

This item is the archived peer-reviewed author-version of:

Preparing and comparing subtitles for quasi-experimental and experimental research in audiovisual translation studies

Reference:

Van Hoecke Senne, Schrijver Iris, Robert Isabelle.- Preparing and comparing subtitles for quasi-experimental and experimental research in audiovisual translation studies

Translation spaces : a multidisciplinary, multimedia, and multilingual journal of translation - ISSN 2211-372X - 11:1(2022), p. 113-133

Full text (Publisher's DOI): <https://doi.org/10.1075/TS.21038.VAN>

To cite this reference: <https://hdl.handle.net/10067/1892600151162165141>

Preparing and comparing subtitles for quasi-experimental and experimental research in audiovisual translation studies

Senne M. Van Hoecke, Iris Schrijver & Isabelle S. Robert
University of Antwerp

Empirical research on cognitive processing in AVT has been on the rise in recent years. A number of overarching works have recommended more standardised approaches and methodological frameworks to contribute to more streamlined, replicable, reproducible and valid future AVT research. To date, the issue of comparability of research materials (e.g., clips, subtitle tracks, comprehension questionnaires) and, more specifically, how to achieve comparability in quasi-experimental and experimental studies, particularly those involving repeated measures, has received little attention. This paper aims to address this knowledge gap by proposing a common-sense ten-step preparatory process for quasi-experimental and experimental subtitling studies. This preparatory process has previously been used in the S4AE project. The paper will focus on the final four steps, consisting of the preparation and comparison of multiple subtitle tracks. These steps were conceptualized taking into account the present research on subtitle parameters and the obstacles encountered while preparing comparable subtitle tracks.

Keywords: audiovisual translation (AVT), subtitling, methodology, cognition, comparability

1. Introduction

The need for subtitling is growing rapidly following the increased importance of overall accessibility, inclusivity and equity, the commercial pressure to reach larger, multilingual and multicultural audiences, and legal measures, such as the recent EU Accessibility Act and the renewed EU Audiovisual Media Service Directive in Europe. Consequently, research into subtitling and audiovisual translation (AVT) in general has never been more relevant. Over the years, studies

into AVT has examined a wide range of topics, but there is one relatively new focus that might be of great interest for the sudden surge in practical use of subtitling: empirical research on cognitive processing in AVT, or what Díaz Cintas (2020) calls cognitive and empirical AVT studies. Such empirical research can allow us to effectively test the impact of new practices and verify old assumptions and theories, provided it is scientifically sound, replicable and reproducible. In view of this scientific robustness, Orero et al. (2018) recommend standardised experimental approaches and methodological frameworks, which are, however, still largely missing in the field. There is a body of overarching works that list various methodologies for experimental AVT reception research (e.g., Doherty & Kruger, 2018; Kruger et al., 2016; Kruger, Szarkowska, & Krejitz, 2015; Orero et al., 2018). These position papers refer to many previously conducted AVT studies and recommend approaches, measurement tools and research designs. While these papers can be used as guidelines for more streamlined future AVT research, little attention is devoted to comparability of research materials, e.g., subtitles, video clips or comprehension test questions. The production of comparable materials is not addressed in detail in these papers, with mentions of comparability being limited to “If various fragments are compared, they should be similar in terms of complexity, speech rate, genre, etc. so as not to create confounding variables” (Orero et al., 2018, 112). Such quasi-experimental or experimental subtitling studies using several tasks and/or measurements in time can provide valuable insight into AVT, provided they are carefully thought out and meticulously prepared (Van Hoecke, Schrijver, & Robert, 2022).

This article proposes a common-sense ten-step process to prepare a quasi-experimental or experimental subtitling study that involves multiple conditions, tasks and/or measurements in time, requiring comparable research materials (e.g., subtitle tracks or clips). Step 1 to 6 of this process, which consist of preparing and comparing materials (e.g., comprehension tests, video fragments, etc.), and validating the materials have been previously discussed in Van Hoecke et al. (2022). In this article, we will focus specifically on steps 7 to 10, which concern the process of preparing multiple comparable interlingual and intralingual subtitle tracks. In these steps, we have given priority to “comparability over quality”.

By way of introduction, the article first draws up a theoretical framework in Section 2, which gives an overview of subtitle parameters that are expected to be relevant for the production of similar subtitles. As the ten-step process was developed within the Subtitles for Access to Education (S4AE) project, Section 3 first contextualises the project and briefly discusses its methodology. Section 4 then continues with the subtitling process, after which the paper concludes with key points from the process in Section 5.

2. Theoretical framework

To ensure the quality of subtitles, scholars like Karamitroglou (1998) and Ivarsson and Carroll (1998) prescribed general subtitling guidelines for practitioners. Though these works were not based on empirical data or scientific research, they are still widely seen as seminal in the profession. Such guidelines do not, however, address language-specific issues. Consequently, some broadcasting companies, businesses or streaming services can be seen developing their own adapted guidelines (e.g., BBC or Netflix). In view of this article's aim, a number of conventions were selected that are of great importance for subtitling (Gottlieb, 2012) and that have direct implications for comparability between different subtitle tracks, namely (1) reading speed; (2) reduction; (3) segmentation; and (4) linguistic complexity.

The first key component is reading speed, also referred to as subtitle speed or presentation rate. Reading speed, which is generally expressed as characters per second (CPS) or words per minute (WPM), is defined as the time an average viewer of a particular audience needs to comfortably read a full two-line subtitle. A full two-line subtitle depends on the maximum CPL. What this exact maximum is, however, is a matter of contention. D'Ydewalle and his colleagues tend to adhere to a maximum of 32 characters and spaces per line in a number of their studies (D'Ydewalle & De Bruycker, 2007; D'Ydewalle, Van Rensberger, & Pollet, 1987), Karamitroglou (1998, 2) mentions "around 35 characters" and Kruger, Hefer and Matthew (2014) adhere to a maximum of 37 CPL in one of their studies. It is clear that the maximum CPL varies across studies and, as Díaz Cintas and Remael (2014) mention, this is also the case in the media. Most standard television subtitles have a maximum of 37 CPL, in the movie industry the norm seems to be 40 CPL and there are cases where only 33 or 35 characters are allowed. The maximum CPL may fluctuate slightly, what does appear to be a common convention in subtitling is the so-called six-second rule. The six-second rule states that it should be possible to read a full two-line subtitle comfortably in six seconds and that shorter subtitles should be timed proportionally. Based on the study of Díaz Cintas (2003), the ideal CPL for the six-second rule is 72 CPL, which translates into a reading speed of subtitles of 12 CPS or approximately 144 WPM. This ideal reading speed has, however, started to receive some criticism lately. Gottlieb (2012), for example, states that the average reading speeds increased over time, especially in subtitling countries. Various commercial TV stations and the movie industry already adhere to a reading speed of 14–16 CPS and streaming companies, like Netflix, even going up to a maximum of 20 CPS for adult programs in English. While some studies show viewers are also able to cope with these faster subtitle reading speeds of up to 20 CPS (Szarkowska & Gerber-Moron, 2018;

Szarkowska & Bogucka, 2019), others reveal an increase in subtitle reading speed might lead to more words being skipped and the viewer skimming the subtitles instead of actually reading them (Kruger, Wisniewska, & Liao, 2022; Liao et al., 2021). The reading speed a particular viewer is capable of is likely to be influenced by the degree of habituation to subtitles and the overall language proficiency as well. With this in mind, it is essential to consider the intended audience during the production of comparable subtitles and the careful consideration of the subtitle reading speed for quasi-experimental and experimental studies. In light of this, Fresno and Sepielak (2020) advise to not only consider the average of the subtitle speed for all subtitles, but also the range of speeds.

The second component one needs to consider when testing or ensuring comparability of subtitle tracks is reduction. To cope with the time-space constraints of subtitling, reduction is considered essential (Gottlieb, 2012). Díaz Cintas and Remael (2014) distinguish two types of reduction: total reduction, i.e., deleting irrelevant information, and partial reduction, i.e., reformulating the message. The amount of reduction in a subtitle also dictates the type of subtitles, namely edited, i.e., content is reduced and simplified, verbatim, i.e., all utterances are included, and standard, i.e., content is slightly edited, subtitles. While the discussion of which is better or more inclusive is very much alive in AVT research (Romero-Fresco, 2009; Szarkowska et al., 2011), it is less relevant for this article. However, what is important to keep in mind is that there are different reading patterns and visual attention distributions for each type. A study by Szarkowska et al. (2011), for example, revealed that viewers spent more time watching the image with edited or standard subtitles than with verbatim subtitles. Verbatim subtitles, on the other hand, were revealed to generally be read faster. The amount of reduction is thus expected to be highly relevant regarding the comparability of different subtitle tracks.

The third component is segmentation. Segmentation takes place on two levels, namely subtitle level, i.e., segmenting over several subtitles, and line level, i.e., segmenting over several lines, also called (subtitle) line-breaks. For both types, the common rule seems to be that each segment, line or subtitle, should ideally be semantically and syntactically self-contained (Díaz Cintas & Remael, 2014, 172; Ivarsson & Carroll, 1998) and “should appear segmented at the highest syntactic nodes possible” (Karamitroglou, 1998, 6). If a sentence does not fit into a single subtitle line, this sentence should be parsed to see which is the most complete syntactical and semantical part that can fit into a single line. A line-break or, for long sentences, segmentation over multiple subtitles at arbitrary, often less coherent points is expected to disrupt reading and be more challenging for the viewer (Perego, 2008, 214). Segmentation plays a significant role in the readability of subtitles. For comparability, it is therefore key that the segmentation in all

subtitle tracks is similar. This does not necessarily mean that the subtitle tracks should have optimal segmentation, but that they should have equal amounts of sub-optimally segmented subtitles, optimal segmented subtitles, etc.

The fourth and last component to consider is the subtitles' linguistic complexity. Evidently, the reduction and segmentation of subtitles influence the final complexity of the subtitles (Perego, 2008; Szarkowska et al., 2011), but the syntactical and lexical complexity also play a vital role. Syntactic complexity has an influence on reading time (Clifton, Staub, & Rayner, 2007), which can be of importance considering the dynamic and fleeting nature of subtitles. With regard to lexical complexity, viewers spend more time on less frequent and more complex words than on frequently occurring words, indicating more effort is required to read lexically complex subtitles (Moran, 2012). With regard to the comparability of subtitle tracks, it is, of course, essential that the lexical and syntactical complexity is relatively similar. For the most part, this complexity will originate from the audiovisual material, which implies that if the material was tested beforehand without subtitles and the complexity was found to be similar, it is more likely that the subtitles will be comparable in this regard as well. However, this may not always be so simple. For edited subtitles, for example, the original is regularly simplified and reduced. Regardless of the source, the degree of simplification may differ across various subtitle tracks disrupting the comparability between them. Another example is the comparative complexity of intralingual and interlingual subtitles. While it might be more straightforward to stick to similar terms and syntactical structures for two languages of the same family, e.g., Dutch and German, it may be more complex for, say, Dutch and Chinese. To our knowledge, there are no clear guidelines on how to produce subtitles of similar syntactical and semantical complexity. Pedersen (2017) proposes the FAR model to assess subtitle quality and Díaz Cintas and Remael (2014) devote attention to the translation process and the transfer of register, style, grammar and lexicon from the original to both interlingual or intralingual subtitles, but both are, of course, more concerned with the quality of the end-product and less with the comparability between separate subtitle tracks. Regardless of the lack of guidelines, linguistic complexity can and should be carefully considered when producing similar subtitle tracks. Practical testing of the subtitle tracks may also shed light on the matter, as illustrated in the following sections.

3. Project background

The ten-step process we discuss in this article and proposed in a previous article (Van Hoecke et al., 2022) was developed within the S4AE project. As this project

will be used as an example in the discussion of the ten-step process, the project's background and aims will first be briefly summarized below.

The project wishes to examine the effects of subtitles on the cognitive load, i.e., the load imposed on a person to complete a task to a certain level, and comprehension of students in an L2 English lecture. It follows a mixed model design that revolves around a central within-subject component. In this design, Dutch (Flemish) students view three different recorded EMI (English as a Medium of Instruction) lectures on philosophy (named P, R, and T).¹ The lectures are provided in three conditions: (1) with intralingual (English) subtitles; (2) with interlingual (Dutch) subtitles; and (3) without subtitles. The viewing of the lectures takes place in an eye tracking laboratory, which allows us to monitor the students' eye movements, measure cognitive load and assess subtitle reading using the Reading Index for Dynamic Texts (RIDT; Kruger & Steyn, 2014). After each lecture the students are required to fill out a psychometric questionnaire on cognitive load from Leppink and van den Heuvel (2015) and a comprehension test. The use of both psychometric questionnaires and eye tracking allows us to assess cognitive load and triangulate the data of both measures, as recommended by Orero et al. (2018). To measure retention, students are asked to complete the comprehension test again one month after the experiment. The collected data are subsequently correlated with the students' biographical data and language proficiency, which is tested one month prior to the experiment.

4. The ten steps

In a previous article (Van Hoecke et al., 2022) we present a ten-step process to ensure the comparability of materials used in quasi-experimental and experimental subtitling studies that involve multiple conditions, tasks and/or measurements in time. The ten steps are as follows:

1. Careful preparation of materials
2. Content and feature analyses
3. First pilot study
4. Reevaluation
5. Optimization
6. Second pilot study
7. Production of comparable subtitles

1. Henceforth, the lectures are named P, R and T as the topic of the lectures are Thomas Piketty, Jean-Jacques Rousseau and Alexis de Tocqueville, respectively.

8. Subtitle analyses
9. Third pilot study with subtitles
10. Finalisation of materials

The first six steps were based on two pilot studies with 75 and 50 participants, respectively, and discussed in Van Hoecke et al. (2022). The present article is based on two more studies with 7 and 6 participants, respectively, and discusses the preparation and comparison of the subtitles (steps 7–10), taking into account relevant research and the theoretical framework presented in Section 2. For clarity purposes, we will first briefly summarize the first six steps.

4.1 The first six steps

To ensure the validity and strengthen the foundations of a quasi-experimental and experimental subtitling study (especially those involving repeated measures), meticulous preparation is required. The ten-step preparatory process we present illustrates a number of obstacles and key elements that we have encountered in the S4AE project (see Section 3) and may serve as a source of inspiration for similar future research. The process is structured in such a way that it gradually introduces and tests the relevant materials for the eventual main study. This is also what is done in the first six steps.

In step 1, the initial materials, which in the case of the S4AE project were the three lectures and the three comprehension tests, are prepared. For a quasi-experimental and experimental study, it is crucial to take into account where a lack of comparability between distinct videos, tools of measurement, etc. could influence the results.

After this initial preparation, step 2 dictates the prepared materials to be analysed before any field-testing is done. Conducting experiments is time-consuming and correction of any issues in the materials that can be found and eliminated beforehand is warranted. Only after the analyses show no major flaws in the materials and the researcher or research team is convinced the materials (and their comparability) is suited for testing, the next step can be taken.

In step 3 the initial materials are tested in practice. It is important to not add too many experimental components, e.g., subtitles, audiovisual source material or post hoc tests, just yet, because, if the results are skewed, it is easier to identify the cause with a small number of components. Furthermore, the process dictates a gradual increase in components to assure comparability and validity for each separate component.

After this first test, step 4 involves analysis of the data in which the focus lies on finding issues that might originate from the experimental materials. For exam-

ple, in the case of the S4AE project, if the data show that participants score significantly higher on one of the comprehension tests, it is possible that this one test or the corresponding lecture is easier than the others, and thus not comparable. It is possible that there are no issues with the materials, in which case the materials do not necessarily need to be optimized. However, if needed, there are various ways to optimize the materials without having to start anew, for example by coding not-comparable comprehension tests using the Item Response Theory (Van Hoecke et al., 2022). This is done in step 5.

Regardless of issues found in step 4 and changes made in step 5, we recommend a second test of the materials (step 6) to ensure no chance-based or sample-related errors. If the data from this second test are promising, the next key component can be added and tested, namely the AVT.

4.2 Step 7: Production of comparable subtitles

The production of comparable subtitles should be as much a careful and thorough process as the production or selection of the visual materials and tools of measurement (step 1). In the S4AE project three recorded EMI lectures are used since there are three conditions (no subtitles, English subtitles and Dutch subtitles). This means that for English and Dutch three comparable subtitle tracks need to be produced. This comparability needs to be present between all three English and all three Dutch subtitle tracks separately and between the English and Dutch subtitle track of each lecture.

Based on the theoretical framework discussed in Section 2, we composed a small set of practical rules that could provide an initial anchor for creating similar subtitle tracks (in the same language and between the two languages). Considering the density of the lectures, the expected language proficiency of the intended audience and recent research on subtitle speeds and word skipping for fast subtitles (Kruger, Wisniewska, & Liao, 2022; Liao et al., 2021), we set the maximum subtitling speed to 15 CPS. In terms of subtitle length, we allowed a maximum of 40 CPL. With these longer subtitles, we were also able to keep the reduction in the English subtitles to a minimum, which made the English subtitles near-verbatim yielding standard subtitles. Lastly, we preferred two-line subtitles over one-line subtitles, as it has been shown that viewers spend proportionally more time on one-line subtitles than on two-line subtitles (D'Ydewalle & De Bruycker, 2007). Additionally, it reduces the difference in total number of subtitle lines between the lectures. By using predominantly two lines, the same font type and size, near-verbatim/standard subtitles and by positioning the subtitles on the bottom centre for all lectures, the subtitle area and appearance were expected to be similar.

After setting up this small guideline, the English subtitles for all three lectures were produced first, since the source texts were written and recorded in English. The comparability of the lectures and lecture transcripts had already been tested and confirmed in the first two steps of the preparatory process. Near-verbatim/standard English subtitles were therefore expected to carry over this comparability. However, subtitles are still distinctly different from a static text, so extra attention was paid to segmenting and reducing the subtitles of all lectures similarly. It is recommended to analyse the subtitles in one language, in this case the English language, before continuing with the subtitles in the other language(s). For structural purposes, however, the analyses of the subtitles are discussed in Section 4.3.

After the English subtitles were produced and found to be comparable in the initial analyses, the Dutch subtitles were made. For the Dutch subtitles, we disregarded the original English transcript of the lecture and used the initial English subtitles as a template. The main aim here was to make the Dutch subtitles match the English in terms of complexity, but also retain the subtitle spotting, duration and segmentation, including sub-optimally segmented parts. Although quality is important, the main goal here was not optimal quality, but comparability of the subtitles in all aspects. In a final effort to make both languages comparable, we reevaluated the English subtitles based on the Dutch subtitles. If certain words or nuances were omitted, pronouns were used or slight segmentation shifts were made during the production of a Dutch subtitle, we corrected the corresponding English subtitle and applied the same changes, keeping in mind idiomatic structures in both languages. Reductions were required, which made the English subtitles slightly less verbatim, but strengthened the similarities between the subtitle tracks of both languages. For cases in which the Dutch segmentation was sub-optimal, the segmentation in English was also altered. This may have led to worse segmentation for the English subtitles, but, again, the goal was comparability of subtitles and, consequently, a more equal number of sub-optimally segmented subtitles.

4.3 Step 8: Subtitle analyses

To give an initial indication of the comparability of the subtitles, all six subtitle tracks were analysed. For each comparison within and between languages, we looked at the four components discussed in Section 2.

The first component concerned the reading speed. In this analysis, we included the separate parameters such as the number of one-line and two-line subtitles, CPL, CPS and subtitle duration. As can be seen in Table 1, the mean CPL and mean subtitle duration was relatively similar for all six subtitle tracks,

implying similar visual presence of the subtitles on screen. The mean CPL was around 23 to 25 characters. The Dutch subtitles always had a marginally higher mean CPL. Lecture T featured slightly more CPL and longer durations than the other two lectures, but the overall mean CPS was similar for all lectures at approximately 12.5 CPS. More importantly, the variability in CPL and CPS as measured by the standard deviation is comparable both between lectures and between languages, indicating that the CPL and particularly the CPS remains relatively constant throughout the lecture.

Table 1. Subtitle parameters*

	Lecture P		Lecture R		Lecture T	
	ENG	DU	ENG	DU	ENG	DU
Total number of subtitle lines	101	101	103	103	96	96
(1-line/2-line)	(6/95)	(6/96)	(3/100)	(3/100)	(5/91)	(5/91)
Mean CPL	23.58	24.24	23.33	23.63	25.11	25.81
Std. dev. CPL	7.10	7.81	7.89	8.23	7.84	8.27
Mean CPS	11.92	12.23	12.49	12.61	12.15	12.49
Std. dev. CPS	1.96	1.88	1.47	1.78	1.78	1.90
Mean subtitle duration	3.88	3.88	3.68	3.68	4.01	4.01
Std. dev. subtitle duration	1.137	1.137	1.004	1.004	0.974	0.974

* The lowest p -value found in Mann-Whitney U tests comparing the CPL, CPS and subtitle duration of subtitle tracks between languages was $p=0.112$ for the English and Dutch CPS in Lecture T.

The second component revolved around the reductions that were made during the production of the subtitles. Using Díaz Cintas and Remael's (2014, 151–171) classification of condensations and reformulations, the subtitle tracks were analysed and compared. We first looked at the English subtitles only. Here we saw a total of 14 reductions for lecture P and 19 each for R and T. Of these reductions, there were 5 total reductions/omissions for P, 7 for R and 6 for T. It is important to note that most omissions were limited to single words, such as adverbs or adjectives. Some examples are 'quite simple' > 'simple' in P or 'is commonly referred to' > 'is referred to' in R. Considering that these omissions are often only single words and that their frequency is similar across the three English subtitle tracks, these were not considered an issue. Regarding the partial reductions, these were generally also limited to the use of pronouns instead of full names, or of shorter synonymous words. Only for lecture T a subtitle could be observed that changed the form of the original soundtrack entirely: 'This is not to say that there are no social-economic classes in a democracy. Of course there are.' > 'However, there are

social-economic / classes in a democracy, of course.' While this was the only subtitle track to include such a relatively major change, it was expected not to influence the results, since the comprehension test did not contain a question about this specific piece of information. Additionally, the same syntactical structure was used in the Dutch subtitle, which meant that interlingual comparability was not problematic. We then analysed the Dutch subtitles. As the English subtitles were used as a template for the production of the Dutch subtitles, all initial reductions made would also be included in the Dutch subtitles. If an additional reduction was necessary for the Dutch subtitle, we tried to further reduce the English subtitles, but this was not always possible without making significant structural changes. Consequently, there were still minor differences in reductions between the Dutch and English subtitles. For lecture P, we highlighted 23 reductions, 2 of which were omissions/additions, R had 22 reductions with 3 omissions/additions and T had 21 reductions, 5 of which were omissions/additions (multiple reductions could occur within one subtitle). The partial reductions that we highlighted concerned small shifts, which most commonly were changes in word class, e.g., 'exceeds' > *is groter* ['is larger'] in P, or 'In studying' > *Tijdens de studie van* ['During the study of'] in T, changing passive to active voice or vice versa, e.g., 'is constituted by' > *bestaat uit* ['consists of'] in R, and a change in subject, e.g., 'It initiates the tragedy' > *Zo begint het drama* ['The tragedy starts with this'] in R or 'It was up to the nations of his day' > *De naties van zijn tijd moesten* ['The nations of his day had to'] in T. After careful consideration, we expected these differences not to significantly influence the results in future experiments.

For the third component, the segmentation and line-breaks of all subtitle tracks were analysed. We will first discuss the changes in line-breaks between the two languages. For lecture P, we observed a total of 12 shifts in line-breaks comparing the English and the Dutch subtitle track; for R 25 shifts; and for T 12 shifts. To evaluate whether these shifts, and the relatively large number of them in R, would not influence the comparability of the subtitle tracks, we categorized the line-breaks based on what changed (and potentially why it changed). The majority of these shifts seemed to be based on three principles. Firstly, in Dutch the verb is sometimes placed at the end of a sentence while it generally follows the subject in English. To build idiomatic structures, the verb was thus frequently placed in the second subtitle line instead of the first in Dutch. These shifts were never considered problematic, only when they occurred between two subtitles, i.e., the viewer only received the main verb at a later point in time for one of the two subtitle languages. A second reason for these shifts, which caused a number of shifts especially for R, was the change in location of the negation. Whereas the negation generally accompanies the verb at the front of the sentence in English, it is more idiomatic in Dutch to place the negation at the end. A third reason was a change

in word order often as a result of a partial reduction. In cases where the word class was changed or the passive voice was used instead of the active voice, the sentence structure changed which often also resulted in a different line-break. These shifts in line-breaks could not always be matched in both languages, so in some cases there were differences in line-breaks that were sub-optimal in one language but not in the other (see Table 2). Because we expect the large majority of these shifts in line-breaks to have no influence, we conclude that the line-breaks are sufficiently similar and of comparable quality and thus fit for the within-subject experiment. As for segmentation between two subtitles, we observed 6 shifts in lecture P, 2 in R and 4 in T. The reasons for these segmentation shifts were generally the same as the ones we mentioned above. In terms of comparability between the subtitle tracks in one language, we also looked at the total number of sub-optimally segmented subtitles (both line-breaks and segmentation) (see Table 2). We observed 7 sub-optimal line-breaks in lecture P, 2 in R and 4 in T. Lecture T also had one case where the segmentation between subtitles was sub-optimal. In most, if not all, cases the direct cause of this sub-optimal segmentation is the syntactic nodes being too long in either of the languages to fit on one subtitle line only. This implies that more than 40 characters are needed for the entire phrase, word group, etc. Because we attempted to match the segmentation between both languages in every case, some of these sub-optimally segmented subtitles could have been prevented in one language, but, maximizing comparability, this was not done.

Table 2. Subtitle segmentation

		Shifts between English and Dutch			Both languages
		Unproblematic	Unfavourable for English	Unfavourable for Dutch	Sub-optimal
Lecture P	Line-break	9	2	1	7
	Segmentation	4	0	2	0
	Total	15	2	3	7
Lecture R	Line-break	19	4	2	2
	Segmentation	1	0	1	0
	Total	20	4	3	2
Lecture T	Line-break	10	0	2	4
	Segmentation	0	2	2	1
	Total	10	2	4	5

Since some reductions that were made directly resulted in shifts in segmentation as well, we also checked how many “exact” subtitle matches, i.e., no reductions or shifts in segmentation, between both languages were present for each lecture. For lecture P, 60 of 101 subtitles (59.41%) are considered exact matches in English and Dutch, for R 62 of 103 (60.19%) are and for T 57 of 96 (59.38%) are. The number of altered subtitles between the subtitle tracks of each lecture is therefore very similar.

The last component concerned the linguistic, i.e., syntactical and lexical, complexity of the subtitles. We expected the linguistic complexity of the English subtitles to mirror the complexity of the original texts as the English subtitles were mostly verbatim, i.e., matching the original soundtrack. The linguistic complexity of these original texts had been shown to be comparable in step 1 of the process (as reported on in Van Hoecke et al. (2022) and using Perego, Del Missier and Stragà (2018) as a source of inspiration). However, a more objective assessment is recommendable, especially since the argument put forward above does not apply for the Dutch subtitles. Objectively assessing the lexical or syntactical complexity of subtitle tracks is a difficult endeavour. There are readability measures to evaluate the complexity of a static text (e.g., Flesch Reading Ease Formula), but the use of standard readability formulae and readability indices should be warranted when analysing linguistic subtitle complexity. Readability indices have been already criticized for their inaccuracy for shorter texts (Kidwell, Lebanon, & Collins-Thompson, 2011), let alone using them for single sentences or separate clauses. One way of avoiding this challenge when analysing syntactic complexity of subtitles would be to bring the subtitles together in a single text and apply measures like average sentence length or number of clauses. This, however, disregards the segmentation of subtitles. Applying these measures to subtitles, which may consist of one or two sentences, but also just a part of a sentence due to segmentation, will, in our view, therefore not yield any meaningful result. Basic indices to measure lexical complexity such as word length have also received considerable criticism. They have been found to be rather superficial, disregard the entire text structure and overall cohesion and coherence and are not necessarily causally related to linguistic complexity (Kraf & Pander, 2009). Nevertheless, one way to measure lexical complexity in subtitles would be to measure word frequencies. This can be done using the SubtLex corpus for subtitle word frequencies. This corpus exists for multiple languages, e.g., Dutch (Keulers, Brysbaert, & New, 2010), British English (van Heuven et al., 2014) or American English (Brysbaert & New, 2009), and thus also allows the comparison of relative word frequencies across languages. To examine the comparability between the subtitle tracks in our study, both between two languages and within one language, we tokenized and lemmatized the subtitles. Subsequently,

we extracted the logarithmic word frequency scores from the Dutch SubtLex corpus and the American English SubtLex corpus (since American English spelling was used for the subtitles). We lemmatized the subtitles as it would give a more accurate indication of the word frequencies of the lemma itself. We also left out every name, year and number as these frequencies might skew the results. Two Kruskal Wallis tests revealed no significant differences between the English subtitle tracks for all three lectures, $H(2)=1.328$, $p=0.515$, or between the three Dutch subtitle tracks, $H(2)=1.203$, $p=0.548$. This suggests that the lexical complexity based on word frequency is comparable for the subtitle tracks in the same language. We then ran three Mann-Whitney U tests to compare the word frequencies between the English and Dutch subtitle tracks of each video. No significant differences were found comparing the English ($Mdn=4.437$) and Dutch ($Mdn=4.319$) tracks for lecture P, $U=279764$, $z=0.962$, $p=0.336$, $r=0.03$, and none were found when comparing the English ($Mdn=4.742$) and Dutch ($Mdn=4.403$) subtitles for lecture R, $U=301939$, $z=1.799$, $p=0.072$, $r=0.05$. A comparison of the English ($Mdn=4.488$) and Dutch ($Mdn=4.453$) tracks for lecture T, $U=288212.5$, $z=1.528$, $p=0.126$, $r=0.04$, did not yield any statistical difference either. These results suggest that the word frequencies of the Dutch and English tracks for each lecture are also comparable. Comparable syntactic complexity between the English and Dutch tracks is, however, less clear-cut. While an attempt was made to match the structure of the English subtitles, a few minor shifts were still present in all lectures. In turn, comparable segmentation and line-breaks between all subtitle tracks, resulting in relatively similar number of clauses and comparable structures, indicated a baseline syntactical comparability. One important difference in this case, however, is that, as mentioned before, the verb in Dutch is generally placed later on in the sentence. This may lead to a line-break being present between the subject or auxiliary verb and the main verb in Dutch, while they are next to one another in English.

4.4 Step 9: Third pilot study with subtitles

To verify the conclusions drawn from step 8, a pilot study was set up consisting of two small-scale experiments. The first experiment was conducted in October 2020 with 7 students from the 3rd-year of the BA in Applied Linguistics or 1st-year of the MA Interpreting or Translation at the University of Antwerp (only 6 students are considered in the analyses below as 1 participant was excluded based on an eye tracking ratio below 85%). The second experiment took place in March 2021 and included 6 students from the MA Linguistics and Literature or the MA Interpreting at the University of Antwerp. In both experiments, the students viewed all three lectures while being monitored with an SMI RED 250Hz eye

tracker, completed the biographical survey, psychometric questionnaires and the comprehension tests and were also interviewed after the experiment. However, in the first experiment (henceforth called ES) all lectures were subtitled in English and in the second (DS) they were subtitled in Dutch. As the groups in both experiments were too small to include between-group variables, such as the student's English proficiency or prior knowledge of the subject (philosophy), these were not included.

For each subtitle track, we collected three types of data. First, we measured the cognitive load using the validated psychometric questionnaire from Leppink and van den Heuvel (2015). This questionnaire consists of eight questions, in which the students had to rate the complexity of the subtitled lecture on a scale from 1 (low complexity) to 10 (high complexity). The first four questions concerned content complexity, providing insight into the perceived intrinsic load. The last four concerned instructional complexity, i.e., perceived extraneous load, and are expected to reveal the effects of subtitles. Second, we had comprehension scores, which could reveal if any of the subtitle tracks influenced comprehension more than the other tracks. Third, eye tracking data was collected to provide insight into differences in subtitle reading behaviour, which may also influence the cognitive load ratings and comprehension scores. In these eye tracking data we limit ourselves to fixation counts, average fixation durations and dwell times in the subtitles' areas of interest (AOI). These global measures are indicators of processing (Schotter & Rayner, 2012) and have been shown to be measures of cognitive load (Kruger & Doherty, 2016).

We analysed the cognitive load ratings, comprehension scores and eye tracking data in each experiment separately to assess the comparability of the subtitle tracks in each of the languages. In view of the small sample size, we consistently used Friedman's tests to assess within-subject differences. For future research, we recommend linear mixed models in which all subtitles can be treated as separate items in the design. This way, intrinsic variability is accounted for and differences can be more accurately measured. While this is something we intend to use in the main study of the project, we limited the comparability analyses to the Friedman's test as our sample size for this analysis is very limited and we still consider this sufficiently robust with the present goal in mind, namely examining comparability.

As shown in Table 3, no significant within-subject effects were found for the cognitive load ratings, the comprehension scores or the eye tracking variables. These findings indicate that, according to these data, there is no significant difference between the three subtitle tracks in English or in Dutch.

The comparison of the English and the Dutch subtitle tracks would seem like a logical next step. However, such a comparison – based on the differences

Table 3. Within-subject differences

		ES			DS		
		<i>df</i>	<i>Q</i>	<i>p</i>	<i>df</i>	<i>Q</i>	<i>p</i>
Cognitive Load	Total	2	0.609	0.738	2	4.000	0.135
	Intrinsic	2	0.300	0.861	2	5.304	0.070
	Extraneous	2	0.273	0.873	2	0.873	0.293
Comprehension	Scores	2	4.000	0.135	2	1.652	0.438
Eye Tracking	Fixation Count in AOI	2	0.333	0.311	2	2.333	0.846
	Mean Fixation Duration in AOI	2	5.333	0.069	2	5.333	0.069
	Dwell Time in AOI	2	1.000	0.607	2	1.333	0.513

in results of the cognitive load questionnaire, comprehension questionnaire and eye tracking measures – should be approached with caution. Such a comparison would rather study the effects instead of the comparability of the subtitle tracks in different languages. Moreover, when a significant difference is found using the same data collection methods as before, it is practically impossible to determine whether this difference was caused by the subtitle complexity or by other confounding factors, such as the matching or contrast between the soundtrack language and the subtitle language or participants' proficiencies, prior knowledge, and subtitle language preference. In terms of cognitive load, a difference in intrinsic load between the two languages would not be expected, as the change in language of the subtitle track is supposed to not have an influence on content complexity, but rather on instructional complexity (i.e., extraneous cognitive load). With regard to comprehension, any significant difference found between languages may very well be due to participants being more proficient or native speakers in one language, thus understanding more because of the subtitle language. This effect of subtitle language on comprehension has been shown in previous research (Lavour & Bairstow, 2011) and is also one of the research foci in the S4AE project. As for the eye tracking data, any difference found would not necessarily mean that there is a difference between the language tracks either. As previous studies have revealed (Kruger et al., 2014; Hefer, 2013), the reading behaviour of a viewer is not the same for intralingual and interlingual subtitling or for native and foreign language subtitling. For these reasons, statistical comparison of tracks in both languages is not warranted. Significant differences found between the two might not necessarily imply dissimilarities between subtitles, but may simply be caused by different reading patterns in different types (intralingual vs. interlingual subtitles) or different language subtitles.

4.5 Step 10: Finalisation of materials

Step 9 revealed that there were no differences between the subtitle tracks and that the subtitles were perceived to be of adequate quality, i.e., representative for actual subtitles. This means that for this concrete example of the S4AE project, adjustments to the subtitles or the other materials was not considered necessary. However, not all future studies may have this outcome. Therefore, a final optimization of the subtitles (the other materials should already be optimized) could be carried out in step 10 before a potential main study. Ideally, the altered subtitles should then be tested again in practice, but depending on the size of the changes made, this may not be deemed necessary.

5. Conclusion

While an increased number of cognitive AVT studies have been conducted in the past decades (Díaz Cintas, 2020), little attention is devoted to preparing and comparing materials for such studies. Meticulous preparation and practical testing of the research materials is important for any experimental study, but absolutely critical for AVT studies using repeated measures, e.g., multiple clips, multiple language tracks, several measurements in time. This paper builds on our previous proposal (Van Hoecke et al., 2022), in which we lay out a ten-step common-sense preparatory process for quasi-experimental and experimental AVT studies. This paper in particular demonstrates possible key points and obstacles in producing comparable subtitles for such a study. Fundamental is the concept of ‘comparability over quality’. In most cases, a subtitler would be concerned with subtitle quality and thus follow the proposed guidelines of optimal segmentation, reduction, terminology, editing, etc. In the case of a study using repeated measures, quality remains important, but it may be more interesting to reduce more than necessary in certain subtitles or have the line-break at a different, less optimal place if it would make the subtitles of different languages, but also of different clips in the same language, more similar overall. Ensuring the comparability between subtitles in the same language, and especially in different languages, is a complex process. In some cases, e.g., between two languages, it is near impossible to be completely certain there are no significant differences between the two. An effort should be made to examine the comparability in the preparatory phase, but one should keep in mind that these differences can still be accounted for using linear mixed models in the main study.

We hope that these ten steps, inspired by our personal experiences in the S4AE project, may be of use and inspiration for similar future AVT research. Evidently, the ten-step preparatory process may need to be slightly altered to fit specific research goals. Moreover, some of the steps can be refined, e.g., by running linear mixed models to account for the individual subtitles, even when we did not do so in this paper. We also acknowledge that the process of producing comparable subtitles for English and Dutch is most likely more straightforward than doing so for, say, English and Chinese. Nevertheless, this ten-step common-sense preparatory process has shown that, regardless of the complexity of preparing comparable subtitles, it is not impossible.

References

- Brysbaert, Marc, and Boris New. 2009. "Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English." *Behavior Research Methods* 41(4): 977–990. <https://doi.org/10.3758/BRM.41.4.977>
- Clifton, Charles, Adrian Staub, and Keith Rayner. 2007. "Eye movements in reading words and sentences." In *Eye movements: A window on mind and brain*, edited by Roger P.G. Van Gompel, Martin H. Fischer, Wayne S. Murray, and Robin L. Hill, 341–371. Amsterdam: Elsevier. <https://doi.org/10.1016/B978-008044980-7/50017-3>
- Díaz Cintas, Jorge. 2003. *Teoría y práctica de la subtitulación: Inglés-Español*. Barcelona: Ariel.
- Díaz Cintas, Jorge. 2020. "Audiovisual translation." In *The Bloomsbury Companion to Language Industry Studies*, edited by Erik Angelone, Maureen Ehrensberger-Dow, and Gary Massey, 209–230. London: Bloomsbury. <https://doi.org/10.5040/9781350024960.0014>
- Díaz Cintas, Jorge, and Aline Remael. 2014. *Audiovisual Translation: Subtitling*. Manchester: St Jerome. <https://doi.org/10.4324/9781315759678>
- Doherty, Stephen, and Jan-Louis Kruger. 2018. "The development of eye tracking in empirical research on subtitling and captioning." In *Seeing into screens: Eye tracking and the moving image*, edited by Tessa Dwyer, Claire Perkins, Sean Redmond, and Jodi Sita, 46–56. London: Bloomsbury. <https://doi.org/10.5040/9781501329012.0009>
- d'Ydewalle, Géry, and Wim De Bruycker. 2007. "Eye movements of children and adults while reading television subtitles." *European Psychologist* 12(3): 196–205. <https://doi.org/10.1027/1016-9040.12.3.196>
- d'Ydewalle, Géry, Johan Van Rensberger, and Joris Pollet. 1987. "Reading a message when the same message is available auditorily in another language: The case of subtitling." In *Eye Movements from Physiology to Cognition*, edited by J. Kevin O'Regan, and Araine Lévy-Schoen, 313–321. Amsterdam: Elsevier. <https://doi.org/10.1016/B978-0-444-70113-8.50047-3>
- Fresno, Nazaret, and Katarzyna Sepielak. 2020. "Subtitling speed in Media Accessibility research: Some methodological considerations." *Perspectives*. <https://doi.org/10.1080/0907676X.2020.1761841>


- Gottlieb, Henrik. 2012. "Subtitles: Readable dialogue?" In *Eye tracking in audiovisual translation*, edited by Elisa Perego, 37–82. Rome: Aracne.
- Hefer, Esté. 2013. "Reading first and second language subtitles: Sesotho viewers reading in Sesotho and English." *Southern African Linguistics and Applied Language Studies* 31(3): 359–373. <https://doi.org/10.2989/16073614.2013.837610>
- Ivarsson, Jan, and Mary Carroll. 1998. *Subtitling*. Simrishamn: TransEdit.
- Karamitroglou, Fotios. 1998. "A proposed set of subtitling standards in Europe." *Translation Journal* 2(2).
- Keuleers, Emmanuel, Marc Brysbaert, and Boris New. 2010. "Subtlex-NL: A new measure for Dutch word frequency based on film subtitles." *Behavior Research Methods* 42(3): 643–650. <https://doi.org/10.3758/BRM.42.3.643>
- Kidwell, Paul, Guy Lebanon, and Kevyn Collins-Thompson. 2011. "Statistical estimation of word acquisition with application to readability prediction." *Journal of the American Statistical Association* 106(493): 21–30. <https://doi.org/10.1198/jasa.2010.ap09318>
- Kraf, Roger, and Henk Pander. 2009. "Leesbaarheidsonderzoek: Oude problemen, nieuwe kansen [Readability research: Old problems, new opportunities]." *Tijdschrift voor Taalbeheersing* 31(2): 97–123. <https://doi.org/10.5117/TVT2009.2.LEES356>
- Kruger, Jan-Louis, and Stephen Doherty. 2016. "Measuring cognitive load in the presence of educational video: Towards a multimodal methodology." *Australasian Journal of Educational Technology* 32(6): 19–31. <https://doi.org/10.14742/ajet.3084>
- Kruger, Jan-Louis, Esté Hefer, and Gordon Matthew. 2014. "Attention distribution and cognitive load in a subtitled academic lecture: L1 vs. L2." *Journal of Eye Movement Research* 7(5): 1–15. <https://doi.org/10.16910/jemr.7.5.4>
- Kruger, Jan-Louis, María T. Soto-Sanfiel, Stephen Doherty, and Ronny Ibrahim. 2016. "Towards a cognitive audiovisual translology." In *Reembedding Translation Process Research*, edited by Ricardo Muñoz Martín, 171–193. Amsterdam: John Benjamins. <https://doi.org/10.1075/btl.128.09kru>
- Kruger, Jan-Louis, and Faans Steyn. 2014. "Subtitles and eye tracking: Reading and performance." *Reading Research Quarterly* 49(1): 105–120. <https://doi.org/10.1002/rrq.59>
- Kruger, Jan-Louis, Agnieszka Szarkowska, and Izabela Krejtz. 2015. "Subtitles on the moving image: An overview of eye tracking studies." *Refractory: A Journal of Entertainment Media* 25: 1–14.
- Kruger, Jan-Louis, Natalia Wisniewska, and Sixin Liao. 2022. "Why subtitle speed matters: Evidence from word skipping and rereading." *Applied Psycholinguistics* 43(1): 211–236. <https://doi.org/10.1017/S0142716421000503>
- Lavaur, Jean-Marc, and Dominique Bairstow. 2011. "Languages on the screen: Is film comprehension related to the viewers' fluency level and to the language in the subtitles?" *International Journal of Psychology* 46(6): 455–462. <https://doi.org/10.1080/00207594.2011.565343>
- Leppink, Jimmie, and Angélique van den Heuvel. 2015. "The evolution of cognitive load theory and its application to medical education." *Perspectives on Medical Education* 4(3): 119–127. <https://doi.org/10.1007/s40037-015-0192-x>
- Liao, Sixin, Lili Yu, Erik D. Reichle, and Jan-Louis Kruger. 2021. "Using eye movements to study the reading of subtitles in video." *Scientific Studies of Reading* 25(5): 417–435. <https://doi.org/10.1080/10888438.2020.1823986>
- Moran, Siobhan. 2012. "The effect of linguistic variation on subtitle reception." In *Eye tracking in audiovisual translation*, edited by Elisa Perego, 37–82. Rome: Aracne.

- Orero, Pilar, Stephen Doherty, Jan-Louis Kruger, Anna Matamala, Jan Pedersen, Elisa Perego, Pablo Romero-Fresco, Sara Rovira-Esteva, Olga Soler-Vilageliu, and Agnieszka Szarkowska. 2018. "Conducting experimental research in audiovisual translation (AVT): A position paper." *The Journal of Specialised Translation* 30: 105–126.
- Pedersen, Jan. 2017. The FAR model: Assessing quality in interlingual subtitling. *The Journal of Specialised Translation* 28: 210–229.
- Perego, Elisa. 2008. "Subtitles and line-breaks: Towards improved readability." In *Between Text and Image: Updating research in screen translation*, edited by Delia Chiaro, Christine Heiss, and Chiara Bucaria, 211–223. Amsterdam: John Benjamins. <https://doi.org/10.1075/btl.78.21per>
- Perego, Elisa, Fabio Del Missier, and Marta Stragà. 2018. Dubbing vs. Subtitling: Complexity matters. *Target* 30(1): 137–157. <https://doi.org/10.1075/target.16083.per>
- Romero-Fresco, Pablo. 2009. "More haste less speed: Edited versus verbatim respoken subtitles." *Vial-vigo International Journal of Applied Linguistics* 6: 109–133.
- Schotter, Elizabeth R., and Keith Rayner. 2012. "Eye movements in reading: Implications for reading subtitles." In *Eye tracking in audiovisual translation*, edited by Elisa Perego, 83–104. Rome: Aracne.
- Szarkowska, Agnieszka, and Lidia Bogucka. 2019. "Six-second rule revisited: An eye tracking study on the impact of speech rate and language proficiency on subtitle reading." *Translation, Cognition & Behavior* 2(1): 101–124. <https://doi.org/10.1075/tcb.00022.sza>
- Szarkowska, Agnieszka, and Gerber-Morón, Oliva. 2018. "Viewers can keep up with fast subtitles: Evidence from eye movements." *PLoS One* 13(6). <https://doi.org/10.1371/journal.pone.0199331>
- Szarkowska, Agnieszka, Izabela Krejtz, Zuzanna Klyszejko, and Anna Wieczorek. 2011. "Verbatim, standard, or edited? Reading patterns of different captioning styles among deaf, hard of hearing and hearing viewers." *American Annals of the Deaf* 156(4): 363–378. <https://doi.org/10.1353/aad.2011.0039>
- van Heuven, Walter J. B., Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. "Subtlex-UK: A new and improved word frequency database for British English." *Quarterly Journal of Experimental Psychology* 67(6): 1176–1190. <https://doi.org/10.1080/17470218.2013.850521>
- Van Hoecke, Senne M., Iris Schrijver, and Isabelle S. Robert. 2022. "Methodological preparation of a within-subject audiovisual cognition, reception and perception study." *Journal of Audiovisual Translation*, 5(1): 94–128. <https://doi.org/10.47476/jat.v5i1.2022.163>

Address for correspondence

Senne M. Van Hoecke
Department of Applied Linguistics, Translators and Interpreters
University of Antwerp
2000 Antwerp
Belgium

senne.vanhoecke@uantwerpen.be

 <https://orcid.org/0000-0003-0519-576X>

Co-author information

Iris Schrijver
University of Antwerp
iris.schrijver@uantwerpen.be

Isabelle S. Robert
University of Antwerp
isabelle.robert@uantwerpen.be

Publication history

Date received: 13 September 2021
Date accepted: 21 June 2022
Published online: 19 July 2022