). We selected the segments where both the source and target contain f0 or intensity peaks. The second analysis looked at differences in the length of prosodic phrases in terms of the number of words and duration. Again, the expectation was to find a smaller difference in on-screen dubbing compared to the off and mixed categories.

	Frequency			Intensity		
	ON / OFF / MIXED			ON / OFF / MIXED		
Concurrent peaks	294	125	201	259	56	154
Percentage of total	7,06%	10.4%	12.44%	6,22%	4,67%	3,46%

Table A.4: Concurrent peaks between source and target segments for frequency (f0) and intensity.

When it comes to concurrent peaks, the results in Table A.4 show that most concurrent peaks appear in the on-screen category (294 for f0 and 159 for intensity). However, it should be noted that On is the class with by far the most segments. Indeed, when looking at the percentage of concurrent peaks out of the total number of source peaks, On has the lowest percentage of concurrent peaks for f0. The opposite is true for intensity, where On has the highest value.[6] This suggests that intonation may be a characteristic more often imitated in on-screen speech by dubbing actors than frequency which depends on the each actor's voice and acting style. Still, frequency/intensity peaks are replicated in less than 12% of sentences regardless of screen mode.

[6]Differences in both cases are statistically significant at p=0.05 based on a chi-square test.

	ON		OFF		MIXED	
	En vs. Es		En vs. Es		En vs. Es	
Mean number of phrases	1.41	1.39	1.38	1.4	2.03	1.96
Mean duration	1.4s	1.38s	1.48s	1.37s	1.93s	1.77s
SD duration	0.84s	0.78s	0.99s	0.86s	1.1s	1.04s

Table A.5: Prosodic phrase length in mean number of phrases per segment and mean duration.

As for prosodic phrase length, Table A.5 shows that the difference both in terms of length in words and duration in on- and off-screen speech is negligible, while mixed dialogue shows a relatively larger albeit still insignificant difference. We can thus conclude that screen category does not affect voice modulation, at least in the Heroes dubbing corpus; the differences between languages are possibly due to differences in the prosody of English compared to Spanish and among individual speakers. Our findings are in line with (Sánchez Mompeán, 2020), who argues that despite some patterns introduced by source text assimilation, dubbing actors resort to the Spanish prosodic repertoire. Finally, voice modulation may not be possible when translation/synchronisation strategies do not vary depending on screen category, as corroborated by the analysis of textual features above.

Lastly, we investigate **phonetic** features between the source and target segment as an indication of lip sync. As in the previous analyses, we expect that on-screen dubbing will have a greater articulatory similarity between source and target. This can be translated into a higher percentage of phoneme matches, especially bilabials (p, b, m), labiodentals (f, v), open vowels (a, E, @, 6, etc) and lip-rounded vowels (U, w, o, etc) (Miggiani, 2019). These features are the most visually noticeable, and thus are taken into consideration for lip-sync dubbing. Segments were synchronically aligned such that the onset of the speech in both languages matches.

This allowed us to look at the degree of phonetic overlap depending on the screen-value category. The phonetic overlap does not necessarily require a phoneme-to-phoneme correspondence, but a phonetic class-to-phonetic class correspondence would be sufficient to create the effect of lip sync. For example, matching the phoneme *p* to a phoneme of the bilabial class, e.g. *b*, in the target creates the same effect as matching it to the same phoneme.

phonetic class	ON	OFF	MIXED
dento-alveolo-postalveolar-c	17.2	17.4	17.2
pause	16.2	15.0	12.0
back-round-v	8.0	8.3	7.8
mid-open-v	7.8	8.0	7.5
velar-c	7.1	6.8	6.9
bilabial-c	6.0	5.5	5.4
close-v	4.7	4.5	4.5
labiodental-c	1.2	1.0	0.7

Table A.6: Percentage of phonetic class correspondence between source and target for the three classes. 'c' stands for consonant and 'v' for vowel.

Results for the three classes are presented in Table A.6. We found no statistically significant differences between the On and Off groups for most phonetic classes. However, the match percentage for pauses is higher for On-screen segments. Respecting the pauses is important in on-screen dubbing, since they correspond to a closed mouth or separate prosodic phrases and thus the existence of a voice would be noticed as a discrepancy by the viewers. The lowest overlap is found for the mixed class, confirming that it is the odd one out. Yet, the percentages of phonetic overlap are in general low. One explanation is that the dubbing in Heroes does not give emphasis to lip sync as in classical dubbing. It is possible that Heroes, being a series and not a featured film, has more flexible synchronisation requirements, what has been called *Simil sync*. Simil sync is halfway between dubbing

and voice over: like dubbing but unlike voice over, it respects isochrony; unlike dubbing but like voice over, it does not respect lip sync (Sileo, 2020). These findings corroborate our observations during the annotation phase, which observed the presence of isochrony but no perfect lip sync.

Our analysis based on textual, prosodic and phonetic characteristics did not reveal any differences between on- and off-screen dubbing. Our findings are rather disappointing and contrary to the literature in AVT, which suggests that translation for dubbing has particular characteristics imposed by the synchronisation requirements. However, due to the limited nature of the study, the findings cannot be generalised to on-/off-screen dubbing. Our data consists in 20 episodes of Heroes, dubbed in only one target language. The degree of synchronisation applied to a dubbing translation depends on several factors, such as the translation instructions, the time given to translators to complete the task and the working conditions. For example, series have less synchronisation requirements and even inside the series genre, the degree of synchronisation depends on the target audience, the language (variety) and the cost of production (Luyken et al., 1991). Due to copyright issues, dubbing translators may often translate without access to the video, therefore without knowing when they have to apply synchronisation. All these external factors influence the translation strategies and, as a result, the characteristics of on-/off-screen dubbing at all linguistic levels.

A.3 The two shades of dubbing in NMT

We now work on the hypothesis that, if there is no difference between on/off-screen dubbing in Heroes, as suggested from the corpus analysis above, an MT system should achieve comparable performance across the different classes. If this is not the case, there should be different dubbing

strategies mirrored in the translation, possibly due to synchronisation constraints. To validate this hypothesis, we test the performance of an NMT model for on/off-screen dubbing.

A.3.1 Experimental setting

Data: We create a test set by randomly selecting 400 sentences from each class, paying attention to avoid data leakage to the training set. From the remaining sentences, 10% from each class are sampled to form the development set. Since the number of sentences among the classes is unbalanced, we over-sample the two smaller classes (OFF and MIXED) until they reach the size of the ON class, resulting in a training corpus of 9,500 sentences. In order to obtain reliable results, we perform three rounds of cross-validation and report the mean and standard deviation of BLEU scores.

Base model: We pretrain an NMT model on a total of 67M parallel sentences from the OPUS project,[7] containing OpenSubtitles (Lison and Tiedemann, 2016), Europarl (Koehn, 2005), GlobalVoices[8], MuST-C (Di Gangi et al., 2019a), WIT^3 (Cettolo et al., 2012). All data are segmented into subword units using SentencePiece (Kudo and Richardson, 2018) with a 40K joint vocabulary. The model is based on the Transformer (big) architecture (Vaswani et al., 2017) of the fairseq toolkit (Ott et al., 2019). It is trained with label smoothed cross-entropy and 0.1 label smoothing. For optimisation, we use Adam (Kingma and Ba, 2015) with an initial learning rate of $1x10^{-7}$, which increases linearly up to 0.005 for 4000 warm-up steps, and then decreases with the inverse square root of the training step. Dropout is set to 0.3 for all layers except for the attention layer, where dropout is set to 0.1. The base model achieves a BLEU score of 31.47 on the WMT'13 English→Spanish test set (Bojar et al., 2013).

[7] http://opus.nlpl.eu/

[8] http://opus.nlpl.eu/GlobalVoices-v2017q3.php

Proposed models: We customise the base NMT model to the dubbing domain by fine-tuning on the Heroes data. Given the scarcity of training material, we concatenate data from all classes and follow two different strategies: 1) simply fine-tuning on all the dubbing data (*ft-heroes*) and 2) fine-tuning with *target forcing*, as in multilingual translation (Johnson et al., 2017). This technique has been shown to be beneficial not only for mixing data from different languages, but also to achieve style transfer (Niu et al., 2017) and to control attributes of the target text, such as politeness (Sennrich et al., 2016a). Since the third class (MIXED) is essentially a mix of ON and OFF, we experiment with two tagging strategies for target forcing: a) we distinguish the 3 classes with the tags <ON>, <OFF> and <MIX> (*ft-3tags)*, and b) we use only the <ON> and <OFF> tags, by annotating the ranges for on/off-screen text in the MIXED class (*ft-2tags*). For example, based on the second strategy, the sentence "No problem, robot", would be annotated as *<OFF> No problem, <ON> robot.* The second strategy is inspired by the use of dubbing symbols in the actual dialogue lists (Chaume, 2012), therefore it explores the possibility of a direct application in the dubbing industry.

Additionally, we apply the previous work attempting to integrate dubbing constraints in NMT (Saboo and Baumann, 2019) to our *ft-2tags* model (*syll-rescore*). We use a beam search of 20 and select the hypothesis with the closest source/target syllable ratio[9] to the target syllable ratio observed in our data (0.9). Compared to previous work, we use a smaller beam (20 instead of 50) because we found that, in 90% of the sentences, at least one candidate matching the target ratio was found among the hypotheses.

A.3.2 Results

The BLEU scores for the models described above are shown in Table A.7.

[9]Computed with the Pyphen package: https://pyphen.org/

Model	Total	On	Off	Mixed
base	21.32 (0.61)	21.30 (0.5)	24.04 (0.44)	20.07 (1.22)
ft-heroes	26.54 (3.09)	25.89 (2.6)	30.40 (4.2)	24.95 (3.09)
ft-3tags	**27.45** (3.1)	25.82 (2.23)	**31.11** (4.01)	**27.01** (3.01)
ft-2tags	**27.50** (3.09)	**26.63** (2.47)	**30.84** (4.3)	26.33 (3.76)
syll-rescore	22.80	23.31	23.71	24.11

Table A.7: BLEU scores for different systems for the entire test set (Total) and the three classes. Mean and StDev (in parentheses) of 3 cross-validation rounds.

On/Off-screen dubbing: Fine-tuning on the Heroes data (*ft-heroes*) gives an increase from 21.30 to 25.89 BLEU for On (+19%), from 20.07 to 24.95 for Mixed (+19%) and from 24.04 to 30.40 for Off (+24%). The score for the Off class is higher under all conditions, showing that off-screen dubbing is less challenging for NMT. Fine-tuning on in-domain data is in general expected to bring performance gains. Based on the evidence shown, we cannot establish whether the model is simply fine-tuned to the style and vocabulary of the Heroes TV series, or to the dubbing task. Hypothetically, if we could fine-tune on another dubbing dataset, we would still likely see performance gains. This would be an indication that, even though the *base* model is trained on corpora of the spoken domain, dubbing is a particular sub-domain with its own particularities in terms of style and orality.

Tagging: Target forcing gives a ∼1 BLEU point increase, showing that it helps the NMT system adapt to different styles and particularities even inside this sub-domain. For the On class, the increase comes only when the Mixed data are separated in On and Off (*ft-2tags*), since this is the only condition when NMT receives more On data. In this condition, there is a slight drop in BLEU for the Mixed class, possibly because the mix of tags increases the complexity of the source sentence.

Syllablic rescoring: When applying isochrony constraints to select the

translation candidate with the most similar number of syllables to the source, the translation performance is compromised. This drop is in line with Saboo and Baumann (2019) who attempted to balance the trade-off between isochrony and translation quality, but without making any on/off-screen distinction. Our results show an interesting tendency; while the drop for On and Mixed when applying synchronisation constraints is 3 BLEU points, for Off the score drops by 7 points. As a result, the translation performance for all the classes flattens to ∼23-24 BLEU points. This finding suggests that for off-screen dubbing, applying synchronisation constraints is not required but rather detrimental. On the other hand, the compromise for the classes containing on-screen dubbing is smaller, which advocates for the presence of some isochrony constraint for On and Mixed. Still, the drop in translation performance indicates that isochrony is not the only constraint affecting the translation, but the need for adaptation to the target language norms and the language of 'dubbese' should be further explored in NMT for dubbing.

A.3.3 Analysis

The BLEU scores and the improvements from fine-tuning are higher for the Off class. Our first hypothesis was that this difference is due to different contents/styles in the on/off test sets; for example, off-screen contains more narratives that are easier to translate, whereas on-screen more informal language. However, the majority of segments come from dialogues between characters, and narratives account only for around 10 segments in the whole corpus. Since the styles of the source segments are comparable, the lower BLEU scores for the On and Mixed class suggest that (partially and fully) on-screen dubbing is highly constrained translation, which depends not only on semantics, but on factors to which the NMT model does not have access, such as phonetics and visemics. Therefore, this difference

should be investigated in the interplay between the source and target in terms of translation solutions.

Model	On	Off	Mixed
Human	347	363	398
ft-2tags	304	339	348
syll-rescore	335	370	374

Table A.8: Perplexity on a general Spanish language model for human dubbing and two MT outputs.

To account for the difference in BLEU, we compute perplexity of the human dubbing and NMT outputs on a 5-gram language model (Heafield et al., 2013) trained on a corpus of general Spanish language (Cardellino, 2019) (Table A.8). A lower perplexity and lower BLEU score for On suggests that there is increased naturalness for on-screen human dubbing, imposed by the need to create realistic dialogues, in line with the factors above. This results in creative translations that rely less on the lexical choice, structure and word order of the source text; NMT systems normally lack this creativity. Indeed, for human dubbing the conformity to the target language norms has been claimed to be more important than following the source text structure or generating a perfect lip-sync (Chaume, 2012; Pavesi, 2008). On the other hand, a higher BLEU and higher perplexity for Off indicate less creative solutions, closer to the source, and therefore more predictable for the NMT system. With a low BLEU and high perplexity, Mixed is once again the most particular class, where the mix of dubbing strategies creates a linguistic hybrid.[10] This tendency is copied in the NMT outputs, which however are less surprising under all conditions, suggesting that they are more plausible as general language, but they do not reflect the orality of dubbed texts. The high perplexity

[10]On a language model of English, no difference in perplexity was found between On and Off, (0.1 points) while perplexity was higher for Mixed (10 points).

for OFF under *syll-rescore* is another indication that syllabic rescoring is harmful, since it leads to translations that are erroneous or not plausible in the target language.

A.4 Conclusions

We have explored for the first time the characteristics of on/off-screen dubbing from a computational perspective. To this end, we annotated the Heroes corpus, distinguishing between the two shades of dubbing. Our annotation is an important starting point towards fulfilling the multimodal requirements necessary for fully functional dubbing engines which will incorporate detection of visuals for identifying the shade and deciding whether to apply synchronisation (Nayak et al., 2020). Our corpus-based analysis showed that synchronisation is hardly discernible at the textual level, since distinctive text features to tell apart on/off-screen dubbing are particularly hard to find. A further analysis of the prosodic and phonetic features in the corpus did not reveal any differences in dubbing actors' voice modulation or lip sync strategies between on/off-screen dubbing, at least in the examined corpus. Despite the lack of distinctive features at the textual, prosodic and phonetic level in our dataset, NMT seems to suffer from a (still) elusive difference between on/off-screen dubbing, as witnessed by lower BLEU scores for ON and MIXED. We have further shown that isochrony constraints significantly detriment NMT performance for off-screen dubbing. Understanding the language of dubbing is still an open problem. We hope that our findings will set a first cornerstone towards the successful integration of NMT, and potentially ST, in dubbing.

Appendix B

Reproducibility checklist

System	
Task	MT, ST, fully automatic subtitling etc.
System Input/Output and granularity	video, audio, text; subtitles, sentences, documents, timestamps; pre-/post-processing
Evaluation criteria	
Type	Quality, conformity, accuracy, productivity, etc.
Aspect	Content (translation), form (length, segmentation), time (reading speed, spotting)
Data	
Language(s)	Language (variety)
Size	Clips, minutes, sentences, tokens, etc.
Format	Video, audio, plain text, xml, srt, etc.
Content	Genre, scripted/unscripted, single/multi-speaker
Participants	
Number	Per language pair/total
Type	Expert, how related to the authors, paid/offered course credit/voluntary, etc.
Expertise	Language skills, subtitling skills, PE skills, expertise with the tool/interface
How recruited	Agency, online, etc.
Remuneration	
Experimental setting	
Type	PE-based, rating, error annotation etc
Tool/user interface	Description and functionalities, video display, audio available
Requirements	Timestamps considered/ignored, format aspects and constraints
Scale or other rating instrument	Size, list or range of possible response values, how presented to evaluators
Information given to evaluators	Training, warm-up sessions, guidelines, instructions, verbatim question/prompt/etc.
Experimental conditions	in the lab, remotely, on crowdsourcing platform, etc

Appendix C

List of publications

Audiovisual Translation

- Alina Karakanta and David Orrego-Carmona. In press. Subtitling in Transition: the case of TED talks. *Translation in Transition: Bridging Human and Machine Intelligence*, American Translators Association series. John Benjamins.

- Alina Karakanta. 2022. Experimental research in Automatic Subtitling: At the crossroads between Machine Translation and Audiovisual Translation. *Translation Spaces*.

- Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. Post-editing in automatic subtitling: A subtitlers' perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 259–268, Ghent, Belgium. European Association for Machine Translation.

- Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. Towards a methodology for evaluating automatic subtitling. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 333–334, Ghent, Belgium. European Association for Machine Translation.

- Luisa Bentivogli, Mauro Cettolo, Alina Karakanta, Matteo Negri, and Marco Turchi. 2022. Extending the MuST-C Corpus for a Comparative Evaluation of Speech Translation Technology. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 361–362, Ghent, Belgium. European Association for Machine Translation.

- Sara Papi, Alina Karakanta, Matteo Negri, and Marco Turchi. 2022. Dodging the Data Bottleneck: Automatic Subtitling with Automatically Segmented ST Corpora. In *Proceedings of AACL 2022*. Online.

- Alina Karakanta, François Buet, Mauro Cettolo, François Yvon. 2022. Evaluating Subtitle Segmentation for End-to-End Generation Systems. In *Proceedings of LREC*. Marseille, France.

- Alina Karakanta, Sara Papi, Matteo Negri, Marco Turchi. 2021. Simultaneous Speech Translation for Live Subtitling: from Delay to Display. In *Proceedings of the 1st Workshop on Automatic Spoken Language Translation in Real-World Settings (ASLTRW)*. Virtual. August 2021. Association for Machine Translation in the Americas, pp. 35-48.

- Alina Karakanta, Marco Gaido, Matteo Negri, Marco Turchi. 2021. Between Flexibility and Consistency: Joint Generation of Captions and Subtitles. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*. Bangkok, Thailand (online): Association for Computational Linguistics, pp. 215–225.

- Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, Marco Turchi. 2021. Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on*

Natural Language Processing (Volume 1: Long Papers). Online. August 2021. Association for Computational Linguistics, pp. 2873–2887.

- Erick García Chávez and Alina Karakanta. 2021. A corpus-based comparison of prosodic features in on/off-screen dubbing. In Castagnoli, S., S. Bernardini, A. Ferraresi, M. Miličević Petrović (eds) *Using Corpora in Contrastive and Translation Studies Conference (6th Edition)*. Bertinoro (Italy), 9-11 September 2021.

- Alina Karakanta, Supratik Bhattacharya, Timo Baumann, Shravan Nayak, Matteo Negri and Marco Turchi. 2020. The Two Shades of Dubbing in Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 4327–4333.

- Shravan Nayak, Timo Baumann, Supratik Bhattacharya, Alina Karakanta, Matteo Negri, Marco Turchi. 2020. See me speaking? Differentiating on whether words are spoken on screen or off to optimize machine dubbing. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*, October 25–29, 2020, Virtual event, Netherlands. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425640

- Karakanta, Alina, Matteo Negri and Marco Turchi. 2020. Point Break: Surfing Heterogeneous Data for Subtitle Segmentation. In *Seventh Italian Conference on Computational Linguistics*, CLiC-It. Bologna, Italy

- Alina Karakanta. 2020. Subtitling in Transition: the case of TED talks. In *Book of Abstracts: Translation in Transition (TT5)*. Kent State University. October 15-17.

- Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. 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.

- Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. Towards Automatic Subtitling: Assessing the Quality of Old and New Resources. In *IJCoL – Italian Journal of Computational Linguistics* vol. 6, n. 1 June 2020.

- Alina Karakanta, Matteo Negri and Marco Turchi. 2020. MuST-Cinema: a Speech-to-Subtitles Corpus. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May 13-15 2020.

- Alina Karakanta, Matteo Negri, Marco Turchi. 2019. Are Subtitling Corpora Really Subtitle-like? In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-It)*, November 2019, Bari, Italy.

Translation studies

- Heike Przybyl, Alina Karakanta, Katrin Menzel, and Elke Teich. 2022. Exploring linguistic variation in mediated discourse: translation vs. interpreting. In *Empirical Investigations in the Forms of Mediated Discourse at the European Parliament.* Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska and Silvia Bernardini (eds). Language Science Press.

- Elke Teich, Jose Martinez Martinez and Alina Karakanta. 2021. Translation, Information Theory, and Cognition. In *The Routledge Handbook of Translation and Cognition.* Fabio Alves, Lykke Arnt Jakobsen (eds). London: Routledge, https://doi.org/10.4324/9781315178127

- Alina Karakanta, Heike Przybyl and Elke Teich. 2021. Exploring variation in translation with probabilistic language models. In *Corpora in*

Translation and Contrastive Research in the Digital Age: Recent advances and explorations. Julia Lavid-López, Carmen Maíz-Arévalo and Juan Rafael Zamorano-Mansilla (eds). Benjamins Translation Library 158, https://doi.org/10.1075/btl.158.12kar, pp. 308–323.

Low-resource languages

- Chao-Hong Liu, Alina Karakanta, Audrey N. Tong, Oleg Aulov, Ian M. Soboroff, Jonathan Washington & Xiaobing Zhao. 2021. Introduction to the second issue on machine translation for low-resource languages. *Machine Translation* 35, 1–2. https://doi.org/10.1007/s10590-021-09265-1

- Chao-Hong Liu, Alina Karakanta, Audrey N. Tong, Oleg Aulov, Ian M. Soboroff, Jonathan Washington & Xiaobing Zhao. 2020. Introduction to the Special Issue on Machine Translation for Low-Resource Languages. *Machine Translation* 34, 247–249. https://doi.org/10.1007/s10590-020-09256-8

- Alina Karakanta, Atul Kr. Ojha, Chao-Hong Liu, Jade Abbott, John Ortega, Jonathan Washington, Nathaniel Oco, Surafel Melaku Lakew, Tommi A Pirinen, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao, eds. 2020. *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*. Suzhou, China: Association for Computational Linguistics.

- Atul Kr. Ojha, Valentin Malykh, Alina Karakanta, and Chao-Hong Liu. 2020. Findings of the LoResMT 2020 Shared Task on Zero-Shot for Low-Resource languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*. Suzhou, China: Association for Computational Linguistics, pp. 33–37.

- Surafel M. Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, Marco Turchi. 2019. Adapting Multilingual Neural Machine Translation

to Unseen Languages. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT)*, November.

- Alina Karakanta, Atul Kr. Ojha, Chao-Hong Liu, Jonathan Washington, Nathaniel Oco, Surafel Melaku Lakew, Valentin Malykh, and Xiaobing Zhao, eds. 2019. *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. Dublin, Ireland: European Association for Machine Translation.