Improving foreign language listening through subtitles

The effects of subtitle language and proficiency on Dutch high school and university students' perceptual learning in English

Nikki van Gasteren

12 April 2019

Master Thesis Linguistics
Radboud University Nijmegen
Supervisor 1: Dr. M.E. Broersma
Supervisor 2: E.R. Felker, MA

Acknowledgements

I would first like to thank my supervisors, Mirjam Broersma and Emily Felker. You were very flexible in supervising me even with my busy schedule including trips abroad and an emigration. Both of you gave me a lot of useful advice and I have learned a lot about how to do good research and how to report it properly. Mirjam, you have taught me the valuable lesson that the best results usually come from putting something away for a bit and thinking about it some more, and these results are often worth it, even if it takes longer to get there. Emily, thank you for teaching me all the basics of Python and linear mixed-effects modelling, including some cool tricks in R, and to not use "a lot" so often.

Secondly, I want to thank all the participants that took part in my study. A special thanks goes out to the Varendonck College, the school that made it possible for me test an enormous amount of high school students. I would especially like to thank Jasper ter Horst and Michiel Maasen, the two teachers that allowed me to test during their classes. You were both very accommodating and helpful, and without you I might not have been able to include high school students in this study. Also, a big thank you to all of the high school students that agreed to take part in my experiment and stayed patient throughout, even though most of the speech sounded like gibberish and the humour was not really to your taste.

Contents

Abstract

Understanding speech in a second language can be hard, as there are many sounds in this language that do not match with the sounds of a listener's native language (Best & Tyler, 2007). Luckily, listeners are able to adapt to foreign speech by shifting their phonetic patterns, using lexical information (Norris, McQueen, & Cutler, 2003). This process is therefore known as lexically-guided perceptual learning (Norris et al., 2003). Previous studies have found that watching videos in a foreign languages with subtitles can aid this process of perceptual learning (Birulés-Muntané & Soto-Faraco, 2016; Charles & Trenkic, 2015; Mitterer & McQueen, 2009). Two of these studies also showed that while foreign-language subtitles improve perceptual learning, native-language subtitles harm perceptual learning (Birulés-Muntané & Soto-Faraco, 2016; Mitterer & McQueen, 2009). All of these results are derived from experiments using participants with a high level of proficiency in their second language. However, the results might be different when using participants with a lower proficiency level.

This study investigates the influence of watching subtitled videos on perceptual learning in a foreign language, and whether language proficiency related to this influence. Four subquestions are investigated in this study: 1. What is the influence of the subtitle language on perceptual tuning in English of Dutch university students? 2. What is the influence of the subtitle language on perceptual tuning in English of Dutch high school and university students? 3. Does language proficiency modulate this possible influence? 4. If there is a learning effect, will this effect generalize to speakers who were not heard during the exposure phase?

These four subquestions were investigated by having Dutch high school and university students watch an English video. This video was accompanied by either English, Dutch or no subtitles. The participants' perception of English was measured using a dictation task before and after watching the video. These scores therefore give an estimation of the amount of perceptual learning that took place while watching the video. Both the video and the sentences where spoken by speakers with a Glaswegian accent.

The results show that using English subtitles and no subtitles leads to similar scores of speech perception. They also show that using Dutch subtitles might lead to less perceptual learning taking place. This seems to be the case mostly for listeners with a lower proficiency level. Listeners with a low proficiency level had lower scores in the Dutch subtitle condition than in the other conditions. The results were however not completely straightforward under closer inspection, making it difficult to interpret them.

It can therefore be concluded that the subtitle language relates to the

amount of perceptual learning that takes place when watching a foreign language video. Confirming previous research (Birulés-Muntané & Soto-Faraco, 2016; Mitterer & McQueen, 2009), native-language subtitles do not seem to work as well as using same-language subtitles or no subtitles at all. The subtitle language seems to matter most strongly for listeners of a lower proficiency level. Listeners with a lower proficiency level showed less perceptual tuning taking place than listeners with a higher proficiency level when using native subtitles. Listeners with a higher proficiency, however, did not seem to be impacted by the subtitle language.

1 Introduction

1.1 General introduction

Learning a foreign language is time-consuming and effortful. Therefore, it would be very interesting for second language learners to find a method for learning a new language with very little effort, while enjoying it at the same time.

One such method could be to watch films or television shows in a foreign language. This is a method that is recommended regularly, for example in online language learning communities (Kreisa, n.d.; Myers, n.d.). It is often advised to also turn on the subtitles, preferably in the target language, as this presumably helps with knowing what words are supposed to be said. The combination of video and subtitles is often called 'bimodal input', as the input is provided through two different modes of communication (Bird & Williams, 2002). Because the foreign language input is processed through two different modes, it is assumed that this makes it easier to learn the language (Clark & Paivio, 1991). Furthermore, watching the videos with subtitles does not cost any additional effort to the viewer: the subtitles are viewed and processed automatically (Koolstra, Peeters, & Spinhof, 2002).

However, it is still unclear if watching videos with subtitles actually improves foreign language ability. Many studies have focused on the effect of subtitles or bimodal input on comprehension of the video. Watching videos with same-language subtitles was found to improve comprehension of the video content (Alamri, 2016; Yoshino, Kano, & Akahori, 2000). However, these studies only measure how well subtitles can help with understanding the contents of the video, but not comprehension of the foreign language itself. Other studies have looked at the influence of subtitles on vocabulary learning and found that watching subtitled videos can increase foreign language vocabulary (Bird & Williams, 2002). These studies, however, only tell us something about the vocabulary and general listening comprehension aspects of foreign language learning. We do not know whether watching videos with subtitles could also improve speech comprehension: purely being able to recognize the correct sounds and words of the foreign language.

Only very few studies have looked specifically at the effect of subtitles on speech comprehension in a foreign language. Three of these studies found similar results. All three studies investigated the effect of watching subtitled videos on lexically-guided perceptual learning (Norris et al., 2003): the ability to shift phonetic categories using lexical information. Mitterer and McQueen (2009) investigated Dutch students who listened to an English video, featuring either an Australian or Scottish accent. They found that

students improved their English speech comprehension of the same accent by a larger degree if they had watched the video with English subtitles than if they used no subtitles. Using Dutch subtitles even led to less improvement than watching the video without any subtitles. Charles and Trenkic (2015) found similar results. Chinese students, living in the United Kingdom, who watched an English video in a standard British accent with English subtitles, improved their speech comprehension by a larger degree than students who watched the same video without any subtitles. These findings were confirmed again in a third study, by Birulés-Muntané and Soto-Faraco (2016). They found as well that Spanish students who watched an English video in a standard British accent with English subtitles improved their English speech comprehension more than students who had watched the same video with Spanish or without any subtitles.

The studies by Birulés-Muntané and Soto-Faraco (2016); Charles and Trenkic (2015); Mitterer and McQueen (2009) all show the same results: subtitles in the target language can help improve perception of a foreign language. However, their results do not answer the question whether subtitled videos can be used for earlier stages of language learning. All three studies used participants who had already reached a very high proficiency level in their target language. Therefore, we only know what the influence of subtitled videos is on speech perception of experienced language learners. If we would want to use watching subtitles videos as a language learning method, it is necessary to know how subtitled videos influence perceptual learning for learners of all proficiency levels.

1.2 Speech perception

1.2.1 Speech perception in the native language

Speech does not consist of neatly divided words but is a mostly continuous stream of sounds. Silences between words are not reliable clues to define word boundaries. Therefore, multiple steps have to be taken to be able to make sense of this stream: the stream has to be segmented into separate words, the individual sounds have to be distinguished, and the proper words have to be recognized. These processes do not operate individually but can also influence each other.

When listening to speech, listeners recognize certain characteristics of the sounds. Using these characteristics, they can recognize the sounds using the phonetic categories that exist in their native language. However, these separate sounds do not yet have any meaning to them, as the listener first has to distinguish which sounds belong to which words. Spoken utterances contain

few reliable markers for word boundaries (Norris, McQueen, & Cutler, 1995). Nevertheless, listeners are able to identify words from the continuous speech stream. According to Mattys, Jusczyk, Luce, and Morgan (1999), there are two kinds of cues that listeners can use to segment this stream. The first kind of cue that listeners can use is prosodic cues, like the stress pattern in speech. For example, according to Fletcher (1991) the accent of a word is always on the second syllable in French (as cited in Vihtnan, DePaolis, and Davis, 1998). Therefore, if a syllable is stressed, the listener can derive from this that the word ends after this syllable. The other cue that Mattys et al. (1999) mention is the phonotactic rules of a language. Phonotactic rules dictate which clusters of sounds are allowed in a language. If a listener encounters a cluster that is not allowed, there must by either a syllable or word boundary somewhere in between these sounds.

But to make sense of all these sounds it is also necessary to recognize the words. Only segmenting the sounds is not enough, retrieving the meaning is also necessary. It is assumed that spoken word recognition happens through spreading activation (see for example McClelland and Elman, 1986; Cutler, 2012; Frauenfelder and Tyler, 1987). This means that multiple word candidates are activated in the network that forms the mental lexicon. All possible word candidates compete with each other and the word with the highest activation is ultimately selected and recognized.

When attempting to recognize words, the initial group of candidates is selected based on the onset of the word (Allopenna, Magnuson, & Tanenhaus, 1998). For example, if a word starts with /b/, the word 'beaker' is still a possible candidate but 'flower' is not. The more information is given about a word, the smaller the group of word candidates will become. The candidates that are still plausible options will inhibit the activation of the candidates that do not conform to the new information. For example, if /beak/ has been presented, 'beaker' is still a candidate but 'beetle' is no longer a possible candidate. Ultimately, the candidate with the highest activation is recognized as the word (Huettig & McQueen, 2007). The word recognition process itself can also be another cue to segment the speech. This rests on the assumption that the competition between word candidates begins at many different points in the input stream and multiple candidates are processed at the same time (McClelland & Elman, 1986). If a word has been recognized in a cluster of sounds, it is clear where the onset and offset of this word are within the speech stream. This also gives information about where the previous word ended and where the next word starts. The word recognition process can also provide information about what phonemes have to be recognized: if not all sounds have been distinguished but there is only a single candidate left, the yet unrecognized sounds can now be recognized.

At least three different factors influence how much activation possible word candidates gain. Firstly, according to Dahan, Magnuson, and Tanenhaus (2001), words that are more frequent in a language get more activation. Because the words are frequent, there is a higher chance that these words will be the correct option. Therefore the words are said to have a higher "resting activation": they are already at a higher level of activation by default. Secondly, Yee and Sedivy (2006) argue that words that are semantically similar to the target word get more activation. When trying to recognize the word, not only the phonological information but also the semantic information is already available. This activation feeds back into the word recognition process, which may cause words that are related to the target word to get more activation, even if they do not conform with the phonological information that is already presented. Finally, according to Huettig and McQueen (2007), words for objects that have a similar shape as the target word get more activation. Both the meaning of the word and features belonging to this word are activated, again increasing the activation through feedback.

There are various models that try to explain how spoken word recognition works and try to capture the interplay between the processes of segmenting, phoneme recognition and word recognition. One of these models is the Merge model of spoken word recognition by Norris, McQueen, and Cutler (2000) which they explain as follows. Lexical (word) information and prelexical information (all information before the lexical stage) can "jointly determine phonemic identification responses". Prelexical processing continually feeds information to the lexical level. This happens strictly in a bottom-up fashion: information can only go forward from the prelexical processing to the lexical level, and information cannot be fed back to the prelexical stage. At the lexical stage, the prelexical information can be used to activate possible word candidates. Simultaneously, the same information is available for "explicit phonemic decision making": to decide what specific phonemes are in the speech signal. This decision making stage also receives information from the lexical level. Both lexical and lexical information can be merged to decide which phonemes actually represent the input. At both the lexical and the decision making level, there is competition between the word candidates and phonemic candidates respectively.

1.2.2 Lexically-guided perceptual tuning

There are three main difficulties that can arise when trying to comprehend speech. Firstly, different speakers produce the same word in a slightly different way. Even when a single speaker produces one word multiple times, the word is produced slightly differently each time (Blumstein & Stevens, 1981).

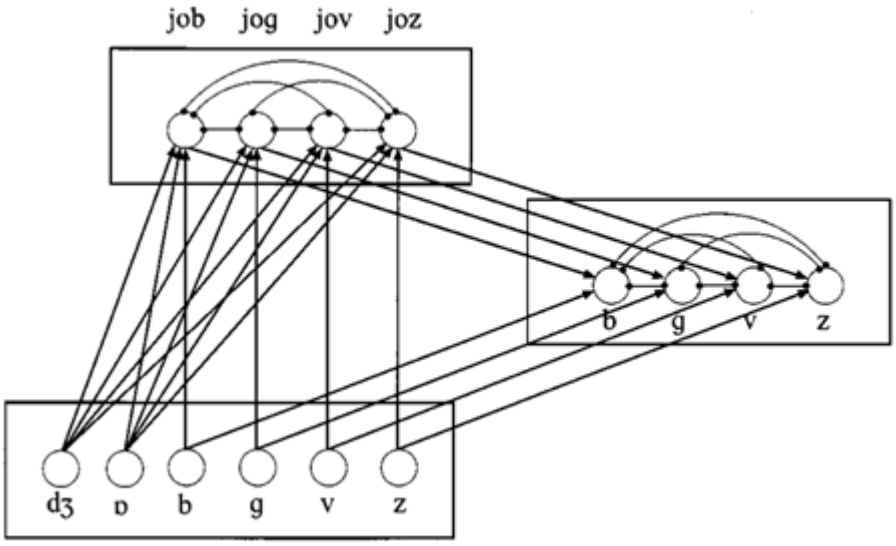


Figure 1: The Merge model of spoken word recognition (Norris et al., 2000)

A listener still has to be able to perceive this word, even when the pronunciation varies. Another difficulty might occur when listening to a speaker with a foreign accent (Bradlow & Bent, 2008). In this case, the variation that the speaker uses might be unexpected for the listener, for example if the speaker inserts speech sounds from their native language that the listener does not know. A third type of difficulty may arise when listening to speech in a non-native language (Best & Tyler, 2007). The perceptual system must be flexible enough to deal with variation in the speech signal to be able to process these variations (M. Baese-Berk, 2018).

Luckily, listeners are able to adapt to these difficulties. Despite learning the native phonetic categories at a young age, the categories can still change after they have been learned (Samuel & Kraljic, 2009). Language users can recalibrate their speech categorization to deal with new variation in speech input.

Shifts in perception can be caused by language-specific phonetic patterns, using both lexical and syntactic knowledge, but it can also be caused by speaker-specific phonetic patterns (M. Baese-Berk, 2018). In both cases, a period of exposure is necessary to adapt to these patterns. Norris, McQueen, and Cutler (2003) were the first to investigate this speaker-specific shift of phonetic patterns, which they describe as "lexically-guided perceptual learning". The core idea of lexically-guided perceptual learning is that the shift in perception is guided by lexical information. Lexical information can tell

a speaker what the word is supposed to be, and therefore what sounds the word has to consist of. The listener can use this information to determine what phonemes are supposedly uttered by a speaker. The listener can then update their phonetic categories accordingly.

In an experiment, Norris et al. (2003) had participants listen to words and non-words uttered by a single speaker. The words contained an ambiguous phoneme [?], halfway between the phonemes [s] and [f]. Participants were assigned to one of three different conditions and depending on the condition, this ambiguous phoneme occurred in different spots in the words. In one condition, the sound would always occur in the final position of non-words. In the second condition, the sound would replace the [f] in words ending on an [f]. In the third condition, the sound would replace the [s] in words ending on an [s]. After an exposure phase, listeners had to rate phonemes on the [s]-[f] continuum. The listeners from the [f] condition had the tendency to rate a larger proportion of the continuum as [f] compared to listeners from the non-word condition, while the listeners from the [s] condition did the opposite. This shows that even after a brief exposure period, listeners are able to shift the boundaries of their phonetic categories.

At least two follow-up studies replicated these findings. McQueen, Cutler, and Norris (2006) found that perceptual learning also generalizes to novel words that did not occur during the exposure or training. Eisner and McQueen (2006) found that perceptual learning is relatively robust and can last for a long time after the exposure.

Sometimes, shifting the phonetic categories of your native language is not enough to understand speaker specific phonetic categories. When trying to understand a second language, phonetic categories that do not exist in a listener's native language might occur in the speech signal (see paragraph 1.2.3). This can also happen when listening to the native language, when it is produced by a speaker with a foreign accent. In this case, the phonetic categories are influenced by the speaker's native phonetic contrasts. Bradlow and Bent (2008) have shown that it is difficult to adapt to these foreign accents, and that adaptation does not happen under all circumstances. In an experiment, Bradlow and Bent (2008) found that only participants who had been exposed to the same accent during both exposure and the post-test improved their understanding of this accent. Moreover, this adaptation seemed to be speaker-specific for the participants who had listened to a single speaker during exposure. However, the participants who had been exposed to the multi-speaker condition had managed to develop a speaker-independent adaptation to the foreign accent. Therefore they concluded that exposure to a specific foreign accent is necessary to adapt phonetic categories to this accent.

The same effect was replicated in another study by M. M. Baese-Berk, Bradlow, and Wright (2013). M. M. Baese-Berk et al. (2013) found that the effect did not only occur in lab conditions but was also found in listeners who had much of experience with foreign accents in their everyday life. This was also found by Witteman, Weber, and McQueen (2013), who showed that Dutch native speakers who had regular experience with native German speakers of Dutch were able to adapt to strong German accents better than speakers who did not have this experience.

It is however still unclear what listeners are adapting to exactly when listening to foreign accented speech (M. Baese-Berk, 2018). One possibility is that listeners adapt to general properties of non-native speech, which is only possible if non-native speakers of different native languages use similar strategies when speaking in their non-native language. Another possibility is that listeners expand their phonetic categories so they include a larger variety of speech sounds. These hypotheses have however not been tested as of yet.

The discussed studies show what happens when shifting the boundaries of phonetic categories in the native language. However, this process looks slightly different for phonemes in a non-native language.

1.2.3 Speech perception in a second language

When listening to a foreign language, non-native listeners run into multiple problems. These problems can occur at the phoneme level but also at the word level.

Categorical perception is language-specific. Therefore, naïve non-native listeners, or functional monolinguals, have difficulty recognizing and categorizing phonemic contrasts that do not occur in their native language (Best & Tyler, 2007). However, not all contrasts are equally difficult. The different types of contrasts are explained by Best (1993, 1994) in the Perceptual Assimilation Model (PAM) (as cited in Best and Tyler, 2007). Contrasts that are "assimilated" as phonetically similar to contrasts in the native language, are easy to discriminate. Some other phonemes are also easy to distinguish, especially if new phonemes are not at all similar to any phonemes present in the native language (e.g. African click sounds can be easy to recognize for native English speakers). However, if a foreign contrast does not assimilate to a native contrast, it is categorized as an existing contrast that may or may not be a good fit for the foreign contrast. Contrasts that map onto two different native categories are easy to discriminate, while contrasts that map onto a single category are hard to discriminate. Sometimes, one phone maps onto a category while the other does not, which leads to in-between

results. Phonemes that cannot be categorized are easy to distinguish from those that can be categorized. It may however be difficult to discriminate between multiple uncategorized phones if they are phonetically similar.

While PAM explains the difficulties naïve listeners face when discriminating phonemes in a foreign language, more experienced second language learners face slightly different problems. In their case, the phonological systems of the L1 and L2 are not completely separate. According to Flege (1995), the Speech Learning Model (SLM) aims to explain how second language learners tune into the phonology of their second language (as cited in Flege, MacKay, and Piske, 2002). Instead of assimilating pairs of phonemes, single phonemes can be assimilated into existing categories or new categories can be created. If a listener encounters a phoneme that is identified as an existing phoneme, it is "equated" to the existing category. If a phoneme is phonetically distant enough from existing categories, a new category can be created. There is a maximum capacity for categories however: as more categories are created, the phonetic space gets more crowded and the categories get closer to each other. New sounds therefore have a higher chance of being assimilated. If a sound is at first assimilated to an existing category but is audibly different, the representation will be modified over time, resulting in a "composite" category of L1 and L2.

In addition to difficulties at the phoneme level, problems can also occur at the word level. Weber and Cutler (2004) describe two main problems that can occur at the word level: interlingual and intralingual competition. Because of this competition, second-language listeners have a larger pool of activated word candidates. Interlingual competition means that not only word candidates from the L2 lexicon are considered, but also candidates from the L1. Additional problems are caused by the listener not being able to distinguish the non-native phonetic contrasts. This means that a non-native listener also considers word candidates from the L2 that would not be considered candidates by native speakers (intralingual competition), because the onset does not match exactly. Moreover, non-native listeners can even consider words that do not exist in the L2 because of these missing phonetic contrasts. This leads to what Weber and Cutler call "phantom word activation".

It is clear that non-native listeners face many difficulties during listening, both at the phoneme level and the word level. However, they can become better L2 listeners as they receive more exposure to the L2. It is not always easy for a learner to receive more exposure to the L2, for example when the L2 is not spoken in the country that the learner lives in. There are however some possibilities for second language learners to create more exposure to the L2 by themselves, for example by watching television shows in the target language.

1.3 Subtitles and second language learning

1.3.1 General use of subtitles

In many European countries, television programs and films are imported from other countries. These programs and films are therefore often in a different language than what is spoken in the country itself. To make them available to the general public, they have to be translated in some way. There are three ways in which this is usually done: subtitling, lip-sync dubbing and voice-over or lectoring. Most European countries use either dubbing or subtitling, and sometimes a mix of both (Kilborn, 1993).

There are two main types of subtitling: same-language subtitles (also called within-language, bimodal or intralingual subtitles), and translations or standard subtitles. Intralingual subtitles are primarily aimed towards the deaf and hard-hearing (sometimes called 'closed captions') (Burnham et al., 2008; De Linde & Kay, 1999). These often also include descriptions of non-dialog audio. These subtitles are also used for clarifying speech that is spoken in a strong regional accent or in video fragments with large amounts of noise.

Subtitles are usually not a direct translation of all the spoken dialogue. For example, Dutch subtitles for English television programmes contain about 30% fewer words than the original dialogue, according to Koolstra et al. (2002). This is done to reduce the amount of text and to avoid redundant information. Also, it is difficult to literally translate idiomatic expressions and metaphors, which means that these have to be replaced, adapted, extended or omitted completely (Pedersen, 2017).

Contrary to popular belief, reading subtitles does not require much effort (Koolstra et al., 2002). Van Driel (1983) claims that it is often assumed that subtitles could harm the viewing experience in multiple ways (as cited in Koolstra et al., 2002).The subtitles take up space on the screen, which could mean that not all of the picture is visible enough. Also, if viewers are reading the text, they might not be able to pay attention to the pictures. This could even mean that viewers get tired by continually reading the text.

All of these claimed disadvantages of subtitles have been opposed by research. Firstly, the subtitles are placed on the bottom of the screen, while the main focus of the pictures is usually in the middle (Koolstra et al., 2002). Subtitles are also not displayed at all times throughout the video, and even when they are, it is usually possible to 'look through them'. Secondly, eye-tracking studies have shown that viewers do not consciously read the subtitles. d'Ydewalle, Van Rensbergen, and Pollet (1987) found that the viewers gaze moves towards the subtitles as soon as they are presented. Gielen (1988) showed that viewers usually look at the screen at a point just above

the subtitles, which makes it possible to read the subtitles and see the most important events on the screen (as cited in Koolstra et al., 2002). Moreover, d'Ydewalle, Praet, Verfaillie, and Rensbergen (1991) showed that viewers are able to constantly move their gaze back and forth from the subtitles to the picture. In fact, viewers' eye movements were similar when watching native videos that were subtitled in a foreign language (d'Ydewalle et al., 1991) and when watching native videos with same-language subtitles (d'Ydewalle et al., 1987), suggesting that attention is automatically drawn towards the subtitles. Subtitles are also processed automatically and efficiently, as reading is usually faster than listening (d'Ydewalle et al., 1991). The study by Gielen (1988) showed that viewers are able to quickly recognize the correct subtitles (as cited in Koolstra et al., 2002). A follow-up by Koolstra, van der Voort, and d'Ydewalle (1999) showed that children are also able to recognize subtitles quickly, and that the ability to do this increases with higher reading comprehension.

1.3.2 Effects of subtitles on native language abilities

Some studies have shown that watching videos with subtitles can have positive effects on native language abilities. Linebarger, Piotrowski, and Greenwood (2010) investigated the effects of watching videos with closed captions on the literacy development of children living in poverty. In their study, they investigated third-grade American pupils who spoke English as a native or a second language. The children were assigned to one of two groups. One group watched six videos with captions, while the other group watched uncaptioned videos. Linebarger et al. (2010) found that during the post-test, children who watched the videos with captions performed better on a word recognition task and a word reading task. These children also learned the meanings of the words better. A similar finding came from two other studies (Kothari, Pandey, & Chudgar, 2004; Kothari, Takeda, Joshi, & Pandey, 2002). In these studies, Indian children who watched educational song programs with subtitles were found to have improved literacy rates compared to children who watched the programmes without subtitles.

Watching subtitled videos does not only improve the native language ability of children, but also of adolescents and even adults. Firstly, Davey, Parkhill, et al. (2012) found that watching subtitled videos could help with improving reading comprehension and vocabulary in teenagers from families with a low socioeconomic status. Similarly, Griffin and Dumestre (1993) found out that watching subtitled videos could aid sailors in improving their vocabulary and reading skills. Secondly, adults who are already highly literate can benefit from captions or subtitles, although not necessarily specif-

ically for language. Brasel and Gips (2014) found that when adults watch captioned television commercials, they remember the brand names better. Moreover, Steinfeld (1998) discovered that students who watched captioned recordings of lectures were better at remembering the contents of these lectures than students who watched the same lectures without any captions.

These studies all show that subtitles can aid people in improving various domains of their native language skills. However, subtitles can also be of help when learning a second language.

1.3.3 Subtitles and listening comprehension in the L2

During the 1980s, captions aimed at the deaf and those hard of hearing became more prominent on television and in language learning classrooms. Language teachers began to use them as a resource to help students improve second language literacy and listening comprehension (Vanderplank, 2010). Various studies have been conducted in which the effects of same-language subtitles and standard subtitles on memory and listening comprehension have been investigated.

For example, Yoshino, Kano, and Akahori (2000) looked at how well Japanese students could remember the contents of English videos. Students watched the same video twice, one time with Japanese and one time with English subtitles. After watching the version with English subtitles, the students were better at remembering the contents of the video than after watching with Japanese subtitles. Moreover, students who watched the video with Japanese subtitles did not remember the contents better than students who only received audio input.

Same-language subtitles do not only aid in better recall of the video contents, they can also improve listening comprehension. Huang and Eskey (1999) investigated the effect of subtitled videos on listening comprehension of students who had learned English as a second language. Students who watched an English video with English captions performed better than students who watched the video without captions at both a reading comprehension and a listening comprehension test.

Similar effects were found in three more studies. The first study is the study by Alamri (2016). In their first experiment, native speakers of Arabic who had learned English as a second language watched a video with either English, Arabic or no subtitles. In the first experiment, participants who had watched the video with English subtitles performed better at a comprehension post-test than participants who had seen the video with either Arabic or no subtitles. In a second experiment, participants had to watch an unsubtitled pretest video before watching a video with again either English, Arabic

or no subtitles, and they watched another unsubtitled video four weeks later. Participants who had seen the video with English subtitles performed better at the post-test than the other participants, suggesting that watching the video with subtitles had led to a long-term learning effect as well.

The second study that found similar results was done by Hayati and Mohmedi (2011). In their study, students watched an English video of about five minutes every week for a period of six weeks. The video was accompanied by either English, Persian or no subtitles. After each session, the participants did a short listening comprehension task. When averaging the listening comprehension scores of all six sessions, the participants who watched the video with English subtitles performed better than the other groups, and the group with Persian subtitles performed better than the group without subtitles. However, it is unclear how these scores changed over the course of the six weeks, as these results were not provided by the authors.

The final study was done by Yang and Chang (2014). Students watched an English video with either full keyword captions, reduced keyword captions or annotated keyword captions. The latter type contained a pictorial symbol that assigned the keyword to one of four reduction categories: assimilation, liaison, reduced sound and omitted sound. Before and after watching the video, the participants did a general listening comprehension task and a task in which they had to recognize reduced forms of words. All groups improved on both of the tasks, but the participants who had watched the video with the annotated keyword captions performed the best on both tests.

These studies show us that watching subtitled videos can improve comprehension of videos, and that these videos can also aid language learners in improving their listening comprehension abilities, also for new videos and over a longer period of time. However, they do not show whether participants also learn vocabulary or phonology of the foreign language.

1.3.4 Subtitles and vocabulary learning in the L2

According to Koolstra, Peeters, and Spinhof (2002) it is often assumed that people from so-called "subtitling countries" are better at foreign languages (i.e. English) than people from "dubbing countries". The reason for this assumption is that being able to listen to the original, foreign speech, provides additional exposure to the foreign language. The subtitles can provide more information about which words are supposed to be said. Studies investigating the influence of watching subtitled videos on vocabulary learning have so far had mixed results.

Three studies have looked at the effects that watching subtitled videos might have on the vocabulary of children. The first study was done by

d'Ydewalle and Van de Poel (1999) and looked at vocabulary learning in children. Native Dutch children of 8-12 years old watched a video with subtitles. The children who watched the video with a foreign speech track and Dutch subtitles acquired some receptive vocabulary both in the written and auditory domain, while the children who watched the video with Dutch speech but foreign subtitles only learned to visually recognize words. There was however no condition with both foreign speech and subtitles in this study. The videos were also not full motion videos but were still-motion videos.

The second study was conducted by Koolstra and Beentjes (1999). In their study, Dutch native children watched either an English video with no subtitles, an English video with Dutch subtitles, or a Dutch video without subtitles. The children in the Dutch subtitle condition learned more vocabulary than children in the no subtitle condition or spoken Dutch condition.

Finally, there was a study that used self-report data to investigate the effects of watching subtitled videos on vocabulary learning in children. Kuppens (2010) investigated primary school pupils and compared their self-reported exposure to subtitled foreign videos and other media to their scores on an oral translation tests (one Dutch to English and one English to Dutch). On average, the children that spent more time on watching subtitled English videos performed better on both translation tests. The effect was larger for female pupils than males.

Four other studies have looked at the influence of subtitles on L2 vocabulary learning in adults. Three of these however did not actually show videos but instead used pictures. For example, Bird and Williams (2002) found that bimodal presentation helped with remembering both words and non-words compared to audio-only presentation. Bisson, Van Heuven, Conklin, and Tunney (2013, 2015) found similar results in a slightly different set-up. Participants who had to listen to words and had a task to learn the words explicitly or implicitly both improved their foreign language vocabulary.

Finally, one study did look at subtitled videos and adult L2 vocabulary learning. Mousavi and Gholami (2014) compared adults who watched a subtitled video in a foreign language to adults that read a plain text in the foreign language. The group that had watched the subtitled video acquired more vocabulary than the group that had read the text.

1.3.5 Subtitles and speech comprehension in the L2

Only very few studies have looked at the effects of subtitles on speech perception of a second language. These studies will be discussed in full.

The first study investigating the effect of subtitles on foreign speech perception was done by Mitterer and McQueen (2009). In their study, partici-

pants watched a video of 30 minutes. The video was spoken in English with either a Scottish or an Australian accent. This video was accompanied by either English or Dutch subtitles or no subtitles at all. All participants were monolingual, native speakers of Dutch who had learned English as a second language. After watching the video, participants did a shadowing task. In this task, they heard sentences from either the video they just watched, or sentences from a different part of the movie/episode that they did not watch before. The sentences were scored by a rater, giving a point for each word that was repeated correctly. Also, participants listened to sentences from the other video (with the other accent) to act as a non-exposure control group for the other video. Mitterer and McQueen (2009) found that participants who had watched the video with English subtitles performed better at the post-test than participants who had watched the video wit Dutch subtitles or without any subtitles. Moreover, participants who had Dutch subtitles during exposure seemed to score worse than participants who did not have any subtitles during exposure. They did find a difference between items that the participants had heard before and new items. For post-test sentences that had also occurred in the video, participants who had seen the video with Dutch subtitles performed similarly to participants who did not have any subtitles during exposure. In contrast, for post-test sentences that were completely new, a negative effect for the Dutch subtitle group was found. All participants also listened to sentences that were spoken with the accent that they did not hear during the exposure phase. The results from these sentences showed that it is necessary that exposure with subtitles only improves speech perception for that specific accent.

From this study, it seems that subtitles in the target language aid speech perception in a second language. Additionally, subtitles in the mother tongue do not seem to have any effect, or even have a negative effect compared to not having any subtitles. Mitterer and McQueen (2009) explain these findings in the following way. The subtitles presented on the screen are processed automatically, and the phonological knowledge about the words in the subtitles is automatically retrieved. When watching an English video accompanied with English subtitles, the phonological knowledge matches the phonological information that is transferred through the speech. However, when watching the video with Dutch subtitles, there is a mismatch between the information from the subtitles and the information retrieved from the speech signal. This inconsistency makes it more difficult for the listener to retune their perception compared to having consistent information (in the case of English subtitles) or only the information from the speech signal (in the case of having no subtitles at all).

Some aspects of using subtitles to improve speech perception were not

yet investigated in this study. Instead of a pre-test, a non-exposure control group was used. Therefore we only know differences on a group level, but not how foreign speech comprehension of the individual participants improved. Secondly, a multi-speaker video was used for the exposure phase but only one speaker was used in the post-test. Hence, we cannot know whether the improvement in speech perception was speaker-specific or was also generalized to other speakers. Additionally, the English proficiency level of participants was not measured. This means that we can only make assumptions about the exact proficiency level of the participants. Finally, they solely tested that are assumed to have a very high level of proficiency in English already. Therefore we cannot generalize these findings to second language learners with a lower proficiency level in their target language.

Following the study by Mitterer and McQueen (2009), a study appeared that investigated the use of subtitled videos and foreign speech perception in children. Ghorbani (2011) did a case-study on one child. One Iranian 12-year-old boy watched a selection of 20 cartoons over a period of two years. He was allowed to watch each cartoon as often as he wanted and could select either English, Persian or no subtitles himself. After "mastering each cartoon", his language proficiency was tested using various different tests that measured perception and production. After two years, the boy was reported to be fluent in English with nativelike pronunciation.

While the results of this study seem promising, it is difficult to draw any sound conclusions from them. Firstly, the study was a case-study with only one child. It is certainly possible that subtitled videos have aided this child in learning English, but it is not possible to generalize these findings. Also, the study does not report on what specific kinds of tests were used and what the results were for each separate test. Furthermore, no statistical analysis has been reported. The progress was judged holistically by the author, who is not a native speaker of English himself. While the quality of this study is questionable, it does show that watching subtitled videos might have a positive influence on learning a foreign language and that watching the videos can be a pleasant experience for a learner.

Charles and Trenkic (2015) have followed up on the study by Mitterer and McQueen (2009). They wanted to look further into two aspects of the study. Firstly, instead of using television shows or movies, they used an educational video as they found this was better suited for university students. Secondly, they used a design in which participants were exposed to the video over multiple sessions. Participants in their study watched a new video every week for four consecutive weeks. After each session, there was a post-test. Their participant group consisted of university students with an average score of 7 on the IELTS listening test, roughly equivalent to the low end of the

C1 level of the CEFR (IELTS, n.d.). To investigate the learning effect, a listening shadowing test was used as a post-test. Like Mitterer and McQueen (2009), Charles and Trenkic (2015) found that English subtitles aided the participants in their speech perception. They found this result for each of the four weeks of testing.

This study also suggests that subtitles could help improve foreign speech comprehension. However, a few notes about the design have to be made. As opposed to Mitterer and McQueen (2009), Charles and Trenkic (2015) did not have a native-subtitle condition. Therefore, we can only draw conclusions about the effects of same-language subtitles and not about native-language subtitles from this study. Another difference was that the participants who were tested in this study had been living and studying in the United Kingdom for an average of eight months. This means that, in contrast to the participants in the Mitterer and McQueen (2009) study, the participants had been exposed to English daily during these months. Therefore they might have already adapted more to the English accent than the participants of Mitterer and McQueen (2009) before taking part in the study. Finally, like Mitterer and McQueen (2009), Charles and Trenkic (2015) used only university students in their study. The proficiency level of the participants was relatively high, so the results of this study cannot be generalized to participants with a lower proficiency level.

Another follow-up study was done by Birulés-Muntané and Soto-Faraco (2016). Like Charles and Trenkic (2015), they adapted the study by Mitterer and McQueen (2009), but adapted another variable instead. The video they used during the exposure phase was double the length, an hour in total. Participants did not only take a listening test afterwards but also a vocabulary test. Furthermore, instead of a control group without any exposure, they used a pre-test. The vocabulary test was inconclusive, but the results of the listening test again confirmed the finding by Mitterer and McQueen (2009). Participants who had watched the video with English subtitles scores better than the other participants, and participants who had watched with native/Spanish subtitles improved less than participants without any subtitles.

This study answers some questions that were left after the studies by Mitterer and McQueen (2009) and Charles and Trenkic (2015). By using a pre- and post-test, the actual progress per participant could be measured, instead of comparing the participants to a non-exposure group that did view an English video but with a different accent. Additionally, we now know that the effect is similar for half an hour and an hour of exposure. Also, the participants tested in this study had a slightly lower proficiency level in English than in the previous two studies. While the proficiency level was only

slightly lower, we now know that subtitles can also aid language learners of a slightly lower proficiency level. Finally, Birulés-Muntané and Soto-Faraco (2016) did not use a shadowing task like the previous studies, but used a gap-test. Participants listened to a 180-word excerpt, for which they were also provided with a written transcript. From this transcript, 24 words had been removed, creating 1-word gaps. They were then tasked with filling in the gaps while listening to the excerpt. While this did make it possible to make scoring more objective, using such a task also has its disadvantages. The participants could read most of the text, which might have made it easier to use the context to predict what words had to fit the gaps. The participants were therefore only required to comprehend single words, instead of complete phrases or sentences.

1.4 Current research

1.4.1 Research gap

While we know that subtitles possibly have a positive influence on learning a foreign language, not much is known about the effects of watching subtitled videos on speech comprehension in a foreign language. The research by Mitterer and McQueen (2009), Charles and Trenkic (2015) and Birulés-Muntané and Soto-Faraco (2016) has shown us that same-language subtitles seemingly contribute positively to perceptual tuning in a foreign language, while native-language subtitles do not provide additional help or might even harm perceptual tuning. However, these are only very few studies, so it is still hard to generalize the findings to all language learners in general or to advise language learners to use watching subtitled videos as a method to improve their foreign language abilities.

Moreover, all three studies have only used participant groups who are already very proficient in their second language, with levels ranging from B2 to C2. This means that the results may only apply to language learners who are already very experienced in their language and not necessarily for beginning or intermediate learners. It could be argued that it is even more important to find out what the effects of watching subtitled videos are for this group, as this group of language learners might be even more inclined to want to use this method to improve their foreign language.

Perceptual tuning of listeners with a lower proficiency level could be influenced by watching subtitled videos in three different ways. The first possibility is that the effects for learners with a lower language proficiency level are similar to the effects that subtitled videos have for learners with a high language proficiency in their L2. Another possibility is that learners with

a lower proficiency level improve even more than experienced learners when watching subtitled videos. In this case, there is more room for improvement in the foreign language. The final possibility is that less experienced learners do not improve tune into the L2 speech as much as more experienced learners do when watching subtitled videos. This would be the case if there is a certain threshold that learners have to reach before being able to use watching videos as a language learning method. It might be necessary to have for example a minimum knowledge of the L2's vocabulary, to be able to use the subtitles to tune into the language. There does however not seem to be any existing literature suggesting that such a threshold exists.

The language of the subtitles can also have a few different effects on comprehension of foreign speech. The first possibility is that similarly to the previous studies, learners would get the most out of watching videos if they watched the videos with same-language subtitles. Mitterer and McQueen (2009) propose that same-language subtitles provide additional phonological information, while native-language subtitles provide information that is inconsistent with the information retrieved from the speech signal. This theory could also be applied to listeners with a lower proficiency level. For them, same-language subtitles also provide additional information and native-language subtitles might be confusing. It is however possible that the English subtitles do aid perceptual tuning. For example, it might be a possibility that if a listener's foreign language proficiency is still too low, the listener does not possess much phonological information about the language yet. It is also possible that the speech already provides enough input and the subtitles do not add any additional information. In both cases, watching the video with English subtitles would not aid perceptual tuning more than when watching the video without any subtitles. The same consideration has to be made for Dutch subtitles. It is possible that Dutch subtitles also result in less perceptual tuning than other subtitle conditions, as predicted by the theory of Mitterer and McQueen (2009). However, it is also possible that using Dutch subtitles leads to even worse results for listeners with a low proficiency compared to listeners with a high proficiency. Low proficiency listeners might be more dependent on the subtitles to understand the video, and might therefore receive more distracting information. The opposite might also be possible: somehow, listeners with a low proficiency level are influenced less by the contradicting information of the Dutch subtitles, and therefore they tune in to the English speech more than high proficiency listeners. This scenario seems however unlikely, if the theory by Mitterer and McQueen (2009) is correct.

1.4.2 Research question and hypothesis

The general research question of this thesis is as follows: *What is the influence of watching subtitled videos on perceptual tuning in a foreign language, and does language proficiency modulate this influence?*

The subquestions of this research question are:

1. What is the influence of the subtitle language on perceptual tuning in English of Dutch university students?

2. What is the influence of the subtitle language on perceptual tuning in English of Dutch high school and university students?

3. Does language proficiency modulate this possible influence?

4. If there is a learning effect, will this effect generalize to speakers who were not heard during the exposure phase?

The hypotheses for four subquestions are:

1. English subtitles aid perceptual tuning as compared to no subtitles or Dutch subtitles. Dutch subtitles have either no effect or a negative effect. This result is expected because this question is a replication of previous studies (Birulés-Muntané & Soto-Faraco, 2016; Charles & Trenkic, 2015; Mitterer & McQueen, 2009), in which this result was found as well.

2. For learners of all proficiency levels, English subtitles aid perceptual tuning as compared to no subtitles or Dutch subtitles. Dutch subtitles have either no effect or a negative effect. This result is expected because, according to the theory of Mitterer and McQueen (2009), English subtitles provide the same phonological information as the speech signal. Combining the information from the text and the speech signal can therefore lead to improved speech comprehension. Dutch subtitles on the other hand provide a mismatch in information.

3. Students with a lower proficiency level might improve after exposure, but cannot use English subtitles as well as students with a higher proficiency level. This is expected because I assume that there is some kind of threshold proficiency level that is needed to fully use subtitles for lexically-guided perceptual tuning. It is also expected that students with a lower proficiency level are hindered more by Dutch subtitles than students with a higher proficiency level. This is hypothesized because

I assume that students with a lower proficiency level will rely more on the subtitles to understand the video than students with a higher proficiency level. Because the students with a lower proficiency level focus more on the subtitles, they might also receive more contradictory information that is derived from these subtitles.

4. Students will tune in to the speaker from the video the most. If there is a learning effect, it might generalize to new speakers, but students will have more difficulty understanding the new speakers. This result is expected because the video is a single speaker video. Following the results of Bradlow and Bent (2008), single speaker exposure will not help in understanding new speakers of the same accent.

1.4.3 General outline of the current study

The current research was roughly based on the methods by Mitterer and McQueen (2009). In the current study, participants had to watch a video of about 15 minutes. All speech in the video was in English, produced by a speaker with a Glaswegian accent. This video was accompanied by either English or Dutch subtitles or no subtitles at all. Before and after watching the video, participants did a listening pre- and post-test. In these tests, they listened to English sentences, also produced by speakers with a Glaswegian accent, and had to write down what they heard. After the post-test, they took part in a brief comprehension test and a background questionnaire.

The participant group consisted of Dutch university students and Dutch high school students from different levels and cohorts, who had learned English as a second language. The high school students were from different levels of the Dutch education system (HAVO/VWO). Some of the VWO students were in a bilingual programme (TTO). Students from both the 4th and 5th year of high school participated. No direct group to group comparison was made in this study as the individual differences within a group were quite large, but recruiting participants from a wide range of levels of education ensured a wider range of proficiency levels in the participant pool.

During the exposure phase, the participants watched excerpts from a sketch show called *Limmy's show*. This is a television show by the BBC, featuring multiple sketches in each episode that are often dark or bizarre in nature. The main character in the show is 'Limmy', played by Brian Limon, or other characters played by the same actor. Some minor characters are played by different actors. All characters in the show speak with a Glaswegian accent. This show was chosen for the exposure phase to make the video entertaining for both university and high school students. Also, Charles and

Trenkic (2015) showed in their first experiment that comedy was the easiest genre to listen to. All of the excerpts featured only the main character.

Before and after the exposure phase, the participants did a dictation task. During each dictation task, the participants listened to sentences that were also taken from *Limmy's Show*. The sentences did not occur in the video that was used during the exposure phase. The participants could listen to each sentence twice and had to write down what they thought was being said in each sentence. Half of the sentences were spoken by the main character while the other half was produced by two different characters from the show.

The task was presented using the online survey software Qualtrics. This made the experiment portable to the schools, and also made it possible for participants to go through all the different tasks by themselves without any help from the experimenter.

The methods differ from previous studies in a few aspects. The first change was the use of different participant groups. In previous studies, only university students were tested. These were all learners who had already reached a high level of English proficiency, between the B2 and C2 level of the CEFR (Council of Europe, 2001). In this study, both university students and high school students participated. By doing this, the total range of proficiency levels was widened as compared to previous studies. Although the exact proficiency levels of the participants were not measured, it is estimated that the levels are between the low end of the B1 up to the C2 level across the complete pool of participants (Europees Referentiekader Talen, n.d.).

Secondly, different video and sentence materials were used. Moreover, the video was only half as long. This was necessary to be able to test high school students, as the total length of the experiment had to fit within the time scheduled for one class. However, it is still expected that this is enough exposure to the speech in the video to lead to perceptual learning. Multiple studies have shown that perceptual learning can already take place after only a minute of exposure (Clarke & Garrett, 2004; Sidaras, Alexander, & Nygaard, 2009). Therefore, the effect in this study might be smaller, but is expected to still be present.

Furthermore, the comprehension test is different from previous studies. The post-test used by Mitterer and McQueen (2009) was adapted and a pre-test was added, like was done by Birulés-Muntané and Soto-Faraco (2016). This study used a dictation task as opposed to the sentence repetition task. This change was made so it was possible to test multiple participants at the same time in the same room, as this was necessary to test the high school students. Furthermore, this allowed for automated and more objective scoring instead of relying on raters. Secondly, the comprehension test was not only presented as a post-test, but also as a pre-test (like Birulés-Muntané

and Soto-Faraco, but they used a gap task). This made it possible to measure the progression/perceptual learning of each individual participant, instead of only knowing their final result. This was deemed important as there may be large differences between individual participants.

The final change was the use of multiple speakers in the pre- and post-test. Mitterer and McQueen (2009) and Birulés-Muntané and Soto-Faraco (2016) both presented a multi-speaker video. Mitterer and McQueen opted to use sentences spoken by the main character for their post-test, while Birulés-Muntané and Soto-Faraco used sentences produced by two characters who did not appear in the video. Charles and Trenkic (2015) used single-speaker videos, but the speaker was different for each exposure. All of the speakers occurred during one of the post-tests. In this study, a single-speaker video is combined with a multi-speaker pre- and post-test. During the exposure phase, only one speaker is heard. In the pre- and post-test, sentences are uttered by multiple speakers, one of which is the same speaker as during the exposure phase and two of which are new speakers with the same accent. This change was made to make it possible to investigate whether any effects also transfer to other speakers of the same accent or whether they are speaker-specific when using a single-speaker exposure phase.

The general outline for the rest of the thesis is as follows. Chapter 2 will be about the methods that were used for the experiment. Chapter 3 will discuss the statistical analysis and the results that followed from this analysis. In Chapter 4, these results will be interpreted and discussed. Some limitations of the study and ideas for future research are also included in this chapter.

2 Methods

2.1 Participants

A total of 102 participants took part in the study. The participant group consisted of high school students and university students. 31 of the participants were male, 71 were female. The average age was 18.74 years (sd = 3.16, range = 14.90-27.08). See Table 2.1.1 for information about the genders and ages in each subgroup of participants.

All participants were required to be monolingual, native speakers of Dutch, to have no language problems including dyslexia, have normal or corrected-to-normal vision, have normal or corrected-to-normal hearing, and have no attention problems.

High school students were recruited from a school the province of Noord-Brabant in the Netherlands. This school offers a regular Dutch track and a bilingual English track. The high school students participated in the study as a class activity for their English classes. They did not receive compensation for participating. All students were in either the 4th or 5th grade of high school. The 4th grade students were from either the HAVO or VWO track. The 5th grade students were from either the VWO or the TTO track. There were no requirements for the high school students to participate in the study, as the experiment was presented as a class activity. However, high school students that did not meet the criteria were excluded from analysis. The high school students were asked for their consent to use the data after completing the task. The data of high school students that did not consent to the use of their data was deleted.

University students were recruited through the SONA recruitment system of Radboud University. Students were only eligible to participate if they were between 18 and 30 years old and were registered as a student at Radboud University at the time of participation. Students of all programmes offered by Radboud University were allowed to participate, as long as they did not study English language and culture. The students gave informed consent before participating in the experiment. They received either a gift card of €7.50 or participant credit for their participation.

An additional 25 participants were tested but had to be excluded from analysis because of technical issues, not answering enough comprehension questions correctly (see paragraph 2.2.3 for an explanation), skipping part of the video, completing the dictation task in Dutch instead of English, because their responses in the background questionnaire showed that they did not fulfil the participant criteria, or because they did not give permission to use their data.

Table 1: Gender and age information of the participants

Education type	n	Age (mean, sd)	Age (min-max)	# of each gender
4HAVO	16	16.00 (0.53)	14.90-16.90	12 female, 4 male
4VWO	14	15.70 (0.40)	15.11-16.43	8 female, 6 male
5VWO	14	16.89 (0.53)	16.08-17.81	8 female, 6 male
5TTO	12	16.63 (0.27)	16.28-17.05	7 female, 5 male
University	46	21.73 (2.32)	18.29-27.08	36 female, 10 male
(Total)	102	18.74 (3.16)	14.90-27.08	71 female, 31 male

2.2 Materials

2.2.1 Dictation task (pre- and post-test)

Before and after watching the video, all subjects performed a dictation task. Both the pre- and post-test consisted of 16 spoken sentences each. The sentences in the post-test were different sentences than those in the pre-test. Sentences for the post-test were selected from the first season of *Limmy's Show*. This is a television show that consists of various sketches, usually performed by the main character Limmy but also featuring other characters. Half of the sentences were produced by characters played by Brian Limon, the leading actor of *Limmy's Show*. The other sentences were produced by two different speakers that also appeared in the sketches, but that did not speak in the video. Both of these speakers were male, native speakers of English with a Glaswegian accent.

Sentences were selected by a native speaker of Dutch with a high proficiency in English (the author). The sentences were selected using the following four criteria, which were all judged subjectively. Firstly, the sentence had to consist of either five or six words to prevent as much influence as possible from differences in short-term memory. Secondly, sentences had to consist of words that were considered as common enough for both the high school students and university students. Sentences were also not allowed to contain words that seemed to be Scottish dialect. Thirdly, only sentences that did not contain a lot of background noise were selected. Finally, sentences had to be produced fluently without any pauses within the sentence.

In some cases, full sentences were used, while in others only phrases were used. Some sentences were taken directly from the sketches, while in some sentences silences or single words were edited out to make the sentence conform to the selection criteria. The final selection of sentences was checked by another native speaker of Dutch with a high proficiency in English and a native speaker of English (the supervisors). The final selection contained both sentences that were easy to understand and sentences that were difficult

to understand (e.g., differences in clarity of articulation), to prevent participants from getting demotivated while also preventing them from reaching ceiling level. The difficulty of the sentences was also judged subjectively.

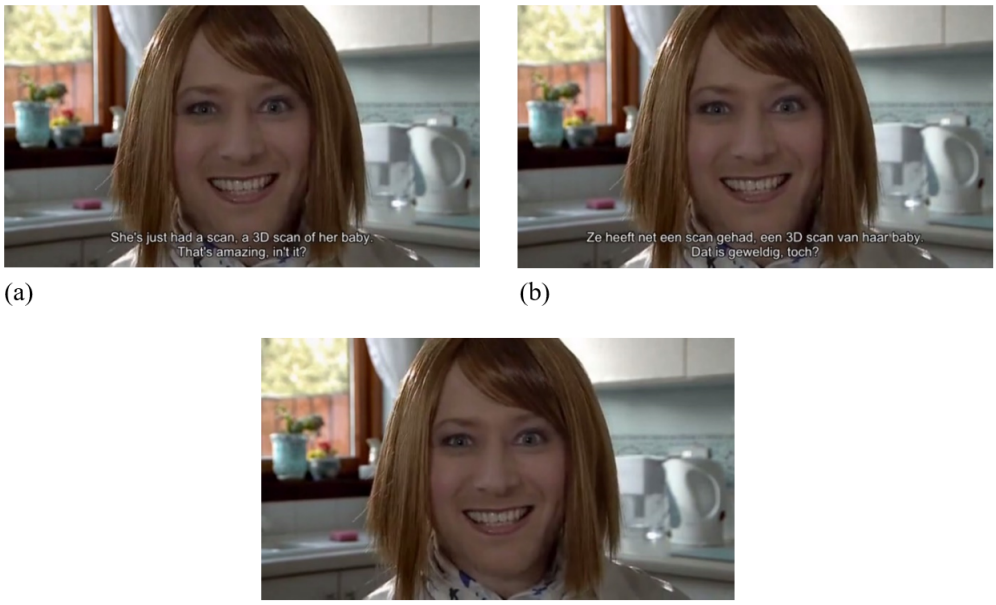
The sentences were divided over two different blocks, one used for the pre-test and the other used for the post-test. The blocks were matched in difficulty as much as possible in two different ways. Firstly, the length of each sentence in both words and syllables was determined. Each block contained roughly the same number of sentences of each length. Secondly, a subjective judgment about the difficulty of the sentences was made by a native speaker of Dutch with a high proficiency in English. Each block contained eight sentences produced by the main character, and eight sentences produced by the two other speakers. Within each block, the sentences were presented to each participant in a random order. A complete list of the sentences can be found in Appendix A.

2.2.2 English video sketches

All subjects watched a video with a duration of 14 minutes and 42 seconds. This video was a compilation of multiple excerpts from the first season of the television show Limmy's Show. The sketches were selected carefully as to not include any strong language, violence, inappropriate humour or alcohol use. Additionally, only sketches in which Limmy was the sole speaker were selected. In some sketches, one or two single phrases were spoken by another character, which were edited out.

The video was accompanied by either English, Dutch or no subtitles. The English subtitles were either adapted from the subtitles used by the BBC when the show was aired (for sketches from episode 1), or adapted from fan made subtitles if these official subtitles were not available (for all the other episodes). All subtitles were checked for mistakes or missing words and phrases, and edited where necessary. Some of the subtitles used a phonological representation of words that were pronounced with a very strong accent, which was not desirable for this particular experiment. These subtitles were therefore replaced with the orthographic representation of what was said. Additionally, repetitions and unfinished utterances that were left out of the original subtitles were added. All English subtitles were checked by another native speakers of Dutch with a high proficiency in English and a native speaker of English (the supervisors).

The Dutch subtitles did not exist yet and were translated by a native speaker of Dutch with a high proficiency in English (the author), using the English subtitles as a guideline. The sentences were translated as closely as possible to the English originals. However, in cases where literal transla-



(a) (b)

(c)

Figure 2: Still of the video in the (a) English subtitle condition (b) Dutch subtitle condition (c) No subtitle condition.

tions would have lead to incorrect or strange phrases in Dutch, more liberal translations were made. All Dutch subtitles were checked by another native speakers of Dutch with a high proficiency in English and a native speaker of English (the supervisors).

For technical reasons, it was not possible to display the video without a pause button. The pause button was therefore present, but participants were instructed to not use this button.

2.2.3 Attention check

An attention check was used to check if participants had actually paid attention to the video. A set of three comprehension questions was presented to the participants immediately after the listening post-test. Each question was about a different sketch and asked what the main point of each sketch was. The questions were all multiple choice questions with three response options. Participants that failed to answer at least two out of the three questions correctly were excluded from the analysis.

The complete list of the comprehension questions can be found in Appendix B.

2.2.4 Background questionnaire

All participants completed a background questionnaire to gain some additional information about all participant groups. Most questions were used to check whether the participants fulfilled the requirements of the study. Participants were also asked to provide information about their experience with the Scottish accent.

The complete background questionnaire can be found in Appendix C.

2.3 Design and procedure

2.3.1 Experiment design

The experiment was presented using the online survey platform Qualtrics. All tasks were included in a single survey so participants could go through all tasks themselves at their own pace. A different version of the survey was made for each different subtitle condition. By doing this, the conditions could be assigned in a pseudo-random manner by the experimenter.

Participants first received the first part of the dictation task (pre-test). Each sentence was presented on a different page. This page was displayed for a maximum of 60 seconds before continuing automatically. A countdown timer was visible for the participants. On this page, a button was presented to play the sentence. Participants could play the sentence twice and then the button would be disabled. Below the button was a text box in which the participants could type their transcription of the sentence. It was possible to type at any time before, during, or after listening to the sentence. The participant could go to the next sentence by pressing a button or waiting for the countdown timer to reach zero.

After completing the pre-test, the participants started the exposure phase by watching the video. The video was embedded in the survey. For technical reasons, it was not possible to embed the video without enabling the play/pause button. The participants therefore had to press the play button themselves and were not allowed to pause the video after doing so. To prevent participants from skipping the video and continuing the survey, the button to go to the next page was disabled for 14 minutes and 42 seconds, the duration of the video. The participants had to click this button to go to the next page.

The participants then had to continue with the second part of the dictation task (post-test). The post-test was presented in the same manner as the pre-test.

After the post-test, the participants had to complete the attention check. All questions in the attention check were multiple choice questions with three

possible responses. One of the three responses could be selected. All questions were presented on the same page. There was no time limit for completing the attention check.

Finally, the participants had to complete the background questionnaire. The background questionnaire consisted of a mix of multiple choice and open questions that could be completed by filling in a text box. Questions about the same theme were presented together on one page.

2.3.2 Procedure for high school students

High school students participated together with their whole class at the same time. The experiment was conducted in the school library of their own school. Other, non-participating students were present in the room but were instructed to be as quiet as possible. The experiment was presented on the school computers in the Mozilla Firefox browser. Participants brought their own headphones or used headphones provided by the school. All participants were assigned to an experiment condition systematically, to make sure that students sitting next to each other did not have the same type of subtitles.

Participants first received a general instruction about the experiment. The general instruction for the experiment was given orally to the whole group at the same time. In this instruction, they were told that they would be doing a number of different tasks. First, they would be doing a dictation task. In this task, they had to listen to a sentence and then write down what they heard. They were allowed to listen to each sentence twice. Participants were encouraged to always give a response, even if they could not understand the sentence. They were also instructed that they would have 60 seconds for each sentence. It was emphasized that participants could work at their own pace and that they did not have to hurry because of the timer. Participants then would have to watch a video. Participants were instructed that this video would be about 15 minutes long, and that they might have either English, Dutch or no subtitles with this video. It was explained to the participants that it was randomly assigned and that they all had to watch carefully, irrespective of the subtitle condition. Participants were instructed to start the video themselves but to never pause the video while it was playing. It was emphasized that participants had to pay attention throughout the entirety of the video. Participants would then be doing a dictation task again with new sentences. The second dictation task would work in exactly the same way as the first dictation task. Participants would then be completing a questionnaire with questions about the video and about their own (language) background. It was emphasized that they had to answer honestly. It was explained that there was no time limit for completing the questionnaire.

The participants were then instructed to please focus on the tasks and to not distract themselves or other participants while doing the tasks. After completing all of the tasks, they had to move to a different area of the library.

Before starting the experiment, the participants were instructed to listen to a test sound. They could use this sound to change the volume to a comfortable level, as they all used their own headphones. The sound could be played as many times as was needed to adjust the volume. After adjusting the volume, the participants could go through all of the tasks by themselves, as described in paragraph 2.3.1.

The high school students received a debriefing through e-mail after all the high school students had been tested, to prevent them from telling possible future participants what the experiment was about.

2.3.3 Procedure for university students

University students were tested either individually or in pairs but in separate booths. The experiment was conducted in a lab of the Centre for Language Studies at Radboud University Nijmegen. The experiment was presented on a lab computer in the Google Chrome browser.

The general instruction was given orally to the participant. If two participants were tested at the same time, they were instructed together. The instructions were the same as the instructions for the high school students, with the exception that the university students were told that they were not allowed to change the volume when listening to the test sound as they were all using the same equipment. The university students also received a debriefing immediately after completing the task. The rest of the procedure was the same as the procedure for high school students.

2.4 Measures

To score the sentences produced by the participants in the dictation task, all punctuation and additional whitespace were removed from the responses, so only words with a single space in between remained. For some words, multiple spellings were allowed. This was done to deal with specific contractions of words (e.g. *"He's"* and *"He is"*). All the different options for one word were therefore converted to one of the options using an automated dictionary.

After cleaning up the sentences, the sentences were automatically scored in two different ways. The first type of score was the lexical accuracy. The lexical accuracy showed how many words of the target phrase were present in the participant's transcription. To avoid reliance on human raters, words

were only counted if they were spelled correctly. The total was then divided by either the number of words in the response or the answer key, depending on which amount of words was higher. Example: if a participant gave the response *"Like they will see over there"* to the sentence that contained *"Like they would say over there"*, the total amount of correct words would be 4 out of 6. The score for the sentence would then amount to 0.67. The lexical accuracy was also used in previous studies, although in a less strict fashion. In these studies, spelling mistakes were either not possible because the task was a speaking task (in the case of Mitterer and McQueen (2009), and Charles and Trenkic (2015)) or light spelling mistakes were allowed (in the case of Birulés-Muntané and Soto-Faraco (2016)).

The second type of score was the orthographic closeness. This distance was calculated by calculating the Levenshtein distance between the correct response and the participant's transcription. The Levenshtein distance is the minimum number of single-character edits (insertions, deletions and substitutions) one would need to transform the participant's response to the correct response Levenshtein (1966). To normalize the Levenshtein distance, the total number of edits was divided by the total number of characters in either the response or the correct answer, depending on which was the longest. This normalization yielded a score between 0 and 1. In this study, the score was then inverted to make the scores and models easier to interpret (a higher score now represents a better performance). Example: if a participant gave the response *"Half a pint of beer please"* to the sentence that contained *"A half pound of bacon please"*, the normalized Levenshtein distance would be 0.39, which would then be inverted to a orthographic closeness score of 0.61.

The orthographic closeness measure was included because it is more precise than the lexical accuracy. With the lexical accuracy, getting part of the word wrong means that the answer is incorrect, and therefore yields a score of 0. When using the orthographic closeness instead, it is possible to award a higher score to partially correct answers than to completely incorrect answers.

The mean score for the items that were presented during the pre-test was calculated for each participant. This score was used as the predictor Pre-test Score. The scores for the items from the post-test were used as the dependent variable. If the Pre-test Score was measured as the orthographic closeness, the dependent score would be expressed as orthographic closeness as well. If the Pre-test Score was however measured as the lexical accuracy, the dependent score would be expressed as lexical accuracy as well. Two separate analyses were performed, one with each type of measure.

3 Results

3.1 Procedure for analyses

The results were analysed using linear mixed effects models in R (R Core Team, 2018) and the lme4 package (Bates, Mächler, Bolker, & Walker, 2015).

The mean Post-test Score was used as the dependent variable. In the first analysis (see 3.2.1 and 3.2.3), the Post-test Score was measured using the orthographic closeness. In the second analysis (see 3.2.2 and 3.2.4), the Post-test Score was measured using the lexical accuracy. The subtitle Condition, Speaker and Pre-test Score were added as fixed effects. Condition consisted of three levels: No Subtitles, English Subtitles and Dutch Subtitles. The No Subtitles condition was mapped onto the intercept. Speaker consisted of three levels: Speaker 1 (speaker who was featured in the video), Speaker 2 and Speaker 3. Speaker 1 was mapped onto the intercept. The Pre-test Score was a continuous variable ranging from 0 to 1, with a higher number indicating a better score at the pre-test. All possible two- and three-way interactions between Condition, Pre-test Score and Speaker were also added as fixed effects. The participants and the trials were included as random factors. By-subject and by-item random slopes were added to the models but were removed again as they did not improve the fit of the models. Visual inspection of the residual plots did not show any obvious deviations from homoscedasticity or normality. T-values higher than 1.96 were considered to be significant.

The optimal model was established using a backward-elimination procedure. In the first iteration of the model, all possible factors and two- and three-way interactions between these factors were included in the model. Predictors that were not significant were removed from the model one by one, starting with the highest orders of interactions, until only significant factors or factors that were involved in an interaction effect remained. To make sure that the model was the most optimal model, it was compared to models with additional factors or with fewer factors using ANOVAs.

3.2 Analyses

3.2.1 Analysis of students only using orthographic closeness

The optimal statistical model revealed a main effect for pre-test score ($\beta = 0.634$, SE = 0.170, t = 3.718), meaning that a higher pre-test score resulted in a higher post-test score. This can also be seen in Figure 3.

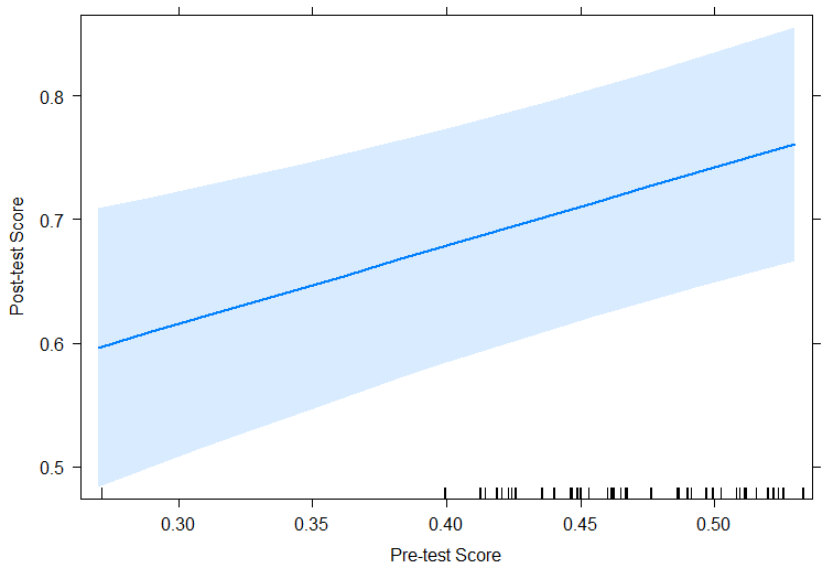


Figure 3: Main effect of the best fitting model for orthographic closeness using only student data.

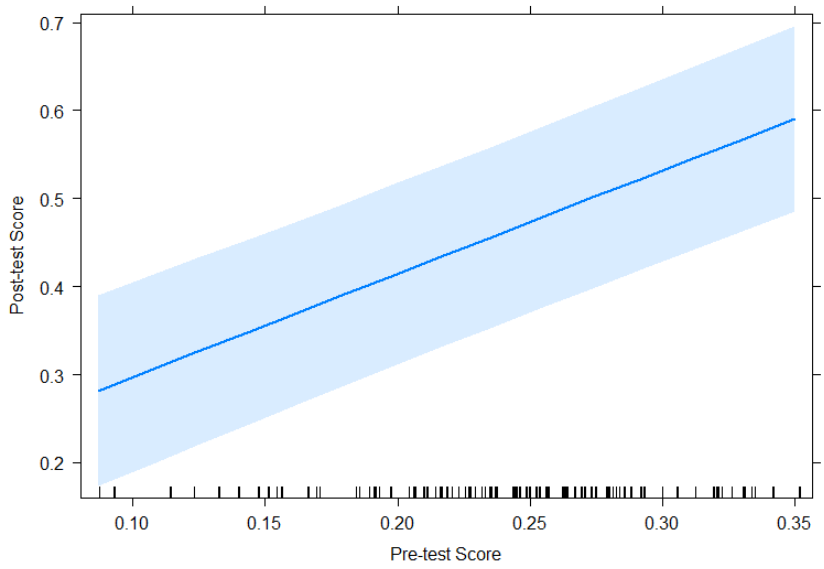


Figure 4: Main effect of the best fitting model for lexical accuracy using only student data.

3.2.2 Analysis of students only using lexical accuracy

The optimal statistical model revealed a main effect for pre-test score (β = 0.839, SE = 0.180, t = 4.673), meaning that a higher pre-test score resulted in a higher post-test score. This can also be seen in Figure 4.

3.2.3 Analysis of all participants using orthographic closeness

The optimal statistical model revealed a main effect for Pre-test Score (β = 0.813, SE = 0.253, t = 3.219). Additionally, there was a main effect of Dutch subtitles compared to baseline (β = -3.02, SE = 0.134, t = -2.257). This means that the scores in the Dutch subtitle condition were lower than in the No subtitle condition. There was no main effect of English subtitles compared to baseline (β = 0.020, SE = 0.134, t = 0.147). A two-way interaction was found between Condition and Pre-test score. This two-way interaction was significant in the Dutch subtitle condition compared to baseline (β = 0.659, SE = 0.303, t = 2.172). This was not the case in the English subtitle condition compared to baseline (β = -0.037, SE = 0.303, t = -0.123). The two-way interaction can be seen in Figure 5. This figure shows that scores were lower in the Dutch subtitle condition than in the baseline condition when the pre-test score of the participant was low, but that they were similar to the baseline condition when the pre-test score of the participant was higher.

Additional ANOVA analyses were done to investigate the interaction effect. First, the effect of condition (three levels: no subtitles, Dutch subtitles and English subtitles) was investigated for the pre- and post-test scores separately. A one-way ANOVA with pre-test score as the dependent variable and condition as the predictor was not significant (F(2,99) = 0.002, p = 0.998). A one-way ANOVA with post-test score as the dependent variable and condition as the predictor was also not significant (F(2,99) = 0.281, p = 0.756). Secondly, the effect of test (two levels: pre-test or post-test) was investigated

Table 2: Beta values, standard errors and t-values of the best fitting model for the orthographic closeness

Effect	Level	β	SE	t
Intercept	No subtitles	0.315	0.120	2.616
Condition	Dutch subtitles	-3.020	0.134	-2.257
Condition	English subtitles	0.020	0.134	0.147
Pre-test Score	-	0.813	0.253	3.219
Condition*Pre-test Score	Dutch subtitles	0.659	0.303	2.172
Condition*Pre-test Score	English subtitles	-0.037	0.303	-0.123

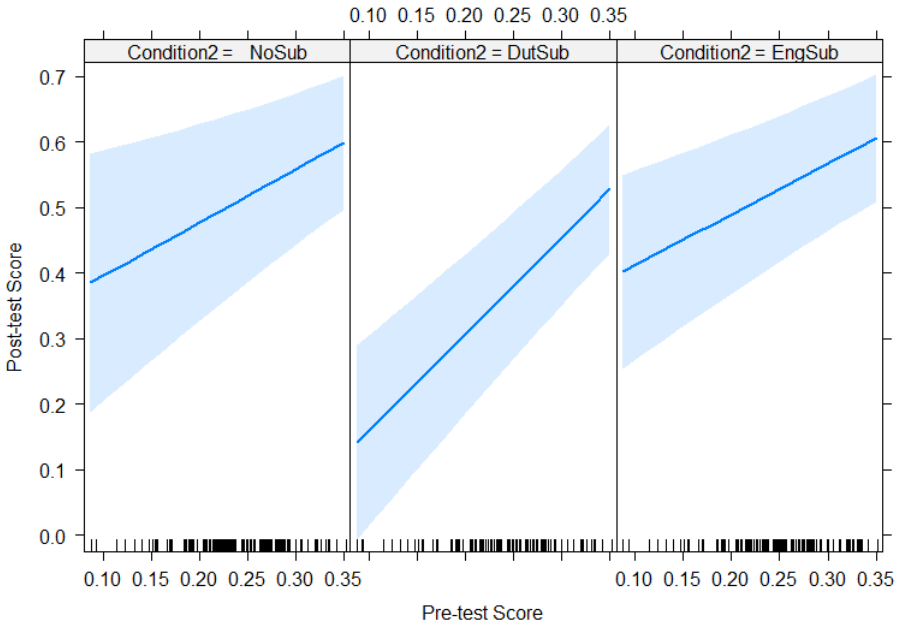


Figure 5: Two-way interactions of the best fitting model for the orthographic closeness.

for all three subtitle conditions separately. A one-way repeated measures ANOVA with all scores from the no subtitle condition as the dependent variable and Test as predictor was significant ($F(1,32) = 395.3$, $p < 0.001$). A one-way repeated measures ANOVA with all scores from the Dutch subtitle condition as the dependent variable and Test as predictor was also significant ($F(1,32) = 259.6$, $p < 0.001$). Finally, a one-way repeated measures ANOVA with all scores from the English subtitle condition as the dependent variable and Test as predictor was also significant ($F(1,35) = 522.7$, $p < 0.001$).

Another model in which the pre-test score was replaced by group (two levels: high school and university) was fit. However, the fit of this model (AIC = -783.19) was worse than the fit of the model using the pre-test score (AIC = -814.96). In this model, the was a main effect of Group ($\beta = 0.123$, SE = 0.019, t = 6.627). The university students scores higher in the post-test than the high school students. There was neither a main effect of Speaker 2 to baseline ($\beta = 0.171$, SE = 0.116, t = 1.470) or Speaker 3 to baseline ($\beta = 0.042$, SE = 0.116, t = 0.365). There was however a two-way interaction between Group and Speaker. This two-way interaction was significant for Speaker 2 compared to baseline ($\beta = -0.078$, SE = 0.022, t = -3.554) but not for Speaker 3 compared to baseline ($\beta = 0.042$, SE = 0.116, t = 0.365).

In this two-way interaction, the scores were higher for Speaker 2 than the baseline speaker in the high school group, but the scores for Speaker 2 were not higher than the baseline speaker in the university group.

Table 3: Beta values, standard errors and t-values of the alternative model for the orthographic closeness

Effect	Level	β	SE	t
Intercept	High school	0.570	0.068	8.395
Group	University	0.123	0.019	6.627
Speaker	Speaker2	0.171	0.116	1.470
Speaker	Speaker3	0.042	0.116	0.365
Group*Speaker	Speaker2	-0.078	0.022	-3.554
Group*Speaker	Speaker3	-0.024	0.022	-1.125

3.2.4 Analysis of all participants using lexical accuracy

The optimal statistical model revealed a main effect for pre-test score (β = 1.174, SE = 0.116, t = 10.120), meaning that a higher pre-test score resulted in a higher post-test score. This can also be seen in Figure 6.

Another model in which the pre-test score was replaced by group (two levels: high school and university) was fit. However, the fit of this model (AIC = -503.16) was worse than the fit of the model using the pre-test score (AIC = -532.35). In this model, there was a main effect of Group (β = 0.118, SE = 0.019, t = 6.377). The university students scores higher in the post-test than the high school students. There was a main effect of Speaker 2 compared to baseline (β = 0.267, SE = 0.121, t = 2.201). The scores for sentences spoken by Speaker 2 were higher than those for the baseline speaker. The was no main effect of Speaker 3 compared to baseline (β =

Table 4: Beta values, standard errors and t-values of the alternative model for the lexical accuracy

Effect	Level	β	SE	t
Intercept	High school	0.326	0.071	4.622
Group	University	0.118	0.019	6.377
Speaker	Speaker2	0.267	0.121	2.201
Speaker	Speaker3	0.112	0.121	0.926
Group*Speaker	Speaker2	-0.080	0.024	-3.349
Group*Speaker	Speaker3	0.001	0.024	0.056

0.112, SE = 0.121, t = 0.926). A two-way interaction between Speaker and Group was also found. This two-way interaction was significant for Speaker 2 compared to baseline (β = -0.080, SE = 0.024, t = -3.349) but not for Speaker 3 compared to baseline (β = 0.001, SE = 0.024, t = 0.056). In this two-way interaction, the scores were higher for Speaker 2 than the baseline speaker in the high school group, but the scores for Speaker 2 were not higher than the baseline speaker in the university group.

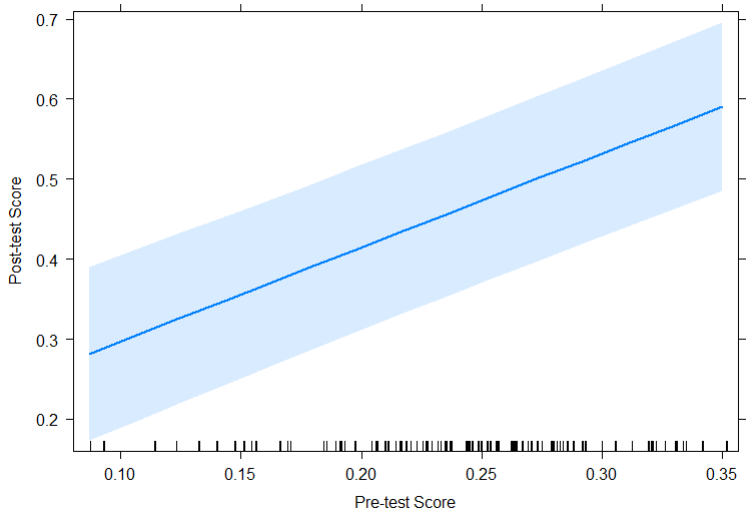


Figure 6: Main effect of the best fitting model for the lexical accuracy.

4 Discussion

4.1 Discussion of subquestions

The four subquestions will first be addressed separately. Afterwards, the main research question will be discussed.

4.1.1 The influence of subtitle language on university students

The first subquestion was: *"What is the influence of the subtitle language on perceptual tuning in English of Dutch university students?"*. The aim of this subquestion was to replicate previous studies (Birulés-Muntané & Soto-Faraco, 2016; Charles & Trenkic, 2015; Mitterer & McQueen, 2009) and to find out whether the subtitle languge modulated the score at the post-test using, as this study uses different materials and different ways to measure the scores. The hypothesis was that participants who had viewed the video with English subtitles would score better at the post-test than participants who did not get any subtitles, and that participants who had viewed the video with Dutch subtitles would score worse at the post-test than participants who did not get any subtitles.

In both the analyses using lexical accuracy and orthographic closeness, there was no main effect of subtitle condition. There was only a main effect of pre-test score. The results are therefore different from the hypothesis and also from previous studies (Birulés-Muntané & Soto-Faraco, 2016; Charles & Trenkic, 2015; Mitterer & McQueen, 2009). It was expected that participants in the Dutch subtitle condition would perform worse at the post-test than the baseline condition (no subtitles), and that participants in the English subtitle condition would perform better than the baseline. Why this result was not found in this study is unclear. It is possible is that either the material or the measures used were not adequate to distinguish between the subtitle groups. Another possibility is that the amount of trials per participant, or the amount of participant in itself, was simply too low. When looking at the full dataset, some effects of subtitle condition were found, as will be explained in the next section.

4.1.2 The influence of subtitle language on high school and university students

The second subquestion was: *"What is the influence of the subtitle language on perceptual tuning in English of Dutch high school and university students?"*. The aim of this subquestion was to find out whether the subtitle language modulated the score at the post-test. The hypothesis was that

participants who had viewed the video with English subtitles would score better at the post-test than participants who did not get any subtitles, and that participants who had viewed the video with Dutch subtitles would score worse at the post-test than participants who did not get any subtitles.

In the analyses using the lexical accuracy, there was no main effect of subtitle condition. The scores in all three subtitle conditions were similar to each other. However, in the analysis using the orthographic closeness to determine the scores, there were some differences. The orthographic closeness measure is more precise than the lexical accuracy measure, as it also measures partially correct answers. Therefore, I will be focusing on these results. There was a main effect of Dutch subtitles compared to the baseline condition (no subtitles). Participants who had seen the video with Dutch subtitles, scored significantly lower on the post-test than participants who had seen the video without any subtitles. There was no difference between participants who had seen the video with English subtitles and the baseline condition. It can therefore be concluded that Dutch subtitles decrease the amount of lexically-guided perceptual tuning that takes place, and that English subtitles do not improve this perceptual tuning. However, note that the effect was not very robust: see section 4.1.3 about the interaction effect with language proficiency.

The results differ slightly from what was hypothesized. The lower scores in the Dutch subtitles was predicted. The same effect was also found in other studies. In the studies by Mitterer and McQueen (2009) and Birulés-Muntané and Soto-Faraco (2016), native-language subtitles lead to lower scores in the post-test compared to the baseline of no subtitles. The current study therefore attests to the theory devised by Mitterer and McQueen (2009). According to their theory, native-language subtitles harm lexically-guided perceptual tuning because the subtitles give contradictory phonological information. Because the written text is automatically processed, the phonological information of the native language is activated. This information is a mismatch with the information from the spoken signal. Therefore, it seems to be more difficult to update the phonetic categories.

It was not predicted that the scores in the English subtitle condition would be the same as the scores in the baseline condition. In other studies (Birulés-Muntané & Soto-Faraco, 2016; Charles & Trenkic, 2015; Mitterer & McQueen, 2009), using same-language subtitles lead to better scores than using no subtitles or using native-language subtitles. It is unclear why this result was not found in the current study. The dictation task might measure a slightly different domain of speech comprehension, and therefore lead to different results than found by the shadowing task (Charles & Trenkic, 2015; Mitterer & McQueen, 2009) and the gap-task (Birulés-Muntané & Soto-

Faraco, 2016). Alternatively, the dictation task might have been too hard or the pre- and post-test might have differed in level. It is also possible that the number of items for the dictation task was too low (an explanation of the limitations of this study can be found in section 4.3). Another explanation would be that same-language subtitles do not actually lead to improved perceptual tuning. However, as multiple studies have confirmed this finding, this would be rather unlikely. A fifth explanation might be that the exposure phase in this study was too short compared to previous studies. The previous studies all used an exposure phase of at least 30 minutes. This study had an exposure phase of only 15 minutes. Although perceptual learning can already occur after very brief exposure (Clarke & Garrett, 2004; Sidaras et al., 2009), it is possible that the exposure phase was still too short to lead to significant results. A final explanation might be that the other studies all used multi-speaker exposure, while this study used single-speaker exposure. Viewing only a single speaker, possibly in combination with a brief exposure, might not lead to enough perceptual tuning.

The results found for this subquestion show that watching a foreign-language video with Dutch subtitles might be harmful for speech comprehension in that foreign language. However, the subtitle language is not the only important factor. An interaction effect between the subtitle language and English proficiency was also found in this study. This result will be discussed in the next section.

4.1.3 The influence of language proficiency

The third subquestion was: *"Does language proficiency modulate this possible influence?"*. The aim of this subquestion was to investigate whether the English language proficiency of the participants would have an influence on the effect of the subtitle language, if such an effect was found at all. The hypothesis was that there would be a different influence of proficiency on the different subtitle conditions. It was expected that participants with a lower proficiency level would not improve as much in the English subtitle condition as participants with a higher proficiency level. This was expected because I assumed that there would be some kind of threshold proficiency level that would be necessary to fully make use of the same-language subtitles. The other expectation was that participants with a lower proficiency level would improve less in the Dutch subtitle condition than participants with a higher proficiency level. This was hypothesized because I assumed that participants with a lower proficiency level would rely on the subtitles more, than participants with a higher proficiency level and therefore be more focused on the contradicting information. Note that the English proficiency of the partici-

pants was not measured directly in this study: the analyses used either the score on the pre-test or the education level (high school or university) as an indirect measure of proficiency.

In the analyses using the lexical accuracy to determine the scores, there was no interaction effect between the pre-test score or group and the score on the post-test. The post-test score and therefore the assumed proficiency level do not seem to modulate the effect of subtitle condition. However, in the analyses using the orthographic closeness, a two-way interaction between pre-test score and the score on the post-test was present. In the model used in this analysis, the post-test score and therefore the assumed proficiency level do modulate the scores on the post-test. This two-way interaction was found for the Dutch subtitle condition as compared to the baseline condition (no subtitles), but not for the English subtitle condition as compared to the baseline condition. In Figure 5 (section 3.2.3), it can be seen that participants with a high proficiency level had similar scores as participants in the other two subtitle conditions. However, participants with a lower proficiency level scored significantly lower than participants with a similar proficiency level in the other two conditions.

However, when investigating this interaction effect between pre-test score and subtitle condition more closely, it does not become clearer what it means exactly. When looking at the pre- and post-test separately, there were no differences between the different subtitle conditions for either test. When looking at the difference between pre- and post-test of the different subtitle conditions separately, there were also no differences between the differenct conditions: in all of them, the post-test score was significantly higher than in the pre-test. Creating a model with only condition or only the interaction between condition and pre-test score as a condition also shows that these predictors on their own are not significant. These models however do not fit the data as well as the model that includes both the condition and the interaction between condition and pre-test score.

These results show that using same-language subtitles might not have the same effect for all listeners, but it is not really clear how this possible proficiency effect manifests itself. From the linear mixed-effects model and Figure 5, it seems like participants with a low English proficiency have more trouble with the Dutch subtitle condition than other participants with a low English proficiency have in the other subtitle conditions, but a closer inspection of the differences between the conditions does not confirm the results.

The hypothesis was therefore only partially correct. While the results give some indication that the perceptual tuning of listeners with a low proficiency level might be obstructed more in the native subtitle condition, they are

not completely conclusive. The results also give no indication that listeners with a high proficiency level can benefit more from the English subtitles than listeners with a lower proficiency. There are a few possible explanations for this. Firstly, it is possible that the proposed "proficiency threshold" for using subtitles for perceptual tuning does not exist. Listeners of all proficiency levels are then able to use subtitles in the same way. It is also possible that this threshold does exist, but that the proficiency level of the participants in this study was already high enough to be above the threshold. This could be confirmed by repeating the study with participants with a lower proficiency level. It is also possible that the lack of an interaction effect is due to the fact that the difference between the subtitle conditions was very small. See section 4.4 for more limitations of this study that can give a possible explanation.

4.1.4 Generalization to new speakers

The fourth and final subquestion was: *"If there is a learning effect, will this effect generalize to speakers who were not heard during the exposure phase?"*. The aim of this subquestion was to find out whether it is possible to tune into a language from listening to a single speaker, or whether it is necessary to watch multi-speaker videos. The hypothesis was that participants would score lower for sentences spoken by the new speakers than for sentences spoken by the speaker that was also speaking during the exposure phase. This result was expected because other studies (e.g. Bradlow and Bent, 2008) found that exposure to a single speaker only led to tuning into this specific speaker, while exposure to multiple speakers improved perception of an accent in general.

In the two analyses using the orthographic closeness to determine the scores, there was no main effect of speaker. There were also no interaction effects involving speaker. However, in the analyses using the lexical accuracy to determine the scores, there was a main effect of speaker when using group as an indirect measure of proficiency. There was also a two-way interaction between speaker and pre-test score in the first analysis and a two-way interaction between speaker and education level in the second analysis, in both the analysis using education level and the analysis using the pre-test score.

Looking at the results in detail, however, they seem to be inconclusive. According to the model using the education level as an indirect measure of proficiency, one of the new speakers (speaker 2) seems to be easier to understand than the speaker from the video (speaker 1). The other speaker (speaker 3) is however not easier to understand. Moreover, the two-way interaction shows that the main speaker and speaker 3 are more difficult to understand for high school students, but not for university students. The

model that uses the pre-test score as an indirect measure of proficiency did not have a main effect of speaker but just the two-way interaction, again showing that the main speaker and speaker 3 are harder to understand for high school students.

It seems very likely that the effects that were found for speaker were due to noise or problems in the construction of the dictation task. The models in which an effect of speaker was found, did not fit the data as well as the models in which speaker was not a significant factor. It is possible that the sentences that the sentences spoken by the different speakers were not of equal difficulty levels. Alternatively, some speakers might have been easier to understand than others, for example due to clearer articulation or a slower speech rate. It is also possible that there were differences in difficulty level between the pre- and post-test, and that these differences were not equal for the different speakers.

4.2 General discussion

The main research question was: *"What is the influence of watching subtitled videos on speech perception in a foreign language, and does language proficiency modulate this influence?"*. The aim of this research question was to find out whether the subtitle language influences how well listeners can tune into a foreign language when watching a video, and to investigate whether proficiency might modulate this effect.

The results of the second subquestion have shown that watching a video with English subtitles or without any subtitles leads to similar results. Previous research has found that watching a video with English subtitles lead to better speech comprehension than watching a video without subtitles (Birulés-Muntané & Soto-Faraco, 2016; Charles & Trenkic, 2015; Mitterer & McQueen, 2009). It is unclear why these results were not found in this study. However, the results also show that watching a video with Dutch subtitles leads to worse results than watching a video without any subtitles. This is in accordance with the results from previous studies (Birulés-Muntané & Soto-Faraco, 2016; Mitterer & McQueen, 2009). Mitterer and McQueen (2009) theorize that native-language subtitles might decrease the amount of perceptual tuning that takes place, because these subtitles give contradictory phonological information. It is possible that because the subtitles are read automatically, the phonological information that is relevant to Dutch is activated. However, this information contradicts the information from the English speech signal. This might make it more difficult to update the phonetic categories. To conclude, it might be better to watch foreign-language videos with no subtitles or with same-language subtitles instead of native-

language subtitles, if a listener wants to learn to understand this foreign language better.

Furthermore, the results of the third subquestion have shown that the effect of subtitle condition might not be similar for all listeners. Listeners with a lower proficiency level seem to be impacted more by the Dutch subtitles, while listeners with a higher proficiency level are not. It seems to be the case that listeners with a lower proficiency level either rely on the subtitles more, or have more difficulty suppressing the phonological information from their native language because of these subtitles. However, these results were difficult to interpret and did not seem to hold up completely under closer inspection. To conclude, a listener has to take their foreign language proficiency into account when selecting the subtitle language. Low proficiency listeners might do well to opt for same-language subtitles or no subtitles, while the subtitle language does not matter as much for high proficiency listeners.

Finally, the results of the final subquestion seem to be unclear. From the data in this study, it cannot be determined whether exposure to a single speaker can generalize to better understanding of new speakers.

4.3 Limitations

While this study aimed to improve some of the shortcomings that were found in previous studies, this study was also not without its limitations.

Firstly, the materials were not pre-tested before being used in this study. While this was probably not a big issue for the video used during the exposure phase, it may have led to unwanted differences in the pre- and post-test. The sentences that were used in the dictation task, were selected by a native speaker of Dutch that is proficient in English (the author). The selection was however done in a holistic and subjective manner. While this speaker may have felt that the sentences of the pre- and post-test were similar in difficulty, it is possible that one of either tests was harder than the other. The possible difficulty difference between pre- and post-test could have been combatted by using a counterbalanced design (i.e., switching which items were given in the pre-test and which in the post-test for half of the participants). The same "difficulty problem" could also be the case for the selection of sentences for the three different speakers. It is possible that not all speakers were comparably hard to understand. It is possible that one or two speakers were easier to understand than the other speaker(s).

Secondly, the high school students and university students were tested in vastly different settings. This may have influenced their scores on the pre- and post-test, as well as their ability to focus on the video. The university

students did the task individually in a sound-proof booth wearing noise-cancelling headphones with good audio quality. The high school students, however, were tested in groups of 15 to 30 students at a time in the school library. Even though the participants were mostly quiet throughout the experiment, it is possible that this setting may have been distracting. Other students that did not participate in the study were also present in the school library while the experiment took place, which may have caused distraction. Additionally, various other sounds could be heard in the school library, like the school bell and a teacher in a classroom next-door. Moreover, the high school students all brought their own headphones and could adjust the volume themselves. This may have caused differences in the audio quality and therefore the intelligibility of the speech in the video and the dictation task.

Another possible issue is that the experiment had to be changed in multiple ways to accommodate for testing the high school students. Two fairly important deviations had to be made as compared to previous studies. The first deviation regards the duration of the experiment. The total duration of the experiment including instructions had to fit within the duration of one class. Lessons at the high school were typically 50 minutes, while previous studies used tasks of 60 to 90 minutes. Those studies did also not include a pre-test, which was expected to take up some additional time. The duration of the experiment was reduced by limiting the duration of the video and therefore the exposure time, as well as presenting fewer trials in the post-test compared to previous studies. This means that the amount of trials per participant was also relatively small. The second deviation regards the means of responding to the post-test. Two previous studies used a spoken sentence repetition task (Charles & Trenkic, 2015; Mitterer & McQueen, 2009) while one study used a written one-word gap listening task (Birulés-Muntané & Soto-Faraco, 2016). Testing all high school students individually would have been too time-consuming for the scope of this study. However, testing students in groups would not have been possible if a spoken post-test was used. To make the post-test usable in groups while still keeping the level of detail that is measured by a shadowing task, this study used a written dictation task. While this task is very similar to a shadowing task, it is possible that a slightly different domain of speech comprehension was measured by changing to this task. Differences in the results could therefore be caused by differences in the task that was used.

Finally, some other methodological differences make it difficult to compare the results of this study and the results of previous studies directly. Firstly, this study was the first to use a single-speaker exposure instead of a multi-speaker exposure. It is therefore hard to compare the results of this study to previous studies, as differences in the results may have been caused by the

change from a multi-speaker to single-speaker exposure phase. Secondly, the outcome measures were calculated differently. In this study, both the word error rate and the orthographic closeness were used as outcome measures. The word error rate would be the most similar measure to the measures used by Mitterer and McQueen (2009), Charles and Trenkic (2015), who all calculated scores by taking the sum of the correctly produced words. However, as both tests used a shadowing task, the scores were less strict than the word error rate in the current study. In the study by Birulés-Muntané and Soto-Faraco (2016), a written post-test was used, but they allowed for small spelling mistakes. The word error rate in this study might have been influenced by spelling mistakes or phonological transcriptions. Scoring was done automatically and therefore these mistakes could not be taken into account. However, automatic scoring might have made the word error rate more objective than in previous studies, as they all relied on subjective judgments of raters. The orthographic closeness was not used in any of the previous studies and therefore it is difficult to compare the outcome using this score to previous studies. Lastly, the subtitles in this study might have been of a different quality than the subtitles in previous studies. Previous studies all used the official subtitles that belonged with the videos used in those studies, and translated the subtitles to the native language for the native-subtitle condition if there were no official subtitles for that language. However, such subtitles were not available for the video that was used in the current study, and user-generated subtitles had to be used instead. While the subtitles were thoroughly checked and edited, they may have been different from the usual style and quality of subtitles used on television.

4.4 Ideas for future research

Research on the topic of subtitles and speech comprehension has been very scarce. Therefore, we would already benefit from any additional research into this topic.

This study has been the first to investigate whether language proficiency modulates the effect that subtitled videos have on perceptual learning. While language proficiency was only measured indirectly by using a pre-test, it shows that differences in proficiency can also lead to different effects of subtitled videos on perceptual learning.

Some additional changes could be made in future research. Firstly, now that we know that proficiency influences the effect of subtitled videos on perceptual learning in a foreign language, it could be an idea to expand the range of proficiency levels. In the current study, participants with estimated proficiency levels ranging from B1 to C2 took part. It would be interest-

ing to see what effects subtitled videos can have on perceptual learning in beginning language learners, with an A1 to A2 level. These learners can potentially benefit tremendously from having an additional learning source. Especially for learners that learn a foreign language not spoken in their direct environment, being able to use subtitled videos could provide them with large amounts of spoken input.

Secondly, this study was the first study about the benefits of subtitles to use a single-speaker exposure phase as opposed to a multi-speaker exposure phase. Unfortunately, the analysis yielded unclear results regarding the generalization of the learning effect to new speakers. While the effects of using single-speaker versus multi-speaker exposure on perceptual learning have been investigated in the domain outside of subtitles, it would be interesting to see whether these findings for more general perceptual learning can also be generalized to perceptual learning with the help of subtitled videos.

Furthermore, all studies investigating a similar research question used an English video. No studies have been done using a different target language. Therefore, we only know the effects of watching subtitled videos in English, but it is possible that the influence of watching subtitled videos can be different when learning another language than English. To find out whether watching subtitled videos can be beneficial when learning other language, it is necessary to include other target languages in future studies. It is possible to take an additional step and include a target language that has a different alphabet or different writing system, or a language that is read from right to left as opposed to left to right, and see whether the effects that have been found for English materials still hold up in these situations.

Moreover, only a handful of different native languages have been compared. Both in the current study and in Mitterer and McQueen (2009), the native language of participants was Dutch. In Birulés-Muntané and Soto-Faraco (2016), the native language of participants was Spanish. The participants tested by Charles and Trenkic (2015) had Chinese as a native language, but this study did not include a native-subtitle condition. It would be interesting to expand the scope of future studies to include a variety of native languages, or even compare groups of participants with different native languages directly in one study. By doing this, it is possible to investigate whether native speakers of different languages can all benefit from watching subtitled videos or whether this is limited to native speakers of certain languages.

It would also be interesting to look at using a combination of two subtitle languages: both the native and the target language. Some language learners have developed tools that allow them to watch television series online using two sets of subtitles (). Whether these tools actually work has not been

investigated yet.

Finally, it might be interesting to look at the effects of combining dubbing and subbing for a single exposure. With the rise of online video-streaming services, language learners have discovered the possibility of watching their favourite shows dubbed in their target language instead of their native language to create video materials that they already know the contents of in their target language (Amarenselise, n.d.). It is possible to add target language subtitles as well. By using a video that is both dubbed and subbed, it is possible to create new materials that have a matching spoken and written signal. Some learners seem to think that they can benefit from the target-language subtitles in this case as well (Amarenselise, n.d.). However, the video does not completely match the sounds that are being spoken, as the original audio has been replaced. It would be interesting to find out whether these videos that combine dubbing and subbing have similar effects on perceptual learning as videos that only use subtitles.

4.5 Conclusion

This study investigated the influence of watching subtitled videos on perceptual learning in a foreign language, and whether language proficiency related to this influence. Four subquestions were investigated in this study: 1. What is the influence of the subtitle language on perceptual tuning in English of Dutch university students? 2. What is the influence of the subtitle language on perceptual tuning in English of Dutch high school and university students? 3. Does language proficiency modulate this possible influence? 4. If there is a learning effect, will this effect generalize to speakers who were not heard during the exposure phase?

These four questions were investigated by having Dutch high school and university students watch an English video featuring a speaker with a Glaswegian accent. This video was accompanied by either English, Dutch or no subtitles. The participants' perception of English was measured using a dictation task before and after watching the video.

The results showed that using English subtitles and no subtitles leads to similar scores of speech perception. They also showed that while using Dutch subtitles still leads to better scores than before watching the video, the scores are lower than when watching the video without any subtitles. This seems to be the case mostly for listeners with a lower proficiency level. Listeners with a low proficiency level had lower scores than their peers in the other subtitle conditions when using Dutch subtitles, while listeners with a high proficiency level scores similarly to listeners in the English and no subtitle conditions when using Dutch subtitles. The results were however

not completely straightforward and therefore difficult to interpret.

It can therefore be concluded that listeners should take the subtitle language into account when a they are watching foreign language videos. It seems to be the case that native subtitles harm perceptual learning, mostly for listeners with a lower proficiency level. Follow-up studies might be able to get clearer results and investigate the effects of subtitle language and proficiency further.

References

Alamri, S. F. (2016). *The effects of English, Arabic, and no subtitles on Arabic ESL learners' listening comprehension* (Unpublished doctoral dissertation). California State University, Fresno.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439.

Amarenselise. (n.d.). *How to watch Netflix in another language.* Retrieved 12/04/2019, from `http://foreignlanguagecollective.com/how-to-watch-netflix-in-another-language/`

Baese-Berk, M. (2018). Perceptual learning for native and non-native speech. In K. D. Federmeier & D. G. Watson (Eds.), *Current topics in language* (Vol. 68, p. 1-29). Cambridge, MA: Academic Press.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*, EL174–EL180.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Monro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (p. 13-34). Amsterdam, The Netherlands: John Benjamins.

Bird, S. A., & Williams, J. N. (2002). The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling. *Applied Psycholinguistics*, *23*, 509–533.

Birulés-Muntané, J., & Soto-Faraco, S. (2016). Watching subtitled films can help learning foreign languages. *PloS One*, *11*, e0158409.

Bisson, M.-J., Van Heuven, W. J., Conklin, K., & Tunney, R. J. (2013). Incidental acquisition of foreign language vocabulary through brief multimodal exposure. *PLoS One*, *8*, e60912.

Bisson, M.-J., Van Heuven, W. J., Conklin, K., & Tunney, R. J. (2015). The role of verbal and pictorial information in multimodal incidental acquisition of foreign language vocabulary. *The Quarterly Journal of Experimental Psychology*, *68*, 1306–1326.

Blumstein, S. E., & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, *10*, 25-32.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native

speech. *Cognition*, *106*, 707-729.

Brasel, S. A., & Gips, J. (2014). Enhancing television advertising: Same-language subtitles can improve brand recall, verbal memory, and behavioral intent. *Journal of the Academy of Marketing Science*, *42*, 322–336.

Burnham, D., Leigh, G., Noble, W., Jones, C., Tyler, M., Grebennikov, L., & Varley, A. (2008). Parameters in television captioning for deaf and hard-of-hearing adults: Effects of caption rate versus text reduction on comprehension. *Journal of Deaf Studies and Deaf Education*, *13*, 391–404.

Charles, T., & Trenkic, D. (2015). Speech segmentation in a second language: The role of bi-modal input. In Y. Gambier, A. Caimi, & C. Mariotti (Eds.), *Subtitles and language learning: Principles, strategies and practical experiences* (p. 173-197). Bern, Switzerland: Peter Lang.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*, 149–210.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*, 3647–3658.

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge, United Kingdom: Press Syndicate of the University of Cambridge.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Cambridge, MA: MIT.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*, 317–367.

Davey, R., Parkhill, F., et al. (2012). Raising adolescent reading achievement: The use of sub-titled popular movies and high interest literacy activities. *English in Aotearoa*, *78*, 61-71.

De Linde, Z., & Kay, N. (1999). Processing subtitles and film images: Hearing vs deaf viewers. *The Translator*, *5*, 45–60.

d'Ydewalle, G., Praet, C., Verfaillie, K., & Rensbergen, J. V. (1991). Watching subtitled television: Automatic reading behavior. *Communication Research*, *18*, 650–666.

d'Ydewalle, G., & Van de Poel, M. (1999). Incidental foreign-language acquisition by children watching subtitled television programs. *Journal of Psycholinguistic Research*, *28*, 227–244.

d'Ydewalle, G., Van Rensbergen, J., & Pollet, J. (1987). Reading a message when the same message is available auditorily in another language: The case of subtitling. In J. O'Regan & A. Levy-Schoen (Eds.), *Eye*

movements from physiology to cognition (p. 313-321). Amsterdam, The Netherlands: Elsevier.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*, 1950-1953.

Europees Referentiekader Talen. (n.d.). *Niveaus havo/vwo.* Retrieved 25/03/2019, from `http://www.erk.nl/docent/streefniveaus/havo/`

Flege, J. E., MacKay, I. R., & Piske, T. (2002). Assessing bilingual dominance. *Applied Psycholinguistics*, *23*, 567–598.

Frauenfelder, U. H., & Tyler, L. K. (1987). The process of spoken word recognition: An introduction. *Cognition*, *25*, 1–20.

Ghorbani, M. R. (2011). Watching cartoons with subtitles improves children's foreign language acquisition. *US-China Foreign Language*, *9*, 241–246.

Griffin, R., & Dumestre, J. (1993). An initial evaluation of the use of captioned television to improve the vocabulary and reading comprehension of navy sailors. *Journal of Educational Technology Systems*, *21*, 193–206.

Hayati, A., & Mohmedi, F. (2011). The effect of films with and without subtitles on listening comprehension of EFL learners. *British Journal of Educational Technology*, *42*, 181–192.

Huang, H.-C., & Eskey, D. E. (1999). The effects of closed-captioned television on the listening comprehension of intermediate English as a second language (ESL) students. *Journal of Educational Technology Systems*, *28*, 75–96.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*, 460–482.

IELTS. (n.d.). *How should the CEFR be used by recognising institutions wishing to set language ability requirements?* Retrieved 25/03/2019, from `https://www.ielts.org/ielts-for-organisations/common-european-framework`

Kilborn, R. (1993). Speak my language: current attitudes to television subtitling and dubbing. *Media, Culture & Society*, *15*, 641–660.

Koolstra, C. M., & Beentjes, J. W. (1999). Children's vocabulary acquisition in a foreign language through watching subtitled television programs at home. *Educational Technology Research and Development*, *47*(1), 51–60.

Koolstra, C. M., Peeters, A. L., & Spinhof, H. (2002). The pros and cons of dubbing and subtitling. *European Journal of Communication*, *17*,

325–354.

Koolstra, C. M., van der Voort, T. H., & d'Ydewalle, G. (1999). Lengthening the presentation time of subtitles on television: Effects on children's reading time and recognition. *Communications*, *24*, 407–422.

Kothari, B., Pandey, A., & Chudgar, A. R. (2004). Reading out of the "idiot box": Same-language subtitling on television in India. *Information Technologies & International Development*, *2*(1), 23-44.

Kothari, B., Takeda, J., Joshi, A., & Pandey, A. (2002). Same language subtitling: a butterfly for literacy? *International Journal of Lifelong Education*, *21*(1), 55–66.

Kreisa, M. (n.d.). *The ultimate guide to using subtitles for foreign language learning.* Retrieved 25/03/2019, from `https://www.fluentu.com/blog/foreign-language-subtitles/`

Kuppens, A. H. (2010). Incidental foreign language acquisition from media exposure. *Learning, Media and Technology*, *35*, 65–85.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklandy*, *10*, 707–710.

Linebarger, D., Piotrowski, J. T., & Greenwood, C. R. (2010). On-screen print: the role of captions as a supplemental literacy tool. *Journal of Research in Reading*, *33*, 148–167.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*, 465–494.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113-1126.

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS One*, *4*, e7785.

Mousavi, F., & Gholami, J. (2014). Effects of watching flash stories with or without subtitle and reading subtitles on incidental vocabulary acquisition. *Procedia-Social and Behavioral Sciences*, *98*, 1273–1281.

Myers, E. (n.d.). *Learn a new language by turning on subtitles.* Retrieved 25/03/2019, from `https://www.rev.com/blog/learn-new-language-subtitles`

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1209-1228.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299-325.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204-238.

Pedersen, J. (2017). How metaphors are rendered in subtitles. *Target: International Journal of Translation Studies*, *29*, 416–439.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*, 1207–1218.

Sidaras, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, *125*, 3306-3316.

Steinfeld, A. (1998). The benefit of real-time captioning in a mainstream classroom as measured by working memory. *Volta Review*, *100*, 29–44.

Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching*, *43*, 1–37.

Vihtnan, M. M., DePaolis, R. A., & Davis, B. L. (1998). Is there a "trochaic bias" in early word learning? Evidence from infant production in English and French. *Child Development*, *69*, 935–949.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*, 1–25.

Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, & Psychophysics*, *75*(3), 537–556.

Yang, J. C., & Chang, P. (2014). Captions and reduced forms instruction: The impact on EFL students' listening comprehension. *ReCALL*, *26*, 44–61.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1-14.

Yoshino, S., Kano, N., & Akahori, K. (2000). The effects of English and Japanese captions on the listening comprehension of Japanese EFL students. *Language Laboratory*, *37*, 111–130.

Appendix A: list of sentences

Test	Item	Speaker	Phrase
Pre-test	1	Main	She should be here any minute
Pre-test	2	Main	She was an old man
Pre-test	3	Main	Like they would say over there
Pre-test	4	Main	Another arrow on the ground
Pre-test	5	Main	Get it out the way
Pre-test	6	Main	But it is a difficult question
Pre-test	7	Main	Just for your sense of humour
Pre-test	8	Main	He gave me my big break
Pre-test	9	New 1	Just wait till you hear it
Pre-test	10	New 1	What time's she getting here?
Pre-test	11	New 1	You're always there to listen
Pre-test	12	New 1	I can tell you anything
Pre-test	13	New 2	I mean she deserves a medal
Pre-test	14	New 2	The letters are all laid out
Pre-test	15	New 2	Can you start on Monday?
Pre-test	16	New 2	After the state you were in
Post-test	17	Main	Scratch it in with a compass
Post-test	18	Main	Wrong shape for her eyeball
Post-test	19	Main	Watch how I wave my hands
Post-test	20	Main	Target one sock from each pair
Post-test	21	Main	The answer to that differs greatly
Post-test	22	Main	Nearly ran right into you
Post-test	23	Main	You know what I hate?
Post-test	24	Main	A half pound of bacon please
Post-test	25	New1	A very proud achievement for you
Post-test	26	New1	Because how long has that been
Post-test	27	New1	But how do you know that?
Post-test	28	New1	Are you getting them right now?
Post-test	29	New2	What people would end up doing
Post-test	30	New2	I'm happy for you mate
Post-test	31	New2	I wasn't there to help
Post-test	32	New2	Check the size of that

Table 5: List of all sentences used in the dictation task

Appendix B: attention check questions

1. Wat gebeurt er wanneer de hoofdpersoon zijn kaartenhuis wil afmaken?
    - Hij zet de kaart er te hard op, waardoor het kaartenhuis instort
    - Hij stoot tegen de tafel, waardoor de tafel en het kaartenhuis instorten
    - Hij ademt hard uit, en blaast het kaartenhuis omver
2. Waarom wil de hoofdpersoon weg uit de koffiezaak?
    - Hij vindt de andere bezoekers vervelend
    - Hij is er voor de Wi-Fi, maar die heeft hij thuis ook
    - Hij vindt de thee niet lekker
3. Waarom staart de man in de lunchzaak naar de hoofdpersoon?
    - De hoofdpersoon zit onder de menukaart, en de man is die aan het lezen
    - De man voelt zich erg verbonden met de hoofdpersoon
    - De hoofdpersoon heeft een vlek op zijn shirt

Appendix C: background questionnaire

General background questions

1. Wat is je leerlingnummer?*
2. Wat is je geslacht?
    - Man
    - Vrouw
    - Overig
3. Wat is je geboortedatum? (dd/mm/yyyy)
4. Heb je dyslexie?
    - Ja
    - Nee
5. Heb je een taalstoornis?
    - Ja, ik heb ...
    - Nee
6. Is je gehoor goed?
    - Ja
    - Nee
7. Heb je ooit gehoorproblemen gehad?
    - Ja, ik heb ... gehad
    - Nee
8. Kun je goed zien?
    - Ja
    - Nee
9. Heb je aandachtsproblemen? (bijv. ADHD)
    - Ja
    - Nee

10. Is Nederlands je moedertaal?
    - Ja
    - Nee, mijn moedertaal is ...
11. Heb je naast het Nederlands nog andere moedertalen?
    - Ja, mijn andere moedertalen zijn ...
    - Nee

* This question was replaced by another question for the university students: "Wat studeer je?"

Questions about experience with the Scottish accent

1. De video die je hebt gekeken werd gesproken in een Schots accent. Heb je al eens eerder een Schots accent gehoord?
    - Ja
    - Nee
    - Ik weet het niet
2. Hoe vaak heb je het Schotse accent al gehoord en in welke context?*

* Question only asked if response was "yes" on previous question

Questions about the experiment

1. Is je iets opgevallen tijdens het onderzoek?
2. Waar denk je dat het onderzoek over ging?
3. Heb je nog opmerkingen over het onderzoek?

Appendix D: Python code

This code was used to clean up all the raw data from the dictation task and score it automatically. Running the code created two different files: one with all the scores using the lexical accuracy, and one file with the scores using the orthographic closeness.

```
#libraries
import csv
import Levenshtein as L
import numpy as np
from pathlib import Path

#variables
folder_root = Path("")
DataFileName = folder_root / "AllSentences.csv"
AnswerKeyName = folder_root / "AnswerKey.csv"
SaveFileL = folder_root / "ScoresL.csv"
SaveFileW = folder_root / "ScoresW.csv"

WordReplacements = {"it's": "it is", "humor": "humour",
                    "you are":"you're", "cause":"because",
                    "i am":"i'm", "i was not": "i wasn't"}
SCORE_L = 0
SCORE_W = 1
SCORES = "LW"

# loading functions
def load_data(filename):
    datafile = []
    with open(filename, newline='') as fin:
        reader = csv.reader(fin, delimiter=',')
        header = next(reader)
        for line in reader:
            datafile.append(line)
    datafile = list(map(list, zip(*datafile)))

    PPInfo = datafile[:7]
    PPAnswers = datafile[7:]

    return header, PPInfo, PPAnswers
```

```
def load_key(filename):
    with open(filename, newline='') as fin:
        reader = csv.reader(fin, delimiter=',')
        next(fin)
        answerkey = next(fin).strip().split(',')

    print(answerkey)
    return answerkey

# saving function
def save_data(filename, header, data):
    data = list(map(list, zip(*data)))

    with open(filename, 'w', newline='') as fin:
        writer = csv.writer(fin, delimiter=',')
        writer.writerow(header)
        writer.writerows(data)

# cleaning function
def clean_sentence(sentence):
    clean = ""
    for character in sentence:
        if character.isalpha() \
                or character.isspace() or character == "'":
            clean += character

    clean = ' '.join(clean.split()).lower()
    return clean

# scoring functions
def score_answer(transcription, rightanswer, score_type=SCORE_L):
    if score_type==SCORE_L:
        score = L.distance(transcription, rightanswer)
        normed_score = score / max(len(transcription),
                                   len(rightanswer))
        rounded_score = np.around(normed_score, decimals=2)
        return rounded_score
    elif score_type==SCORE_W:
        score = sum([x in transcription.split()
                     for x in rightanswer.split()])
```

```
        normed_score = 1-(score / max(len(transcription.split()),
                                      len(rightanswer.split())))
        rounded_score = np.around(normed_score, decimals=2)
        return rounded_score

# ReplaceByDict function: replace a column in a
# dataframe using a dictionary
def ReplaceByDict(text, dictionary):
    for key in dictionary.keys():
        text = text.replace(key, dictionary[key])
    return text

# read files
header, PPInfo, PPAnswers = load_data(DataFileName)
AnswerKey = load_key(AnswerKeyName)

# clean data
ProcessedData = []
for item in PPAnswers:
    ProcessedColumns = []
    for answer in item:
        AnswerClean = clean_sentence(answer)
        AnswerReplaced = ReplaceByDict(AnswerClean,
                                       WordReplacements)
        ProcessedColumns.append(AnswerReplaced)
    ProcessedData.append(ProcessedColumns)

# score data
def score_on_columns(score_type):
    ScoredData = []
    for these_pp_answers, this_right_answer \
            in zip(ProcessedData,AnswerKey):
        ScoredColumns = []
        for this_pp_answer in these_pp_answers:
            score = score_answer(this_pp_answer,
                                 this_right_answer,
                                 score_type)
            ScoredColumns.append(score)
        ScoredData.append(ScoredColumns)
    return ScoredData
```

```
data_score_L = score_on_columns(SCORE_L)
data_score_W = score_on_columns(SCORE_W)

# generate output files
L_scores_complete = PPInfo.copy()
W_scores_complete = PPInfo.copy()

L_scores_complete.extend(data_score_L)
W_scores_complete.extend(data_score_W)

save_data(SaveFileL, header, L_scores_complete)
save_data(SaveFileW, header, W_scores_complete)
```

Appendix E: R code

This code was used to transform the data to be used for the statistical analysis and to perform the actual statistical analysis.

```
rm(list=ls())

# include custom libraries

library(dplyr)
library(tidyr)
library(reshape2)
library(lme4)
library(car)
library(effects)

# global variables

L_filename <- "ScoresL.csv"
W_filename <- "ScoresW.csv"
info_filename <- "SpeakerInfo.csv"

# read in a datafile

read_datafile <- function(filename)
{
  input <- file(description=filename,open="r")
  dataname <- read.csv(file=input,header=TRUE,sep=",")
  close(input)
  return(dataname)
}

# process the datafile (filter out excluded participants,
   restructure into long format, sort by participant and
   then by item)

process_datafile <- function(filename,infofile)
{
  scores <- read_datafile(filename)
  scores_filtered <- filter(scores, Exclude == "N")
  scores_long <- gather(scores_filtered, Trial, Score, PRE1
     :POST16, factor_key=TRUE)
  scores_long_sorted <- arrange(scores_long, PPID, Trial)
```

```
  scores_info <- mutate(scores_long_sorted, Test = gsub("([
      A-Z]*)[0-9]*", "\\1",scores_long_sorted$Trial), Nr =
      as.integer(gsub("[A-Z]*([0-9]*)", "\\1",scores_long_
      sorted$Trial)))
  scores_speakerinfo <- mutate(scores_info, Speaker = case_
      when((Nr >= 1 & Nr <= 8) ~ "Speaker1", (Nr >= 9 & Nr
      <= 12) ~ "Speaker2", TRUE ~ "Speaker3"))
  scores_replacecondition <- mutate(scores_speakerinfo,
      Condition2 = case_when((Condition == "NoSub") ~ "_
      NoSub",(Condition == "DutSub") ~ "DutSub", TRUE ~ "
      EngSub"))
  scores_groupinfo <- mutate(scores_replacecondition,
      Groupsimple = case_when(Group == "Student" ~ "
      University", TRUE ~ "Highschool"))

  scores_groupinfo$PPID <- as.factor(scores_groupinfo$PPID)
  scores_groupinfo$Test <- as.factor(scores_groupinfo$Test)
  scores_groupinfo$Nr <- as.factor(scores_groupinfo$Nr)
  scores_groupinfo$Speaker <- as.factor(scores_groupinfo$
      Speaker)
  scores_groupinfo$Condition <- as.factor(scores_groupinfo$
      Condition)
  scores_groupinfo$Condition2 <- as.factor(scores_groupinfo
      $Condition2)
  scores_groupinfo$Groupsimple <- as.factor(scores_
      groupinfo$Groupsimple)

  scores_inverted <- mutate(scores_groupinfo, Score_inverse
      = (1-Score))

  return(scores_inverted)
}

add_means <- function(scores)
{
  scores_pretest <- filter(scores, Test == "PRE")
  mean_pretest_bypp <- group_by(scores_pretest, PPID)%>%
      summarise(Mean_PRE = mean(Score), Mean_PRE_inverse = (
      mean(Score_inverse)))
  merged_table <- merge(scores, mean_pretest_bypp)
  return(merged_table)
}
```

```
# main
run_all_functions <- function(scoretype)
{
  setwd("~/DictationScores")
  if (scoretype == "L")
  {
    L_scores <- process_datafile(L_filename)
    L_complete <- add_means(L_scores)
    return(L_complete)
  }
  else
  {
    W_scores <- process_datafile(W_filename)
    W_complete <- add_means(W_scores)
    return(W_complete)
  }
}

df <- run_all_functions("L")

df <- filter(df, Test=="POST")

### 1-factor models (L) ###

m1 <- lmer(Score_inverse ~ Condition2 + (1|PPID) + (1|Trial
    ), data=df)
m2 <- lmer(Score_inverse ~ Mean_PRE_inverse + (1|PPID) +
    (1|Trial), data=df)
m3 <- lmer(Score_inverse ~ Speaker + (1|PPID) + (1|Trial),
    data=df)
m4 <- lmer(Score_inverse ~ Groupsimple + (1|PPID) + (1|
    Trial), data=df)

### Model using pre-test score (L) ###

m5 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
    (1|PPID) + (1|Trial), data=df)
m6 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
    Condition2*Mean_PRE_inverse + (1|PPID) + (1|Trial), data
    =df)
m7 <- lmer(Score_inverse ~ Condition2 + Speaker + (1|PPID)
    + (1|Trial), data=df)
```

```
m8 <- lmer(Score_inverse ~ Condition2 + Speaker +
    Condition2*Speaker + (1|PPID) + (1|Trial), data=df)
m9 <- lmer(Score_inverse ~ Speaker + Mean_PRE_inverse + (1|
    PPID) + (1|Trial), data=df)
m10 <- lmer(Score_inverse ~ Speaker + Mean_PRE_inverse +
    Speaker*Mean_PRE_inverse + (1|PPID) + (1|Trial), data=df
    )
m11 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Speaker + (1|PPID) + (1|Trial), data=df)
m12 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Speaker + Condition2*Mean_PRE_inverse + (1|PPID) + (1|
    Trial), data=df)
m13 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Speaker + Condition2*Speaker + (1|PPID) + (1|Trial),
    data=df)
m14 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Speaker + Mean_PRE_inverse*Speaker + (1|PPID) + (1|
    Trial), data=df)
m15 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Speaker + Condition2*Mean_PRE_inverse + Condition2*
    Speaker + Mean_PRE_inverse*Speaker + Condition2*Mean_PRE
    _inverse*Speaker + (1|PPID) + (1|Trial), data=df)
#best model for L+pre-test
mLPbest <- lmer(Score_inverse ~ Condition2 + Mean_PRE_
    inverse + Condition2*Mean_PRE_inverse + (1|PPID) + (1|
    Trial), data=df)

### Model using group instead of pre-test (L) ###

m16 <- lmer(Score_inverse ~ Condition2 + Groupsimple + (1|
    PPID) + (1|Trial), data=df)
m17 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Condition2*Groupsimple + (1|PPID) + (1|Trial), data=df)
m18 <- lmer(Score_inverse ~ Condition2 + Speaker + (1|PPID)
     + (1|Trial), data=df)
m19 <- lmer(Score_inverse ~ Condition2 + Speaker +
    Condition2*Speaker + (1|PPID) + (1|Trial), data=df)
m20 <- lmer(Score_inverse ~ Speaker + Groupsimple + (1|PPID
    ) + (1|Trial), data=df)
m21 <- lmer(Score_inverse ~ Speaker + Groupsimple + Speaker
    *Groupsimple + (1|PPID) + (1|Trial), data=df)
m22 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Speaker + (1|PPID) + (1|Trial), data=df)
```

```
m23 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Speaker + Condition2*Groupsimple + (1|PPID) + (1|Trial),
     data=df)
m24 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Speaker + Condition2*Speaker + (1|PPID) + (1|Trial),
    data=df)
m25 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Speaker + Groupsimple*Speaker + (1|PPID) + (1|Trial),
    data=df)
m26 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Speaker + Condition2*Groupsimple + Condition2*Speaker +
    Groupsimple*Speaker + Condition2*Groupsimple*Speaker +
    (1|PPID) + (1|Trial), data=df)

#best model for L+group
mLGbest <- lmer(Score_inverse ~ Groupsimple + Speaker +
    Groupsimple*Speaker + (1|PPID) + (1|Trial), data=df)

df <- run_all_functions("W")
df <- filter(df, Test=="POST")

### 1-factor models (W) ###

m27 <- lmer(Score_inverse ~ Condition2 + (1|PPID) + (1|
    Trial), data=df)
m28 <- lmer(Score_inverse ~ Mean_PRE_inverse + (1|PPID) +
    (1|Trial), data=df)
m29 <- lmer(Score_inverse ~ Speaker + (1|PPID) + (1|Trial),
     data=df)
m30 <- lmer(Score_inverse ~ Groupsimple + (1|PPID) + (1|
    Trial), data=df)

### Model using pre-test score (W) ###

m31 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     (1|PPID) + (1|Trial), data=df)
m32 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Condition2*Mean_PRE_inverse + (1|PPID) + (1|Trial),
    data=df)
m33 <- lmer(Score_inverse ~ Condition2 + Speaker + (1|PPID)
     + (1|Trial), data=df)
m34 <- lmer(Score_inverse ~ Condition2 + Speaker +
    Condition2*Speaker + (1|PPID) + (1|Trial), data=df)
```

```
m35 <- lmer(Score_inverse ~ Speaker + Mean_PRE_inverse +
    (1|PPID) + (1|Trial), data=df)
m36 <- lmer(Score_inverse ~ Speaker + Mean_PRE_inverse +
    Speaker*Mean_PRE_inverse + (1|PPID) + (1|Trial), data=df
    )
m37 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Speaker + (1|PPID) + (1|Trial), data=df)
m38 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Speaker + Condition2*Mean_PRE_inverse + (1|PPID) + (1|
    Trial), data=df)
m39 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Speaker + Condition2*Speaker + (1|PPID) + (1|Trial),
    data=df)
m40 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Speaker + Mean_PRE_inverse*Speaker + (1|PPID) + (1|
    Trial), data=df)
m41 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse +
     Speaker + Condition2*Mean_PRE_inverse + Condition2*
    Speaker + Mean_PRE_inverse*Speaker + Condition2*Mean_PRE
    _inverse*Speaker + (1|PPID) + (1|Trial), data=df)

#best model for W+pre-test
mWPbest <- lmer(Score_inverse ~ Mean_PRE_inverse + (1|PPID)
     + (1|Trial), data=df)

#### Model using group instead of pre-test (W) ###

m42 <- lmer(Score_inverse ~ Condition2 + Groupsimple + (1|
    PPID) + (1|Trial), data=df)
m43 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Condition2*Groupsimple + (1|PPID) + (1|Trial), data=df)
m44 <- lmer(Score_inverse ~ Condition2 + Speaker + (1|PPID)
     + (1|Trial), data=df)
m45 <- lmer(Score_inverse ~ Condition2 + Speaker +
    Condition2*Speaker + (1|PPID) + (1|Trial), data=df)
m46 <- lmer(Score_inverse ~ Speaker + Groupsimple + (1|PPID
    ) + (1|Trial), data=df)
m47 <- lmer(Score_inverse ~ Speaker + Groupsimple + Speaker
    *Groupsimple + (1|PPID) + (1|Trial), data=df)
m48 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Speaker + (1|PPID) + (1|Trial), data=df)
m49 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Speaker + Condition2*Groupsimple + (1|PPID) + (1|Trial),
```

```
    data=df)
m50 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Speaker + Condition2*Speaker + (1|PPID) + (1|Trial),
    data=df)
m51 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Speaker + Group*Speaker + (1|PPID) + (1|Trial), data=df)
m52 <- lmer(Score_inverse ~ Condition2 + Groupsimple +
    Speaker + Condition2*Groupsimple + Condition2*Speaker +
    Groupsimple*Speaker + Condition2*Groupsimple*Speaker +
    (1|PPID) + (1|Trial), data=df)

#best model for W + group
mWGbest <- lmer(Score_inverse ~ Groupsimple + Speaker +
    Groupsimple*Speaker + (1|PPID) + (1|Trial), data=df)

### Comparison of models ###

#activate function that is needed by removing the #
#Anova(model1, model2)
#print(Anova(bestmodel))
#print(summary(bestmodel))

### Plot ###

plot(allEffects(mLGbest), xlab = "Pre-test Score", ylab = "
    Post-test Score", main="")

#### Additional replication ####

#using L score
df <- run_all_functions("L")

df <- filter(df, Test=="POST")

dfalleenstudent <- filter(df, Group=="Student")

#1-factor models
m1001 <- lmer(Score_inverse ~ Condition2 + (1|PPID) + (1|
    Trial), data=dfalleenstudent)
m1002 <- lmer(Score_inverse ~ Mean_PRE_inverse + (1|PPID) +
     (1|Trial), data=dfalleenstudent)
m1003 <- lmer(Score_inverse ~ Speaker + (1|PPID) + (1|Trial
    ), data=dfalleenstudent)
```

```
#models using pre-test score
m1005 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
    + (1|PPID) + (1|Trial), data=dfalleenstudent)
m1006 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
    + Condition2*Mean_PRE_inverse + (1|PPID) + (1|Trial),
   data=dfalleenstudent)
m1007 <- lmer(Score_inverse ~ Condition2 + Speaker + (1|
   PPID) + (1|Trial), data=dfalleenstudent)
m1008 <- lmer(Score_inverse ~ Condition2 + Speaker +
   Condition2*Speaker + (1|PPID) + (1|Trial), data=
   dfalleenstudent)
m1009 <- lmer(Score_inverse ~ Speaker + Mean_PRE_inverse +
   (1|PPID) + (1|Trial), data=dfalleenstudent)
m1010 <- lmer(Score_inverse ~ Speaker + Mean_PRE_inverse +
   Speaker*Mean_PRE_inverse + (1|PPID) + (1|Trial), data=
   dfalleenstudent)
m1011 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
    + Speaker + (1|PPID) + (1|Trial), data=dfalleenstudent)
m1012 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
    + Speaker + Condition2*Mean_PRE_inverse + (1|PPID) +
   (1|Trial), data=dfalleenstudent)
m1013 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
    + Speaker + Condition2*Speaker + (1|PPID) + (1|Trial),
   data=dfalleenstudent)
m1014 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
    + Speaker + Mean_PRE_inverse*Speaker + (1|PPID) + (1|
   Trial), data=dfalleenstudent)
m1015 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
    + Speaker + Condition2*Mean_PRE_inverse + Condition2*
   Speaker + Mean_PRE_inverse*Speaker + Condition2*Mean_PRE
   _inverse*Speaker + (1|PPID) + (1|Trial), data=
   dfalleenstudent)

mLSbest <- lmer(Score_inverse ~ Mean_PRE_inverse + (1|PPID)
    + (1|Trial), data=dfalleenstudent)

#print(summary(mLSbest))
#Anova(mLSbest)
#anova(mLSbest,m1015)

#plot
```

```
plot(allEffects(mLSbest), xlab = "Pre-test Score", ylab = "
    Post-test Score", main="")

#using W score

df <- run_all_functions("W")

df <- filter(df, Test=="POST")

dfalleenstudent <- filter(df, Group=="Student")

#1 factor models

m1027 <- lmer(Score_inverse ~ Condition2 + (1|PPID) + (1|
    Trial), data=dfalleenstudent)
m1028 <- lmer(Score_inverse ~ Mean_PRE_inverse + (1|PPID) +
     (1|Trial), data=dfalleenstudent)
m1029 <- lmer(Score_inverse ~ Speaker + (1|PPID) + (1|Trial
    ), data=dfalleenstudent)

#models using pre-test score

m1031 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
     + (1|PPID) + (1|Trial), data=dfalleenstudent)
m1032 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
     + Condition2*Mean_PRE_inverse + (1|PPID) + (1|Trial),
    data=dfalleenstudent)
m1033 <- lmer(Score_inverse ~ Condition2 + Speaker + (1|
    PPID) + (1|Trial), data=dfalleenstudent)
m1034 <- lmer(Score_inverse ~ Condition2 + Speaker +
    Condition2*Speaker + (1|PPID) + (1|Trial), data=
    dfalleenstudent)
m1035 <- lmer(Score_inverse ~ Speaker + Mean_PRE_inverse +
    (1|PPID) + (1|Trial), data=dfalleenstudent)
m1036 <- lmer(Score_inverse ~ Speaker + Mean_PRE_inverse +
    Speaker*Mean_PRE_inverse + (1|PPID) + (1|Trial), data=
    dfalleenstudent)
m1037 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
     + Speaker + (1|PPID) + (1|Trial), data=dfalleenstudent)
m1038 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
     + Speaker + Condition2*Mean_PRE_inverse + (1|PPID) +
    (1|Trial), data=dfalleenstudent)
```

```
m1039 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
    + Speaker + Condition2*Speaker + (1|PPID) + (1|Trial),
   data=dfalleenstudent)
m1040 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
    + Speaker + Mean_PRE_inverse*Speaker + (1|PPID) + (1|
   Trial), data=dfalleenstudent)
m1041 <- lmer(Score_inverse ~ Condition2 + Mean_PRE_inverse
    + Speaker + Condition2*Mean_PRE_inverse + Condition2*
   Speaker + Mean_PRE_inverse*Speaker + Condition2*Mean_PRE
   _inverse*Speaker + (1|PPID) + (1|Trial), data=
   dfalleenstudent)

mWSbest <- lmer(Score_inverse ~ Mean_PRE_inverse + (1|PPID)
    + (1|Trial), data=dfalleenstudent)

#print(summary(mWSbest))
#Anova(mWSbest)
#anova(mWSbest,m1041)

#plot
plot(allEffects(mWSbest), xlab = "Pre-test Score", ylab = "
   Post-test Score", main="")

#### Additional ANOVAs interaction effect ####

#processing without putting the pre-test scores as a column
    but keeping them as row entries
run_only_processing <-function(scoretype)
{
  setwd("~/DictationScores")
  if (scoretype == "L")
  {
    L_scores <- process_datafile(L_filename)
    return(L_scores)
  }
  else
  {
    W_scores <- process_datafile(W_filename)
    return(W_scores)
  }
}

df2 <- run_only_processing("L")
```

```
df2pretestonly <- filter(df2, Test == "PRE")
pretest_bycondition <- aov(Score_inverse ~ Condition2 +
    Error(PPID), data = df2pretestonly)

df2posttestonly <- filter(df2, Test == "POST")
posttest_bycondition <- aov(Score_inverse ~ Condition2 +
    Error(PPID), data = df2posttestonly)

df2nosubonly <- filter(df2, Condition2 == "_NoSub")
nosub_bytest <- aov(Score_inverse ~ Test + Error(PPID/Test)
    , data = df2nosubonly)

df2dutsubonly <- filter(df2, Condition2 == "DutSub")
dutsub_bytest <- aov(Score_inverse ~ Test + Error(PPID/Test
    ), data = df2dutsubonly)

df2engsubonly <- filter(df2, Condition2 == "EngSub")
engsub_bytest <- aov(Score_inverse ~ Test + Error(PPID/Test
    ), data = df2engsubonly)
```