Title: Using linear mixed models to analyze data from eye-tracking research on subtitling.

Breno B. Silva
Institute of English Studies
University of Warsaw, Poland

David Orrego-Carmona
Department of English Languages and Applied Linguistics
Aston University, UK

Department of Linguistics and Language Practice
University of the Free State, South Africa

Agnieszka Szarkowska
Department of Modern Languages
University of Warsaw, Poland

Abstract

In this paper, we aim to promote the use of linear mixed models (LMMs) in eye-tracking research on subtitling. Using eye tracking to study viewers' reading of subtitles often warrants controlling for many confounding variables. However, even assuming that these variables are known to researchers, such control may not be possible or desired. Traditional statistical methods such as t-tests or ANOVAs exacerbate the problem due to the use of aggregated data: each participant has one data point per dependent variable. As a solution, we propose the use of LMMs, which are better suited to account for a number of subtitle and participant characteristics, thus explaining more variance. We introduce essential theoretical aspects of LMMs and highlight some of their advantages over traditional statistical methods. To illustrate our point, we compare two analyses of the same dataset: one using a t-test; another using LMMs.

Keywords: eye tracking, subtitling, linear mixed-effects models, LMMs, traditional statistical methods, confounding variables, reading, audiovisual translation.

1. Introduction

In recent years, more and more audiovisual translation (AVT) researchers have adopted eye-tracking methods to understand how viewers process subtitled videos and how they are affected by the characteristics of subtitles(Doherty and Kruger 2018). Findings from such studies enable the creation of subtitles that better respond to users' needs and allow

researchers to better understand the cognitive processes behind audiovisual texts. Online data generated with eye tracking combined with offline measures such as questions to assess textual comprehension and task-induced cognitive load (Kruger and Doherty 2018) ultimately allow researchers to understand how viewers engage with subtitled content.

However, the reliability of many of the findings from eye-tracking research on subtitles could be improved with increased scientific rigour (Kruger 2018), including more powerful statistical analysis. One problem with existing eye-tracking research on subtitles is that data obtained from each subtitle (or word within subtitles) is often aggregated (e.g., averaged) before being analyzed statistically, which makes it impossible to examine sentence-level or word-level data. Linear mixed models (LMMs), promoted in this paper as a more robust method of statistical analysis to study the reading of subtitles, do not necessitate aggregated data, thus circumventing some of the limitations of other statistical methods.

This paper seeks to provide a brief introduction to LMMs, especially in the context of eye-tracking research on subtitles. (For a more general, step-by-step introduction on how to perform a linear mixed model analysis see Carson and Beeson 2013; Cunnings and Finlayson 2015.) Here, we aim to show that LMMs may be better suited than other more commonly used statistical methods to investigate the complexities involved in subtitle processing. This focus presupposes that researchers have already taken into account the various methodological issues that may affect cognitive translation studies, such as those described in Mellinger and Hanson (2017), Saldanha and O'Brien (2014), Gile (2016), and Balling and Hvelplund (2015).

We start by drawing attention to many variables that may influence the results of eye-tracking studies on subtitling. Next, we argue that more widely adopted statistical methods in AVT research such as *t*-tests, ANOVAs, and regression analyses (all part of the general linear model; see Field 2017) may be unable to effectively cope with these variables and the complexity of subtitle processing. Then, we introduce essential theoretical aspects of LMMs and underscore some of their advantages over other statistical techniques (although see Section 8 for a brief discussion on the limitations of the method). To illustrate our point with a practical example, we compare the impact of subtitling speed on mean fixation duration (MFD) through two statistical analyses: *t*-tests, representing more commonly used statistical methods, and LMMs, the method we seek to promote.

2. Variables to control for in eye-tracking research on subtitling

Subtitled audiovisual materials consist of several intertwining elements that may be difficult to control in experimental settings. Eye-tracking research is a useful example of this issue. Let us consider a relatively simple eye-tracking study that seeks to compare the effects of two subtitling speeds (the independent variable) on the mean fixation duration on the subtitles. In such a study, several confounding variables may influence the results, e.g., participants' proficiency, and the length and number of lines in the subtitles. Without controlling for such variables, it may be impossible to ascertain whether a difference in conditions (here, the two different speeds) may be attributed solely to the independent variable manipulated. To help raise awareness to this problem, below we address a number of relevant participant-, subtitle-, and word-related variables that may need to be controlled for in experimental settings.

In many eye-tracking study on subtitling, researchers must ensure that participants and subtitles are comparable across or within conditions. For instance, if the language of the film dialogues or the language of the subtitles is not the viewers' first language (L1), participants need to be matched in language proficiency (De Bruycker and d'Ydewalle 2003; Muñoz 2017; Peters 2019; Szarkowska and Bogucka 2019). Proficiency level must be controlled for because, even if it is not related to the research question, it may inadvertently influence the results. Similarly, familiarity with subtitling might also need to be controlled, as previous exposure to this type of audiovisual translation may affect reading patterns and typically confer on experienced viewers an advantage over inexperienced ones (Orrego-Carmona 2015). Other aspects that need to be controlled are participants' age and reading abilities, particularly when comparing viewers with varying levels of reading skills, such as children and adults (d'Ydewalle and De Bruycker 2007; Koolstra, Van Der Voort, and d'Ydewalle 1999), or deaf and hearing viewers (Cambra et al. 2014; Gerber-Morón, Szarkowska, and Woll 2018).

Subtitle characteristics also need to be comparable across conditions, as previous research has demonstrated. One example is the degree of text condensation, a variable that may be confounded with subtitling speed (Burnham et al. 2012; Szarkowska, Silva, and Orrego-Carmona 2021). The reason for such confusion is that, to maintain ecological validity, slow subtitles are condensed, reflecting industry practice to ensure subtitle-to-soundtrack synchronization (Romero-Fresco 2015; Szarkowska et al. 2011). Subtitles also need to be controlled for the number of lines and text segmentation (d'Ydewalle et al. 1991; Gerber-Morón and Szarkowska 2018; Rajendran et al. 2013). Other variables that may affect subtitle

processing include maximum line length (e.g., 40 characters per line) typeface and font size (see Díaz-Cintas & Remael, 2021 for a recent discussion).

Various linguistic characteristics of the subtitles, and particularly of the words within the subtitles, may also inadvertently affect the results. These include grammatical difficulty, idiomaticity, and textual cohesion (Moran 2009). Also, differences in word frequency may influence textual processing. For instance, high-frequency words are typically easier to process, as demonstrated by L2 reading research (e.g., Brysbaert and New 2009; Crossley, Cobb, and McNamara 2013; Peters 2019), and may induce fewer fixations, fewer regressions, and shorter mean fixation durations (Inhoff and Rayner 1986; Rayner et al. 2004). Word length and word type (e.g., nouns, adjectives, cognates, non-cognates, content, and function words) may all also affect subtitle processing and thus the results (Krejtz, Szarkowska, and Łogińska 2016; Liao et al. 2020). For example, psycholinguistic research has shown that longer words tend to be more difficult than shorter ones (Ellis and Beaton 1993); eye-tracking studies have found that longer words induce more fixations and less skipping than shorter words (K. Rayner et al. 2011). These results hold true both for studies on reading printed text as well as subtitles (Krejtz, Szarkowska, and Łogińska 2016). Furthermore, psycholinguistic studies investigating the reading of sentences have shown that cognates are processed faster than non-cognates (e.g., Dijkstra et al. 2010; Lemhöfer and Dijkstra 2004; Schwartz and Kroll 2006), whereas translation studies have demonstrated that subtitles containing a more literal translation are processed faster than the non-verbatim version (see Ghia 2012) Finally, all these textual factors, even if controlled for, may be confounded by the presence of sound: Research has shown that silent reading differs from oral reading (Rayner 2009) and that the process of subtitle reading is affected by both the presence or absence of sound (d'Ydewalle, Rensbergen, and Pollet 1987; Łuczak 2017).

3. *Problems with more common approaches to statistical analyses*

It is probably impossible to control for all the confounding variables that may affect the results of an eye-tracking experiment investigating the processing of subtitles[1]. Nevertheless, the higher the lack of control, the more unexplained variance there is in the statistical analysis (Balling 2008). Unfortunately, popular methods to analyze data such as *t*-

[1] Please note that we are not advocating for the careless inclusion of as many variables as possible in the model. Variable selection should be based on empirical findings, underlying theories, or at a minimum, on a plausible rationale (see Mellinger and Hanson 2017, Chapter 12, for a concise discussion on model selection; see also Field 2017). Also, controlling for too many variables may be problematic, not least because it may reduce statistical power (e.g., Field 2017).

tests, analyses of variance (ANOVAs), or linear regression models, all frequently adopted by AVT researchers, may be rather limiting on how they control for these confounding variables. There are at least two reasons for this.

One reason pertains to how the data are organized and compared. In eye-tracking studies on subtitling, researchers can obtain data for every subtitle or even every word separately. However, the data are then aggregated into averages so that each participant[2] ends up with one measurement for the dependent variable (e.g., one MFD value; see Figure 1). As such, there will be as many rows in the dataset as there are participants in the study. Thus, subtitle-related variables (e.g., number of words or lines per subtitle) and word-related variables (e.g., part of speech, frequency, or cognateness), which would necessitate one row per subtitle or per word (see Figure 2 in Section 4), cannot be included in the model to help control for unexplained variance. As a result, the variation in scores caused by the different characteristics of each individual subtitle or word is lost because of the aggregation.

Participants	Speed: 12 cps	Speed: 20 cps
P01	233	201
P02	322	253
P03	156	141
P04	281	285

Figure 1. Data aggregated per participant illustrating a within-subject design. The values represent hypothetical MFDs. CPS = characters per second.

There is another reason why these more popular statistical methods are less able than more advanced techniques to control for unexplained variance. As with any statistical test, there will always be unexplained variance, even after controlling for all relevant confounding variables. This might be because it was not possible to obtain certain measurements, or because certain sources of variance are simply unknown to researchers. With *t*-tests, ANOVAs, or regression analyses, among others, all this unexplained variance remains unexplained. With more advanced techniques that do not rely on data aggregation, it is possible to attribute much of this variance to differences between participants' performance, or to differences in scores (e.g., MFD) between different words or subtitles (see next section for a more in-depth discussion). This reduces unexplained variance, and hence increases

[2] This is the more common "by-participant" analysis. Data may also be averaged per subtitle (i.e., the average of all participants); still, such by-subtitle analyses also utilize aggregated data.

power (Meteyard and Davies 2020). One such advanced technique is linear mixed models (LMMs), to which we turn now.

4. *An introduction to linear mixed models*

Also known as multilevel models, random coefficient models, or hierarchical linear models (Field 2017), LMMs compute the estimates for each individual score for each item measured (e.g., the MFD for each subtitle or word). This way, each participant will have many rows of data, one per subtitle or per word. Therefore, in a study investigating 20 participants and 50 subtitles, there will be 1,000 data points (i.e., rows); if, for example, each subtitle consists of, on average, five words, and measurements are derived for each word separately, the final dataset might comprise around 5,000 data points. Participants, subtitles, and words may then be entered in the model as separate variables, therefore allowing the analysis to account for per-participant, per-subtitle, and per-word variation in scores (see Figure 2 below).

Participants	Speed	Clip	AOI	Line	FC	WC	MFD
P01	12 cps	GF	S1	1 line	3	5	233
P01	12 cps	GF	S2	1 line	2	4	193
P01	12 cps	GF	S3	2 lines	5	7	281
P01	12 cps	GF	S4	1 line	3	5	170
P01	20 cps	GG	S5	1 line	3	3	201
P01	20 cps	GG	S6	2 lines	7	10	198
P01	20 cps	GG	S7	2 lines	6	9	165
P01	20 cps	GG	S8	1 line	1	2	161
P02	12 cps	GG	S9	1 line	2	2	156
P02	12 cps	GG	S10	2 lines	9	14	320
P02	12 cps	GG	S5	1 line	2	4	184
P02	12 cps	GG	S7	1 line	3	5	192
P02	20 cps	GF	S2	1 line	1	1	172
P02	20 cps	GF	S1	2 lines	7	12	167
P02	20 cps	GF	S11	2 lines	5	10	140
P02	20 cps	GF	S3	2 lines	6	13	183

Figure 2. Per-subtitle data for LMM. Video clips: GF = Grace and Frankie; GG = Gilmore Girls; AOI = area of interest (i.e., the subtitles); FC = fixation count; WC = word count; MFD = mean fixation duration. The values in this figure are not based on real data. They are used for illustrative purposes only.

Also, because the dataset is organized in this manner, LMMs are able to analyze the data at multiple levels[3]. For example, participants with different L1s may be tested with different clips of different genres, at different laboratories and in different countries. These six

[3] In LMMs, data is said to be *nested* (different levels) or *crossed*. For more information see Cunnings and Finlayson 2015; Hajduk 2019).

different levels (i.e., participants, L1s, clips, genres, laboratories, and countries) may be also incorporated into the model to help explain further variance in the data. By contrast, in ANOVAs or regression analyses such variables would either be assumed comparable across conditions or be included as predictors (main effects, simple effects, and interactions), considerably increasing the use of degrees of freedom and greatly decreasing power (Field 2017)[4]. Importantly, a large amount of data is necessary for the analyst to be able to incorporate multilevel data in a model, which is one of the limitations of LMMs (see Section 8 for other limitations).

Still, incorporating the different levels of the data in the LMM is important because multilevel data are not truly independent, thus failing to meet the assumption of independence of observations (Field 2017; Hajduk 2019; Winter 2020), which is central to ANOVAs and regression analyses. It is expected, for instance, that MFD values will be more highly correlated across subtitles within the same participant than between participants, perhaps owing to participants' varying reading abilities, L1s, or education. Thus, by explaining variation in scores between participants, items, or even time of the day – i.e., by controlling for random effects (see below) – the problem with non-independence of observations is overcome (Field 2017; Hedeker 2003). Also, the LMM will account for more unexplained variance in the data, which generates more accurate measurements with enhanced power (Meteyard and Davies 2020). In fact, LMMs are much better able than more popular statistical approaches to find a difference when one exists in the population (i.e., they have more power) without an increased risk of erroneously rejecting the null hypothesis (i.e., a Type I error; Baayen 2008)[5].

4.1. Fixed and random effects

LMM is yet another technique that is part of general linear models, just like *t*-tests, ANOVAs, and regression analyses. The difference is that these more popular methods produce fixed-effect-only models (Balling and Hvelplund 2015; Mellinger and Hanson 2017; Winter 2020), whereas LMMs also include *random effects*, thus the term "mixed". *Fixed effects* are the variables of primary interest to the researcher. They may be entered in the model as factors or covariates. As factors, they are categorical variables and are equivalent to what is commonly called "independent", "predictor", or "explanatory" variables (Hajduk

[4] Essentially, the model would use one degree of freedom, e.g., for every participant, clip, or genre included as predictor.
[5] See Section 8 for some limitations of LMMs

2019; Heck, Thomas, and Tabata 2012; Winter 2020). In the example outlined in Section 2, the speed of subtitles (fast or slow) would be a factor. Fixed effects as covariates are continuous variables that may be included in the model to, for example, account for extra unexplained variance in the data (Hajduk 2019; Singmann and Kellen 2019), similarly to what is done in an analysis of covariance (ANCOVA), or in a regression model. In our example, it would be advisable to control for the number of words (or characters) per subtitle under the assumption that subtitle length may affect fixation durations. This would generate a continuous variable that may be used as a covariate in the model (see Figure 2 above). By contrast, random effects are those contextual (or grouping) factors that are not normally of primary interest to the researcher (Field 2017). Typical examples include participants and items (Balling 2008; Winter 2020), in our case the subtitles. The levels of the random effects (e.g., each participant or subtitle is one level) "are drawn randomly from a larger population of possibilities" (Mellinger and Hanson 2017, p. 234).

It is the random effects that grant LMMs an advantage over more widely used statistical techniques (Winter 2020). First, random effects help overcome the problem of non-independence of observations (see above); they make it possible for the dataset to be organized per subtitle (or word), thus obviating the need for data aggregation, which allows for the inclusion of subtitle-related (or word-related) covariates (see Section 2 and 3; Figure 2). Second, random effects help explain more variance in the data, which decreases overall error and increases power (Hajduk 2019; Meteyard and Davies 2020; see above). This is true in that even if several measurements are used as fixed effects (e.g., covariates) to control for variance in responses, there will always be some error variance related to participants and items that cannot be explained (Meteyard and Davies 2020). The random effects help to account for such unexplained variance in the model. Hence, the inclusion of random effects enables the researcher to determine whether there are differences in the fixed effects irrespective of differences between individual participants, items, or other contextual factors (Balling 2008). In our case, the inclusion of random effects help, for example, to control the idiosyncrasy typical of eye-tracking (Holmqvist et al. 2011): Some participants may have larger individual variations in fixation durations between conditions than other participants.

Apart from participants and subtitles, other (contextual) factors that may be included as random effects include the participants' country of provenance (e.g., different L1s and degree of exposure to subtitles may affect performance), location of the experiment, and time of the day at which the data were collected (see Section 4). Importantly, only categorical variables may be entered in the model as random effects (Singmann and Kellen 2019; Winter

2020); continuous variables must be entered as covariates. Also, random effects need to have at least five to six levels to generate reliable estimates (Bolker 2021; Hajduk 2019; Harrison et al. 2018). This may mean that, in some cases, only participants and subtitles may be included as random effects.

4.2. Random intercepts and slopes

LMMs may estimate *random intercepts*, *random slopes*, or both (Cunnings and Finlayson 2015; Harrison et al. 2018; Meteyard and Davies 2020). Random intercepts model the variance of the random effect in the average of the dependent variable (Bates et al. 2018; Cunnings and Finlayson 2015). Consider a model with two different levels of subtitling speed as the predictor (i.e., fixed effect) and MFD as the dependent variable. By incorporating random intercepts for participants in the analysis, for example, the researcher can model how the MFD values for each participant deviate from the MFD intercept in the population (predicted from the study sample; Winter, 2020). By contrast, fixed-effects-only statistical techniques model only the MFD intercept for the population. Another way to look at the random intercept is by considering the typical scatter plot of a simple linear regression analysis (Balling 2008; Field, 2017). In such a plot, the horizontal axis (X-axis) plots the predictor (e.g., subtitling speed) whereas the Y-axis plots the outcome (dependent) variable (e.g., MFD). The addition of random intercepts for participants to a linear mixed model enables it to plot a regression line for each participant (not only the line of best fit for all participants), and each regression line has a different (or random) intercept (i.e., the point where the regression line crosses the Y-axis; see Figure 3a). Adding random intercepts to LMMs makes sense since participants will almost always vary regarding their performance on the dependent variable. As a matter of fact, a substantial amount of extra variance is often explained this way (Balling 2008; Winter 2020). Still, random intercept models assume that the slope is constant, that is, that the variation for main effects and interactions (i.e., fixed effects) is the same for all participants (or other random effects). Considering our example, this is equivalent to assuming that MFD values will vary similarly for all participants across different subtitling speeds, either increasing or decreasing to similar extents. This is rather unrealistic given the various participant- and subtitle-related variables that may affect such variation.

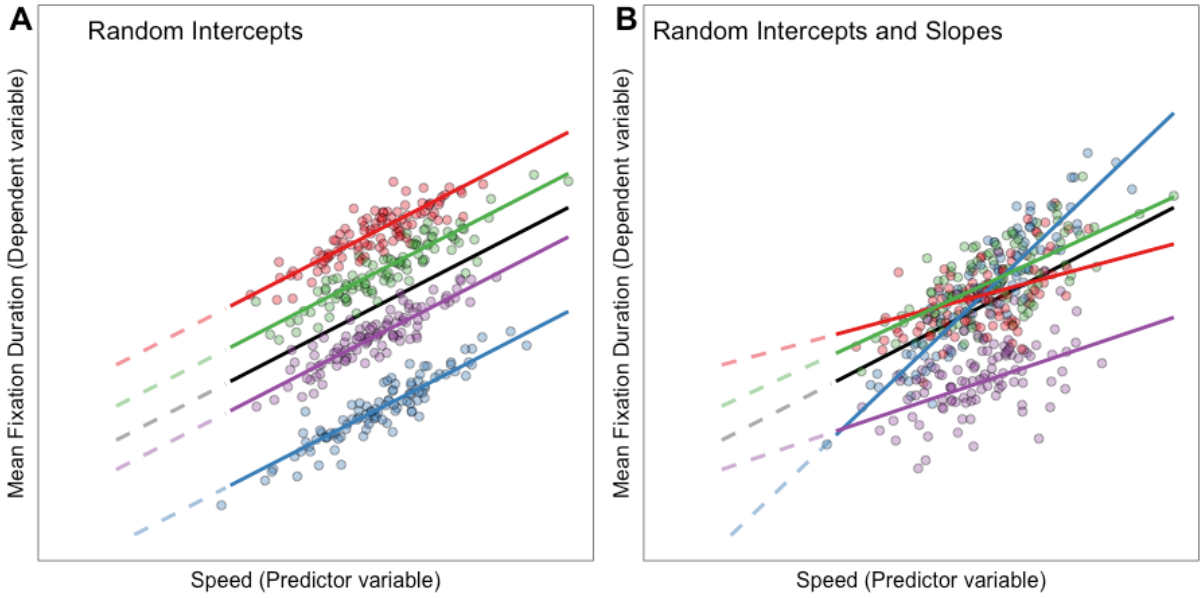

Figure 3. Random intercept (3a) and random intercept and slope (3b) models (adapted from Harrison et al. 2018). The colored lines represent the different participants. The black line represents the line of best fit.

To account for the variation between participants for the main effects and interactions, models need to include random slopes (Bates et al. 2018). In our example, a random slope model would mean that all participants are assumed to have the same starting point (i.e., the same intercept) but to differ in MFD values across speeds (see Balling 2008; Harrison et al. 2018)[6]. This is also unrealistic. A better model may include both random intercepts and slopes (Carson and Beeson 2013; Harrison et al. 2018), thus assuming a different starting point and random variation across fixed effects (e.g., subtitling speed) for each random effect (e.g., participants, subtitles). In this case, it is possible that the intercepts and slopes will correlate, and the model may also account for this. For instance, it is possible that participants with higher intercepts will have a steeper decline in MFD across subtitling speed than participants with lower intercepts. Such a model is illustrated in Figure 3b. A problem with fitting random slopes (and the intercept-slope correlations) is that although they increase the flexibility of the data analysis, they necessitate much more data to generate reliable estimates of separate slopes (or regression lines) for each level of a random effect (e.g., for each participant; Bates et al. 2018; Harrison et al. 2018). Without enough data, the effect may not be found and/or the model may simply not compute (i.e., fail to converge). Then, researchers may need to recruit more participants, or investigate more items (e.g., subtitles), which would require additional

[6] Note that to be possible to add a slope for Participants, the predictor (e.g., Speed) must be a within-participant variable. For example, all Participants watched subtitles at different speeds, so the slope for speed by participants (see Figure 3) is possible. However, in our dataset the subtitles were different at different speeds and clips (see Figure 2), and thus it is not possible to fit the slopes for speed and clip by subtitles (see Brown 2021).

resources. As a result, it may not always be possible to fit all intercepts and slopes that are, in theory, expected to exist. In this case, the model-building process may be driven by the data analysis itself (see below).

4.3. Fitting linear mixed models

Researchers do not agree on exactly how LMMs should be created, or *fit* (Field 2017; Hajduk 2019; Meteyard and Davies 2020). Overall, models may be fit in two ways: a "maximal-to-minimal that converges process" or a "minimal-to-maximal-that-improves fit process" (Meteyard and Davies, 2020, 17). Both ways involve several decisions from the analyst, which may increase the chance of statistical error (see Section 8 for other limitations of LMMs). To help minimize the possibility of error, we discuss below both methods in some detail.

In the maximal-to-minimal model, researchers fit all relevant fixed and random effects at once (Hajduk 2019). Then, one by one, variables that are deemed redundant or with very low estimates are eliminated until the model converges without any errors. This approach has been used in studies on subtitle reading (e.g., Liao et al. 2020). Other researchers who prefer this approach include Barr, Levy, Scheepers, and Tily (2013), Hajduk (2019), and Singmann and Kellen (2019). The maximal-to-minimal process is typically adopted in confirmatory analysis, that is, when there are clear hypotheses being tested (Bates et al 2018; Meteyard and Davies 2020). The maximal-to-minimal model is therefore theory-driven, whereas minimal-to-maximal-that-improves-fit models are more data-driven (see below).

This approach to fitting a model has benefits and drawbacks. One upside is that it includes all variables that may help answer the hypothesis set for the fixed effects (Bates et al. 2018). Even non-significant variables are kept, especially those that are considered essential to investigate the hypothesis (Hajduk 2019). For example, in an eye-tracking study measuring MFD on subtitles, if viewers watch different videos (i.e., clips) within or between conditions, it is expected that performance among clips will vary; in this case, it makes sense to include the effect of different clips in the final model, even if the test was statistically non-significant. A problem with this approach is that it may lead to overparameterized models with less power (Bates et al. 2018). To avoid overparameterization, Bates et al. (2018, 4-5) have recommended a series of steps that may be used to reach a *parsimonious model*, an approach that has become popular among researchers. Essentially, after fitting a maximal model, random effects that explain little to no variance and do not improve fit are removed. The quality of fit is usually assessed by comparing *Akaike's information criterion* (AIC) or

Schwarz Bayesian criterion (BIC) across models, with lower scores meaning better fit models (Meteyard and Davies 2020).

The minimal to maximal-that-improves fit approach to model building is usually recommended for exploratory analysis, that is, when there are no clear-cut hypotheses (Baayen 2008b; Meteyard and Davies 2020). Here, researchers start with a simpler model, either from the fixed or the random effects, until obtaining the most complex model that is theoretically defensible and that can be explained by an improvement in model fit (again, by comparing AIC or BIC across models). Some researchers who follow this approach include Carson and Beeson (2013), Cunnings and Finlayson (2015), Field (2017), Heck et al. (2012), and Szarkowska et al. (2021) This is mostly the approach that will be followed in the example quasi-experiment described below (which is exploratory in nature). Please refer to Section 4.2.1 for more details on the model building process.

5. ***An example quasi-experiment: Comparing t-tests to LMMs***

In this section, we re-analyze a sample from an eye-tracking study on reading subtitles by Szarkowska and Gerben-Morón (2018). The original dataset is available online ([dataset] Silva, Orrego-Carmona, and Szarkowska, 2021a). The purpose of this re-analysis is to illustrate some of the advantages of LMMs over *t*-tests (and other popular statistical techniques such as ANOVAs) when analyzing eye-tracking data on reading subtitles. We provide such illustration by answering the following research question:

RQ. Is there a difference in mean fixation duration between slow (12 cps) and fast (20 cps) subtitle speeds?

5.1. Research Design

The original dataset consisted of Polish, English, and Spanish participants. Here, we re-analyze data from Polish participants only in a within-subject design. Therefore, although the original study used ANOVAs, here we employ *t*-tests (as there is no between-subject variable). The dataset used for *t*-tests is organized in a typical one-row-per-participant wide format (see Figure 1); the dataset used for the LMM analysis is organized in a one-row-per-subtitle long format (see Figure 2). Therefore, the difference between the datasets (apart from the wide and long formats) is that for *t*-tests all measurements for all subtitles were necessarily averaged per participant. Both sets used in our analyses are available online ([dataset] Silva, Orrego-Carmona, and Szarkowska, 2021a). Below, we provide only the

details on research design that are necessary to understand the analyses and results. The original study (Szarkowska and Gerber-Morón 2018) includes more information on research design and data processing.

18 Polish hearing young adults (15 females) with advanced proficiency in English watched English-language videos with Polish interlingual subtitles. The materials analyzed were two self-contained scenes of similar length from *Grace and Frankie* (henceforth GF; 2015, created by Marta Kauffman and Howard J. Morris) and *Gilmore Girls* (GG; 2000, created by Amy Sherman-Palladino), and both were dialogue-heavy and featured two to four people engaged in conversation. The subtitles in both clips were deemed comparable in overall difficulty (i.e., easy to read; Jasnopis readability index: 1 or 2 out of 7 for all subtitles).

In a within-subject design, each participant watched both clips at two different speeds: the 12-character-per-second (cps) condition (slow speed) and the 20-cps condition (fast speed). The dependent variable analyzed was mean fixation duration (i.e., the average, in milliseconds, of all fixations in a subtitle [for LMMs] or in all subtitles per participant per condition [for *t*-tests]). To ensure that the clips did not influence the results, they were counterbalanced across speeds and the order of presentation of the stimuli was randomized using the SMI Experiment Centre. Below, we first analyze the data using a paired-samples *t*-test, a more widely employed technique for data analysis; then, we analyze the data using LMMs. All data were analyzed using IBM SPSS version 27, and the alpha value was set at 0.05.

6. ***Results with a more traditional approach: paired-samples t-tests.***

The descriptive statistics show that the MFD for 12 cps (M = 185.33; SD = 26.35) was higher than the MFD for 20 cps (M = 175.69; SD = 26.21). As typically happens with eye-tracking data (Holmqvist et al. 2011), the dependent variable MFD for 12 cps failed the assumption of normality (z score > 1.96); hence, both dependent variables (MFD for 12 cps and for 20 cps) were normalized with a log transformation. In answering the research question, the *t*-test found a significant difference in MFD between 12 cps and 20 cps, $t(17)$ = 2.451, p = 0.025, 95% CI [.008, .102], Cohen's d = .58, representing a small effect size (see Mellinger and Hanson 2017, Chapter 7 for a discussion on effect sizes). Note also that the 95% CI approximates zero (i.e., it suggests no difference in speed), which reinforces the weak effect size found. Put differently, while there was a statistically significant difference between

12 cps and 20 cps, this difference was minor (i.e., the time viewers spent fixating on subtitles differed only slightly between speeds and might be of little relevance in practice)[7].

7. *Analyzing the data with linear mixed models*

7.1. Fitting the model

As recommended for exploratory data analysis, the LMM reported here followed a minimal-to-maximal-that-improves fit process and used AIC to assess model fit (see Appendix S1 at Silva, Orrego-Carmona, and Szarkowska, 2021b), in accordance with guidelines suggested by Meteyard and Davies (2020). We started by fitting the random effects in a stepwise fashion. First, we entered the intercepts, then intercepts and slopes. Nevertheless, to keep this introduction to LMMs simpler, we did not include the next recommended step, i.e., the intercept-slope correlations[8](see Section 4.2. and 4.3. for the importance of and limitations when including such correlations). If certain effects did not improve fit (i.e., a considerable decrease in AIC), were deemed redundant, had estimates that approximated zero, or caused errors in computation (e.g., failure to converge), they were eliminated from the model. After having established the random effects model, the fixed effects were added. The final model, reported here, is the model with the lowest AIC that did not return any error. Of note, fixed effects that were deemed theoretically essential (see below) to the model were not removed, even if non-significant (Baayen, 2008b).

As mentioned above, the main purpose of this paper is to highlight the advantages of LMMs over more traditional statistical techniques. Due to this narrow scope, we do not include here all stages of the model-building process, only the final model. The interested reader is therefore advised to refer to Appendix S1 (Silva, Orrego-Carmona, and Szarkowska, 2021b) for the step-by-step model-building process, and to Carson and Beeson (2013), using IBM SPSS, and Cunnings and Finlayson (2015), using R (RStudio Team 2021), for introductory papers detailing this process.

7.2. Variables included in the model and data analysis

Regarding fixed effects, both factors and covariates were included in the model, and many decisions needed to be made in the process (see Section 8 for some limitations of

[7] Effect sizes should always be considered when interpreting results. *p* values do not indicate the strength of the effect and may be affected by factors such as variance and sample size.
[8] See Garson (2020) for more details on intercept-slope correlations.

LMMs). The main factor needed to answer the research question was Speed (12 cps and 20 cps). The factor Clip was also included in the model to account for the possible influence of different clips (GF and GG) on MFD. Clip, although suitable as a random effect, was entered as a fixed effect because it had fewer than five levels (Bolker 2021; Hajduk 2019; Harrison et al. 2018). The factor Line (i.e., one or two lines in the subtitles) was added in the model because MFD was expected to be influenced by the number of lines (d'Ydewalle and De Bruycker 2007). The interactions Speed*Clip and Speed*Line were also included as fixed effects. Speed*Clip allows the model to consider the possibility that MFD varies to different extents between clips across speeds; Speed*Line allows the model to consider the fact that the number of lines may influence MFD differently across speeds. The subtitle-related covariates included in the model were Fixation Count and Number of Words because they were expected to explain extra variance in MFD. Of all the fixed effects, only Speed, Clip, and the Speed*Clip interaction were considered essential as they were part of the original design and research question; hence, they were not removed from the final model irrespective of statistical significance and model fit.

The random effects entered in the model were the intercepts for Participants and Subtitles, and the slopes for Speed, Clip, and Line by Participants. Figure 2 above shows the structure of the dataset used for the LMM (for the actual data used see [dataset] Silva, Orrego-Carmona, and Szarkowska, 2021a).

The initial dataset consisted of 2409 cases (rows). The dependent variable MFD met the assumption of homogeneity of variance (after a visual inspection of histograms and the Levene test [$p = .157$]) but failed the assumption of normality ($z > 1.96$). A log transformation improved the distribution, though insufficiently. Hence, to reduce the skewness, we removed all cases below or above 3 *SD*. This removal resulted in the elimination of 23 cases and a final dataset with 2386 cases. The resulting log-transformed MFD dependent variable approximated normality with no apparent influential outlier (see Appendix S2 for data diagnostics from the final model; Silva, Orrego-Carmona, and Szarkowska, 2021b). Finally, all covariates were centered, and hence the intercept corresponds to the grand mean.

7.3. Results

Table 1 shows the descriptive statistics and number of subtitles (i.e., rows in the dataset) by Clip and Speed. It appears that MFD for GG is higher than for GF irrespective of the number of lines and speed; also, 2-line and 12-cps subtitles seems to generate higher MFD than 1-line and 20-cps subtitles across the table.

Table 1
Mean fixation duration by Speed, Clip, and Line.

Clip	Line type (12 cps/20 cps)	12 cps		20 cps	
		M	*SD*	*M*	*SD*
Grace and Frankie (1,006)	1-line subtitles (252/222)	170.99	47.74	160.54	50.44
	2-line subtitles (295/237)	186.16	51.23	171.35	54.09
Gilmore Girls (1,380)	1-line subtitles (203/137)	176.75	52.70	166.28	47.16
	2-line subtitles (439/601)	190.51	56.83	184.24	49.08
Total (2,386)		182.95	53.43	175.24	51.01

Note. Number of subtitles indicated in brackets. Total averages may differ slightly due to rounding.

However, the LMM shown in Table 2 does not corroborate the trend illustrated by the descriptive statistics. Let us focus on the fixed effects. First, to answer the research question, the LMM found no statistically significant difference in Speed ($p = 0.125$). This lack of difference runs counter to the results of the *t*-test (see Section 6), which found a difference ($p = 0.025$), with the MFD on 20 cps subtitles being lower than on 12 cps subtitles. Put differently, when using LMMs, and thus accounting for extra unexplained variance in the data (see below), the difference for Speed disappeared. For example, Line and Fixation Count, subtitle-related variables that significantly contributed to the model ($p < .001$), and explained additional variance, could only be included in the analysis because of the one-row-per-subtitle long format needed in LMMs.

Table 2
Fixed and random effect estimates for MFD comparing 12 cps and 20 cps.

Fixed effects				
	Estimate (Std. error)	*95 % CI*	*p*	Cohen's *d*
Intercept[a]	5.085 (.039)	[5.002, 5.168]	< .001	
Speed	-.093 (.059)	[-.215, .028]	.125	0.33
Clip	.040 (.055)	[-.075, .155]	.477	0.14
Line	.117 (.021)	[.074, .160]	< .001	0.42
Speed*Clip	.028 (.111)	[-.204, .261]	.800	0.10
Fixation Count	-.011 (.002)	[-.015, -.007]	< .001	0.04
Random effects				
	Variance (Std. error)	*p*		*ICC*[b]
Residual	.054 (.002)	<.001		
Participants (Intercept)	.011 (.004)	.011		0.1467
Speed | Participants slope	.004 (.002)	.046		0.0542
Line | Participants slope	.004 (.002)	.055		0.0458
Subtitles intercept	.005 (.001)	<.001		0.0701

Note. Number of data points = 2386; participants = 18. Degrees of freedom estimation: Satterthwaite. Method of estimation: REML (see Hajduk 2019 for difference between ML and REML).
[a]The intercept corresponds to the value of log-transformed MFD when all factors are 0 (Speed = 12 cps; Clip = GF; Line = 1 line) and Fixation Count is at its grand mean (i.e., 5.45). [b]ICC = intraclass correlation coefficients.

The estimates of the fixed effects offer a more fine-grained interpretation of the results. These may be interpreted similarly to the estimates of regression analyses; however, as we log-transformed MFD, the interpretation must account for such transformation[9]. For instance, according to the estimate for Speed, the model predicts that MFD decreases (a negative estimate) by 8.9% from 12 cps (the reference category) to 20 cps when all other predictors are held constant. This decrease is illustrated in Table 1, but according to the model, it does not reach statistical significance and has a weak effect size (Cohen's d = .33).

Turning to the second part of Table 2, the tests for all random effects were significant or approached significance[10]. Regarding the intercepts, this means that the difference in average MFDs for different participants and subtitles was statistically significant. As for the slopes (Speed and Line by Participants), the model found that MFD varied differently for different participants between the two speeds (12 cps and 20 cps) and number of lines (one or two lines). That is, for instance, the decrease (or increase) in MFD between speeds may have been greater for some participants than for others. This considerable variation between participants and subtitles is highly expected given all the participant- and subtitle-related confounding variables – known (see Section 2) and unknown – that may influence the results. Still, this variation in the results was at least partly accounted for by the random effects in the LMM, thus highlighting the advantages of linear mixed models over fixed-effects-only statistical methods.

The intraclass correlation (ICC) provides a better idea of the extra variation explained by the random effects. It describes the proportion of variance explained by the random effects after accounting for the fixed effects (see Carson and Beeson 2013; Heck et al. 2012 for how to calculate the ICC). As an example, the ICC for the Participants intercept (0.1467) shows that this random effect explained an extra 14.67% of variation in the data. Taken together, all random effects explain 31.68% of the variation in MFD values that were left unexplained by

[9] Because MFD was log-transformed, the estimates must be exponentiated ("UCLA Institute of Digital Research" 2021). For instance, the exponentiation of the positive estimate for Clip (.040) is 1.04. This means that the model predicts that MFD increases by 4% from GF (the reference category) to GG when all the other variables are held constant. For negative estimates, e.g. -.093 for Speed, it is necessary to exponentiate and then deduct the result from one. For Speed, this means 1 - 0.911 = 0.089 (8.9%). Except for the intercept, the estimates reflect proportions.
[10] In LMMs, p values are usually not very reliable due to the difficulty in estimating degrees of freedom (Baayen 2008b). Effects whose tests approach significance often contribute to the model and improve fit.

the fixed effects. In other words, the random effects in the model reduced statistical error and increased the power of the analysis (Meteyard and Davies 2020), leading to more reliable results.

8. Conclusions, limitations, and research recommendations

Researchers conducting eye-tracking studies on subtitling need to bear in mind many elements that, if not controlled for, may distort the results. Nevertheless, it may be impossible to control for so many variables, not least because many confounding factors remain unknown. To address this issue, this paper has made the case for using LMMs in statistical analyses in eye-tracking studies on subtitling.

Overall, the advantage of LMMs over more traditional statistical methods is two-fold. First, as it does not rely on aggregated values for participants (or items), the model enables the researcher to include participant-related predictors (e.g., L2 proficiency) together with item-related predictors for every subtitle (e.g., fixation counts, word counts, character counts) or every word (e.g., length, frequency) within the subtitles (Baayen 2008a). Second, the random effects help account for further unexplained variance in the results.

Importantly, linear mixed models are not without limitations (Mellinger and Hanson 2017). First, LMMs necessitate a large amount of data to attain sufficient power. Meteyard and Davies (2020) recommend at least 30-50 participants per condition, and 30-50 items (i.e., subtitles) per participant, totalling 900-2500 data points per condition. Second, LMMs must meet most of the assumptions of linear regression models: i.e., linearity and no multicollinearity; normality and homoscedasticity of residuals (see Mellinger and Hanson 2017; Meteyard and Davies 2020). Note that the assumption of independence of residuals need only be met by the fixed effects, but the inclusion of random effects solves the problem of non-independence. Also, random effects can only be included in the model with a minimum of five to six levels. These constraints in the number of participants, items, and levels of random effects all affect research design. Finally, conducting LMMs analyses involves making considerably more choices than with fixed-effects-only statistical techniques. This higher number of choices warrants significantly more knowledge and judgement from the analyst, requires more (and more informed) decisions from the researcher, and increases the chances of statistical error.

Given the limitations above, the following recommendations regarding the use of LMMs are in order. Studies that produce large sets of data may take advantage of linear mixed models. As we have demonstrated, eye-tracking studies on subtitle reading are a clear

example of when LMMs might be used, and corpus-based studies could follow a similar approach. Any study where data is obtained from multiple items for each participant (e.g., questionnaires or other tests) may make use of LMMs (instead of one score per test, the researcher would consider one score per individual test item). Furthermore, the data must be nested in multiple levels (or grouping factors) or crossed (e.g., when the same items are tested across subjects, as in the study reported here; see Section 4). Yet, if random effects are entered in the model (e.g., intercepts for participants and subtitles), and they are not significant and do not improve model fit, LMMs may not be necessary. In this case, t-tests, ANOVAs, and regression analyses may suffice.

Address for correspondence

Breno B. Silva
Institute of English Studies
University of Warsaw
ul. Hoża 69
00-681
Warsaw, Poland
b.barreto-sil2@uw.edu.pl

Biographical notes

Breno Barreto Silva is Assistant Professor in the Department of Applied Linguistics and Translation Studies at the University of Warsaw. His research interests and projects are: second language acquisition; academic vocabulary; incidental lexical learning (reading and writing); psycholinguistics; bilingualism and multilingualism; cognate vocabulary; eye-tracking research; statistics; vocabulary learning and teaching; vocabulary research methodology; and teaching English to adults.
https://orcid.org/0000-0001-6574-5896

David Orrego Carmona is a lecturer in translation studies at Aston University (UK) and a research associate at the University of the Free State (South Africa). My research deals with translation, technologies and users. It analyses how translation technologies empower professional and non-professional translators and how the democratisation of technology allows translation users to become non-professional translators. I am ultimately interested in how nonprofessionalism impacts social translation, professional translation and translator training.
https://orcid.org/0000-0001-6459-1813

Agnieszka Szarkowska is Head of AVT Lab, one of the first research groups on audiovisual translation. Agnieszka is a researcher, academic teacher, ex-translator, and translator trainer. Her research projects include eye tracking studies on subtitling, audio description,

multilingualism in subtitling for the deaf and the hard of hearing, and respeaking. https://orcid.org/0000-0002-0048-993X

References

Baayen, Harald. 2008a. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

———. 2008b. *Exploratory Data Analysis: An Introduction to R for the Language Sciences*. Cambridge: Cambridge University Press.

Balling, Laura Winther. 2008. "A Brief Introduction to Regression Designs and Mixed-Effects Modelling by a Recent Convert." *Copenhagen Studies in Language*, 175–92.

Balling, Laura Winther, and Kristian Tangsgaard Hvelplund. 2015. "Design and Statistics in Quantitative Translation (Process) Research". *Translation Spaces* 4 (1): 170-187. https://doi.org/10.1075/ts.4.1.08bal

Barr, Dale. J., Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal." *J Mem Lang* 68 (3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth, and R. Harald Baayen. 2018. "Parsimonius Mixed Models." arXiv:1506.04967v2 [stat.ME].

Bisson, Marie-Josée, Walter J. B. Heuven, Kathy Conklin, and Richard J. Tunney. 2014. "The Role of Repeated Exposure to Multimodal Input in Incidental Acquisition of Foreign Language Vocabulary." *Language Learning* 64 (4): 855–77. https://doi.org/10.1111/lang.12085.

Bolker, Ben. 2021. "GLMM FAQ." 2021. http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#reml-for-glmms.

Braverman, Barbara, and Melody Hertzog. 1980. "The Effects of Caption Rate and Language Level on Comprehension of a Captioned Cideo Presentation." *American Annals of the Deaf* 125 (7): 943–48. https://doi.org/10.1353/aad.2012.1118.

Brown, Violet A. 2021. An Introduction to Linear Mixed-Effects Modelling in R. *Advances in Methods and Practices in Psychological Science* 4 (1): 1-19

Brysbaert, Marc, and Boris New. 2009. "Moving beyond Kučera 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): 977–90. https://doi.org/10.3758/BRM.41.4.977.

Burnham, Denis, Kaoru Sekiyama, Gerard Bailly, Pascal Perrier, and Eric Vatikiotis-Bateson. 2012. "Investigating Auditory-Visual Speech Perception Development." In *Audiovisual Speech Processing*, edited by Gérard Bailly, Pascal Perrier and Eric Vatikiotis-Bateson, 62–75. Cambridge: Cambridge University Press.

Cambra, Cristina, Olivier Penacchio, Núria Silvestre, and Aurora Leal. 2014. "Visual Attention to Subtitles When Viewing a Cartoon by Deaf and Hearing Children: An Eye-Tracking Pilot Study." *Perspectives* 22 (4): 607–17. https://doi.org/10.1080/0907676x.2014.923477.

Carson, Robyn J., and Christina M. L. Beeson. 2013. “Crossing Language Barriers: Using Crossed Random Effects Modelling in Psycholinguistics Research.” *Tutorials in Quantitative Methods for Psychology* 9 (1): 25–41. https://doi.org/10.20982/tqmp.09.1.p025.
Crossley, Scott A., Tom Cobb, and Danielle S. McNamara. 2013. “Comparing Count-Based and Band-Based Indices of Word Frequency: Implications for Active Vocabulary Research and Pedagogical Applications.” *System* 41 (4): 965–81. https://doi.org/10.1016/j.system.2013.08.002.
Cunnings, Ian, and Ian Finlayson. 2015. “Mixed Effects Modeling and Longitudinal Data Analysis.” In *Advancing Quantitative Methods in Second Language Research*, edited by Luke Plonsky, 159–81. New York: Routledge.
De Bruycker, Wim, and Gery d’Ydewalle. 2003. “Reading Native and Foreign Language Television Subtitles in Children and Adults.” In *The Mind’s Eye: Cognitive and Applied Aspects of Eye Movement Research*, edited by Jukka Hyönä, Ralph Radach, and Heiner Deubel, 671–84. Amsterdam: North-Holland. https://doi.org/10.1016/B978-044451020-4/50036-0.
Dijkstra, Ton, Koji Miwa, Bianca Brummelhuis, Maya Sappelli, and Harald Baayen. 2010. “How Cross-Language Similarity and Task Demands Affect Cognate Recognition.” *Journal of Memory and Language* 62 (3): 284-301. https://doi.org/10.1016/j.jml.2009.12.003.
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 Imaging*, edited by Jodi Sita, Tessa Dwyer, Sean Redmond, and Claire Perkins, 46–64. London: Bloomsbury.
Duyck, Wouter, Eva Van Assche, Denis Drieghe, and Robert J Hartsuiker. 2007. “Visual Word Recognition by Bilinguals in a Sentence Context: Evidence for Nonselective Lexical Access." *J Exp Psychol Learn Mem Cogn* 33 (4): 663–79. doi: 10.1037/0278-7393.33.4.663.
Ellis, Nick C., and Alan Beaton. 1993. “Psycholinguistic Determinants of Foreign Language Vocabulary Learning.” *Language Learning* 43 (4): 559–617. https://doi.org/10.1111/j.1467-1770.1993.tb00627.x.
Field, Andy. 2017. *Discovering Statistics Using IBM SPSS Statistics*. 5th ed. London: SAGE Publications, Inc.
Garson, G. David. 2020. *Multilevel modelling: Applications in Stata®, IBM® SPSS®, SAS®, R, & HLM™* . New York: Sage.
Gerber-Morón, Olivia, Olga Soler-Vilageliu, and Judith Castella. 2019. “The Effects of Screen Size on Subtitle Layout Preferences and Comprehension across Devices.” *Hermeneus* 22: 157-182. https://doi.org/10.24197/her.22.2020.157-182
Gerber-Morón, Olivia, and Agnieszka Szarkowska. 2018. “Line Breaks in Subtitling: An Eye Tracking Study on Viewer Preferences.” *Journal of Eye Movement Research* 11 (3): 1–22. https://doi.org/10.16910/jemr.11.3.2.
Gerber-Morón, Olivia, Agnieszka Szarkowska, and Bencie Woll. 2018. “The Impact of Text Segmentation on Subtitle Reading.” *Journal of Eye Movement Research* 11 (4): 1–18. https://doi.org/10.16910/jemr.11.4.2.
Ghia, Elisa. 2012. “The Impact of Translation Strategies on Subtitle Reading.” In *Eye Tracking in Audiovisual Translation*, edited by Elisa Perego, 155–82. Roma: Aracne Editrice.

Gile, Daniel. 2016. “Experimental research.” In *Researching Translation and Interpreting*, edited by Claudia V. Angelelli and Brian James Baer, 220–228. New York: Routledge.
Hajduk, Gabriela K. 2019. “*Introduction to Linear Mixed Models.*” https://ourcodingclub.github.io/tutorials/mixed-models/#crossed.
Harrison, Xavier. A., Lynda Donaldson, Maria E. Correa-Cano, Julian Evans, David N. Fisher, Cecily E. D. Goodwin, Beth. S. Robinson, David J. Hodgson, and Richard Inger. 2018. “A Brief Introduction to Mixed Effects Modelling and Multi-Model Inference in Ecology.” *PeerJ* 6: e4794. https://doi.org/10.7717/peerj.4794.
Heck, Ronald H., Scott L. Thomas, and Lynn N. Tabata. 2012. *Multilevel Modelling of Categorical Outcomes Using IBM SPSS*. Quantitative Methodology Series. New York / London: Routledge.
Hedeker, Donald. 2003. “A Mixed-Effects Multinomial Logistic Regression Model.” *Statistics in Medicine* 22 (9): 1433–46. https://doi.org/10.1002/sim.1522.
Holmqvist, Kenneth, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. 2011. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford: Oxford University Press.
Inhoff, Albrecht W., and Keith Rayner. 1986. “Parafoveal Word Processing during Eye Fixations in Reading: Effects of Word Frequency.” *Perception & Psychophysics* 40 (6): 431–39. https://doi.org/10.3758/BF03208203.
Jensema, Carl. 1998. “Viewer Reaction to Different Television Captioning Speeds.” *American Annals of the Deaf* 143 (4): 318–24.
Koolstra, Cees M., Tom H. A. Van Der Voort, and Gery d’Ydewalle. 1999. “Lengthening the Presentation Time of Subtitles on Television: Effects on Children’s Reading Time and Recognition.” *Communications* 24 (4): 407–22. https://doi.org/10.1515/comm.1999.24.4.407.
Krejtz, Izabela, Agnieszka Szarkowska, and Maria Łogińska. 2016. “Reading Function and Content Words in Subtitled Videos.” *Journal Of Deaf Studies And Deaf Education* 21 (2): 222–32. https://doi.org/10.1093/deafed/env061.
Kruger, Jan-Louis. 2018. “Eye tracking in audiovisual translation research”. In *The Routledge Handbook of Audiovisual Translation,* edited by L. Pérez-González*,* 350-366. Routledge. https://doi.org/10.4324/9781315717166-22.
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, Stephen Doherty, and María T. Soto-Sanfiel. 2017. “Original Language Subtitles: Their Effects on the Native and Foreign Viewer.” *Comunicar*, 50 (January): 23–32. https://doi.org/10.3916/c50-2017-02.
Kruger, Jan-Louis, and Stephen Doherty. 2018. “Triangulation of Online and Offline Measures of Processing and Reception in AVT.” In *Reception Studies and Audiovisual Translation*, edited by Elena Di Giovanni and Yves Gambier, 91–109. Amsterdam/New York: John Benjamins. https://doi.org/10.1075/btl.141.06kru.
Lemhöfer, Kristin, and Ton Dijkstra. 2004. “Recognizing Cognates and Interlingual Homographs: Effects of Code Similarity in Language-Specific and Generalized Lexical Decision.” *Memory & Cognition* 32 (4): 533–50. https://doi.org/10.3758/BF03195845.

Liao, Sixin, Jan-Louis Kruger, and Stephen Doherty. 2020. “The Impact of Monolingual and Bilingual Subtitles on Visual Attention, Cognitive Load, and Comprehension.” *Journal of Specialised Translation* 33: 70–98.
Liao, Sixin, Lili Yu, Erik D. Reichle, and Jan-Louis Kruger. 2020. “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.
Łuczak, Krzysztof. 2017. “The Effects of the Language of the Soundtrack on Film Comprehension, Cognitive Load and Subtitle Reading Patterns. An Eye-Tracking Study.” *Institite of Applied Linguistics*. Warsaw: University of Warsaw.
Mellinger, Christopher D., and Thomas A. Hanson (2017). *Quantitative Research Methods in Translation and Interpreting Studies*. New York: Routledge.
Meteyard, Lotte, and Robert A. I. Davies. 2020. “Best Practice Guidance for Linear Mixed-Effects Models in Psychological Science.” *Journal of Memory and Language* 112. https://doi.org/10.1016/j.jml.2020.104092.
Moran, Siobhan. 2009. “The Effect of Linguistic Variation on Subtitle Reception.” Toronto: York University.
Mulder, Kimberley, Ton Dijkstra, and R. Harald Baayen. 2015. “Cross-Language Activation of Morphological Relatives in Cognates: The Role of Orthographic Overlap and Task-Related Processing.” *Frontiers in Human Neuroscience* 9. https://doi.org/10.3389/fnhum.2015.00016.
Muñoz, Carmen. 2017. “The Role of Age and Proficiency in Subtitle Reading. An Eye-Tracking Study.” *System* 67: 77–86. https://doi.org/10.1016/j.system.2017.04.015.
Orrego-Carmona, David. 2015. “The Reception of (Non)Professional Subtitling.” *Department of English and German Studies*. Tarragona: Universitat Rovira i Virgili.
Peters, Elke. 2019. “Factors Affecting the Learning of Single-Word Items.” In *The Routledge Handbook of Vocabulary Studies*, edited by Stuart Webb, 125–42. New York: Routledge.
Ragni, Valentina. 2020. “More than Meets the Eye: An Eye-Tracking Study of the Effects of Translation on the Processing and Memorisation of Reversed Subtitles.” *Journal of Specialised Translation* 33: 99–128.
Rajendran, Dhevi J., Andrew T. Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. 2013. “Effects of Text Chunking on Subtitling: A Quantitative and Qualitative Examination.” *Perspectives* 21 (1): 5–21. https://doi.org/10.1080/0907676X.2012.722651.
Rayner, Keith. 2009. "Eye movements and attention in reading, scene perception, and visual search". *The Quarterly Journal of Experimental Psychology* 62(8); 1457-1506. https://doi.org/10.1080/17470210902816461
Rayner, Keith, and Susan A. Duffy. 1986. “Lexical Complexity and Fixation Times in Reading: Effects of Word Frequency, Verb Complexity, and Lexical Ambiguity.” *Memory & Cognition* 14 (3): 191–201. https://doi.org/10.3758/BF03197692.
Rayner, Keith, Tessa Warren, Barbara J. Juhasz, and Simon P. Liversedge. 2004. “The Effect of Plausibility on Eye Movements in Reading.” *J Exp Psychol Learn Mem Cogn* 30 (6): 1290–1301. https://doi.org/10.1037/0278-7393.30.6.1290.
Rayner, K., Timothy J. Slattery, Denis Drieghe, and Simon P. Liversedge. 2011. “Eye Movements and Word Skipping during Reading: Effects of Word Length and Predictability.” *J Exp Psychol Hum Percept Perform* 37 (2): 514–28. https://doi.org/10.1037/a0020990.

Romero-Fresco, Pablo. 2015. “Final Thoughts: Viewing Speed in Subtitling.” In *The Reception of Subtitles for the Deaf and Hard of Hearing in Europe*, edited by Pablo Romero-Fresco, 335–41. Bern: Peter Lang.
RStudio Team. 2021. *Studio: Integrated Development for R. RStudio*. R. Boston. http://www.rstudio.com/.
Saldanha, Gabriela, and Sharon O’Brien. 2014. *Research Methodologies in Translation Studies*. New York: Routledge.
Schwartz, Ana I., and Judith F. Kroll. 2006. “Bilingual Lexical Activation in Sentence Context.” *Journal of Memory and Language* 55 (2): 197–212. https://doi.org/10.1016/j.jml.2006.03.004.
Singmann, Henrik, and David Kellen. 2019. “An Introduction to Mixed Models for Experimental Psychology.” In *New Methods in Cognitive Psychology*, edited by Daniel Spieler and Eric Schumacher, 1st ed., 4–31. Routledge. https://doi.org/10.4324/9780429318405-2.
Silva, Breno B., David Orrego-Carmona, and Agnieszka Szarkowska. 2021a. Data for paper entitled ‘Using linear mixed models to analyze data from eye-tracking research on subtitling’ *Figshare*. Available at: https://doi.org/10.6084/m9.figshare.14403389.v1 (accessed March 2022).
Silva, Breno B., David Orrego-Carmona, and Agnieszka Szarkowska. 2021b. Appendices for paper entitled ‘Using linear mixed models to analyze data from eye-tracking research on subtitling’ *Figshare*. Available at: https://doi.org/10.6084/m9.figshare.14403428.v2 (accessed March 2022).
Szarkowska, Agnieszka, Izabela Krejtz, Zuzanna Kłyszejko, 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): 363–78. https://doi.org/10.1353/aad.2011.0039.
Szarkowska, Agnieszka, and Olivia Gerber-Morón. 2018. “Viewers Can Keep up with Fast Subtitles: Evidence from Eye Movements.” *Plos One* 13 (6). https://doi.org/10.1371/journal.pone.0199331.
———. 2019. “Two or Three Lines: A Mixed-Methods Study on Subtitle Processing and Preferences.” *Perspectives* 27 (1): 144–64. https://doi.org/10.1080/0907676x.2018.1520267.
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): 101–24. https://doi.org/10.1075/tcb.00022.sza.
Szarkowska, Agnieszka, Breno B. Silva, and David Orrego-Carmona. 2021. Effects of subtitle speed on proportional reading time: Re-analysing subtitle reading data with mixed effects models. *Translation, Cognition & Behavior*, 4(2): 305-330. https://doi.org/10.1075/tcb.00057.sza
“UCLA Institute of Digital Research.” 2021. FAQ How Do I Interpret a Regression Model When Some Variables Are Log Transformed? 2021. https://docs.google.com/document/d/1zHvX_0tOsbfAgz1VgizBh4WWJjQQxIvObYf0AP7Pd1U/edit.

Winter, Bodo. 2020. *Statistics for Linguists: An Introduction Using R*. New York / London: Routledge.
Ydewalle, Géry d', Johan van Rensbergen, 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. K. O'Regan and A. Levy-Schoen, 313–21. Amsterdam/New York: Elsevier.
Ydewalle, Géry d', Caroline Praet, Karl Verfaillie, and Johan Van Rensbergen. 1991. "Watching Subtitled Television: Automatic Reading Behavior." *Communication Research* 18 (5): 650–66. https://doi.org/10.1177/009365091018005005
Ydewalle, Géry d', and Ingrid Gielen. 1992. "Attention Allocation with Overlapping Sound, Image, and Text." In *Eye Movements and Visual Cognition: Scene Perception and Reading*, edited by Keith Rayner, 415–27. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-2852-3_25.
Ydewalle, Géry d', 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.