The Roles of Familiarity, Intelligibility and Attitude in the Processing of L1 Accents

A thesis submitted to the

School of Education, Communication and Language Sciences

for the degree of Doctor of Philosophy

Andreas Krug

July 2023

Abstract

This thesis presents an experimental investigation into the roles of familiarity, intelligibility and attitude in the lexical processing and short-term narrative recall of first language (L1) speech. Lev-Ari and Keysar (2012) suggest that the lexical and semantic representations of second language (L2) speech by L1 listeners are less detailed, potentially due to assumptions of lower linguistic proficiency in L2 speakers. These assumptions would not typically apply to L1 speakers. However, even for L1 speakers, listeners vary in terms of their familiarity with the speakers' accent, how intelligible it is to them and their attitudes towards it, all of which can affect speech processing.

Three online experiments were conducted to research the effects of familiarity, intelligibility and attitude. A lexical decision task was used to measure lexical processing. Recall was measured with a change detection task that included short stories and variations of the semantic proximity of the change. The experiments included manipulations of, first, the participants' familiarity with the speakers' accent; second, listening conditions (through added background noise) and; third, the participants' attitudes towards speakers.

Regarding lexical processing, the results showed a familiarity benefit and a window of opportunity for adaptation, which opened when lexical access was demanding but not too challenging. Negative attitudes were associated with poorer performance. For recall, a familiarity benefit only surfaced in noise. Recall performance further increased when the changed words were semantically related and when participants held negative attitudes towards speakers. These findings can be accounted for within a framework of semantic priming and item-specific versus relational processing. Recall is supported by semantic priming and item-specific processing. The latter seems to be induced by unfamiliar accents (in quiet), adverse listening conditions and negative attitudes. Taken together, this suggests that the less-detailed representations suggested for L2 accents do not generally apply to L1 accents.

Acknowledgements

Doing a PhD has been a challenging, rewarding and humbling experience for me. There are many fantastic people that contributed in their own way to this project and I am very grateful for every one of them in my life. The following (very likely non-exhaustive) list contains some of the greatest minds I was fortunate enough to cross paths with:

- To Ghada and Laurence: Thank you for your patience, guidance and feedback along the way. This PhD would not have been possible without your help from start to finish. I have learned a lot from you and I hope that many future PhD students will benefit from your expertise and realistic approach to academia.
- To all the speakers who lent their voices for the experiments: Thank you for taking time out of your day to do so. I can imagine that there are nicer things to do than reading the same sentence with a strange nonword three times.
- To all the participants who took part in the experiment: Thank you for letting me collect your data and analyse it. These experiments might not have been the most fun but I could not have done them without you.
- To the NINE Doctoral Training Partnership: Thank you for accepting our funding application and for the support throughout the process. I would not have been able to attend conferences, connect with other NINE students and hone my soft skills without your generous support.
- To Dan, Jalal, Damien and Nadine: Thank you for being part of my APR and viva panel. Your feedback over the years and at the very end of my PhD was extremely helpful not only in terms of shaping this thesis but also with regards to life inside and outside academia.
- To the Phonetics and Phonology Research Group: Thank you for listening to my research ideas, methodological problems and mock conference presentations. Your support and companionship were invaluable.
- To the support staff at PCIbex (especially Jérémy Zehr) and LabVanced: Thank you for your patience and help with setting up the experiments. Many hours went into this and it would have been impossible without your help.

- To my writing group, office, teaching and PGR mates April, Bruce, Caitlin, Carol-Ann, Damar, Fengting, Gustavo, Hajar, Jasmin, Joaquín, Lucia, Olivia, Saad, Sascha, Sergio, Sheradan, Sherly, Thamer and Yanyu: Thank you for being the best company and friends I could have wished for in Newcastle. I already have so many fond memories of our time together and I hope that we will be able to add many more in the future. You are all amazing and I wish you all the best on your journeys.

- To Hanain and Maaly: Thank you for being my ECLS buddies throughout COVID and beyond. I still remember the day you first came into our little office and I am beyond grateful that you stayed. I believe that throughout the years we talked about everything from pop culture to PhD misery. You are basic and I love you for it.

- To Febee, Jacob, Jason, Lisa and Mac: Thank you for hopping on to Google Meets to talk about everything and anything, take random quizzes and play Ludo for hours during lockdown. You contributed so much more to this thesis than you might know and I look forward to celebrating with you one day. Maybe at a reunion in Detroit?

- To Barbara, Daniel, Franzi, Isi, Lukas, Pia and Sarah: Thank you for your friendship throughout the years. I can hardly believe that I have known you all for more than ten years now. There are so many moments in life with you that I will cherish forever.

- To my Mum Hedwig, my Dad Otto and my brother Christopher: Thank you for everything you have done for me. Having you in my life means the world to me. Unfortunately, I could not visit home too many times because of lockdowns but I hope to be able to see you more often in the future. Thank you for the trips you made to the UK and for always being there for me.

  Mama, Papa und Christopher, vielen Dank für all das, was ihr für mich getan habt. Es bedeutet mir alles, euch in meinem Leben zu haben. Leider konnte ich wegen der Lockdowns night so oft nach Hause kommen, aber ich hoffe, dass ich euch in Zukunft öfter sehen kann. Vielen Dank für eure Reisen ins UK and dafür, dass ihr immer für mich da seid.

- To Marcus: Thank you for standing by my side and for your love, wisdom and support throughout the years. We might be an old couple now but I love you a bit more every day and I could not have finished this PhD without you. Thank you for becoming my husband, for taking me on trips and for helping me to take my mind off work. You are the best.

Contents

List of Tables

List of Figures

List of Abbreviations

AE	American English
BE	British English
CDA	Contralateral delay activity
Covert-NY	Rhotic listeners born and raised in New York City
DB	Decibel(s)
DH	Different processes hypothesis
DRM	Deese-Roediger-McDermott
ECLS	School of Education, Communication and Language Sciences
ERP	Event-related potential
IAPS	International affective picture system
IAT	Implicit association test
GA	General American
GE	Glaswegian English
LDL	Lexical decision latency
MS	Millisecond
NYC	New York City
NZE	New Zealand English
Overt-NY	Non-rhotic listeners born and raised in New York City
PDH	Perceptual difference hyothesis
PMN	Phonological mapping negativity
NNSR	Non-native speech recognition
RP	Received Pronunciation
S	Second
SD	Standard deviation
SE	Standard Error
SNR	Signal-to-noise ratio
SSBE	Standard Southern British English
TE	Tyneside English
UK	United Kingdom

Chapter 1

Introduction

Interactions between speakers of different accents are ubiquitous in both private and professional settings. As a result, a better understanding of the mechanisms underlying the processing of different accents can foster linguistic theory and effective communication alike. This thesis presents an experimental investigation into the effects of familiarity, intelligibility and attitude on the lexical processing and narrative recall of first language (L1) speech.

The starting point for the current research was a finding which suggests that L1 listeners process and represent second language (L2) speech with less lexical and semantic detail than L1 speech (Lev-Ari & Keysar, 2012). This less-detailed processing mode could be due to assumptions of L2 speakers' lower linguistic proficiency. Thus, lexical and semantic processing becomes more efficient for L2 speech if it is executed in less detail and, thus, glances over potential linguistic inaccuracies. Lower linguistic proficiency would not be expected for L1 speakers but listeners might still vary in terms of how familiar they are with the speaker's accent, how intelligible it is to them and which attitudes they hold towards it. Past research has shown that these factors influence the lexical processing and recall of L1 and L2 speech. For example, lexical processing is generally better for familiar accents, especially in noise (e.g. Adank & McQueen, 2007; Stringer & Iverson, 2019). In addition, adverse listening conditions have been shown to be associated with item-specific processing, which refers to the "encoding [of] items by their features, elements, and distinctive qualities [...] [rather than] in relation to other concepts in memory" and, in turn, fosters recall (Kjellberg, Ljung, and Hallman, 2008; Ljung and Kjellberg, 2008; Storbeck and Clore, 2005, 786). Finally, negative affect, which could reasonably be assumed to be linked to negative attitudes, also induced item-specific

processing while lowering working memory capacity (e.g. Storbeck, 2013; Storbeck & Clore, 2005) and, thus, affecting speech processing.

However, previous studies on the influence of familiarity, intelligibility and attitude differ in terms of the theoretical background, accents, listening conditions and tasks included. The research presented in this thesis uses a single methodology to examine the three factors of interest within three online experiments. Lexical processing was measured by means of a lexical decision task, with a real English word or a nonsense word placed at the end of standardised sentences. Recall was assessed via a change detection task that included two short stories, which were either identical or differed in a single word to be recalled by the participants. This allowed for a thorough analysis of each factor and, importantly, for a comparison of the findings across the three experiments. The present investigation is the first of its kind to consider familiarity, intelligibility and attitude in this comprehensive manner.

The remainder of this introductory chapter is divided into two parts. First, the research questions for each factor of interest are provided. Second, an overview of the structure of this thesis and its seven chapters is given.

1.1 Research Questions

The overarching research question for this thesis is as follows: What are the effects of familiarity, intelligibility and attitude on the lexical processing and recall of L1 speech? This overarching question is broken down into more specific research questions for each factor of interest. They are provided below to show the agenda of the current research and will be contextualised more in the corresponding thesis chapters.

Familiarity:

(1) How does accent familiarity affect lexical processing?

(2) Is there evidence for adaptation in terms of lexical processing?

(3) What is the effect of word frequency on lexical processing?

(4) How do accent familiarity and semantic proximity of the change affect recall?

(5) Overall, is there evidence for less-detailed processing of the unfamiliar versus familiar accent?

Intelligibility:

(1) How do adverse listening conditions and accent familiarity affect lexical processing?

(2) Is there evidence for adaptation in terms of lexical processing?

(3) What is the effect of word frequency on lexical processing?

(4) How do adverse listening conditions, accent familiarity and semantic proximity of the change affect recall?

(5) Overall, is there evidence for less-detailed processing in quiet versus in noise and/or for the unfamiliar versus familiar accent?

Attitude:

(1) How do language attitudes affect lexical processing?

(2) Is there evidence for adaptation in terms of lexical processing?

(3) What is the effect of word frequency on lexical processing?

(4) How do language attitudes and the semantic proximity of the change affect recall?

(5) Overall, is there evidence for less-detailed processing when the speaker is evaluated positively versus negatively?

1.2 Structure of this Thesis

This thesis is structured into seven chapters:

Following this introduction (Chapter 1), Chapter 2 reviews relevant literature on speech processing, as mediated by familiarity, intelligibility and attitude. First, abstractionist versus exemplar-based models as well as a hybrid model of speech and lexical processing are discussed (e.g. Drager & Kirtley, 2016; Marslen-Wilson, 1987; McClelland & Elman, 1986; Nosofsky, 2011; Stevens, 2008; Sumner et al., 2013). While the aim of this thesis is not to argue for one or another model, these theoretical considerations are necessary to set the general context for the current research. Next, the three factors of interest are addressed. For each factor, different operationalisations are provided, before relevant past studies on lexical processing and recall are reviewed.

Chapter 3 is the first experimental chapter and examines the effect of accent familiarity on lexical processing and recall. Lexical processing was measured via a lexical decision task, during which participants had to decide if the final word of a sentence was a real English word or a nonword. A change detection task was used to measure recall. Participants heard pairs of short stories and had to decide if the stories were identical or different. In the latter case, they were asked to recall the changed word between the two stories. Participants from Tyneside completed these tasks for Tyneside English (TE; familiar accent) and New Zealand English (NZE; unfamiliar accent). The lexical processing results showed that lexical decisions became faster and more accurate for trials with TE speakers but faster and less accurate for trials with NZE speakers (see Floccia, Butler, Goslin, & Ellis, 2009; Floccia, Goslin, Girard, & Konopczynski, 2006; Impe, Geeraerts, & Speelman, 2009). In terms of recall, performance was better for the unfamiliar accent and for semantically related rather than unrelated changes between stories (see Clopper, Tamati, & Pierrehumbert, 2016; Frances, Costa, & Baus, 2018; Grohe & Weber, 2018; Lev-Ari & Keysar, 2012).

Chapter 4 is concerned with the experiment on intelligibility. The experimental design was similar to the one of the experiment on familiarity, with two important modifications. First, noise was added to all stimuli to increase task difficulty. Second, participants from New Zealand as well as Tyneside were recruited. This allowed for the analysis of two sets of results: Tynesiders versus New Zealanders in noise and Tynesiders in quiet versus in noise. Under adverse listening conditions, a familiarity benefit emerged in that TE was processed more accurately and more quickly by Tyneside participants (see Adank & Janse, 2010; Adank & McQueen, 2007; Stringer & Iverson, 2019). For the recall task, a similar familiarity benefit was found in that Tyneside participants performed better for their own accent if the stories included a change. Semantic proximity was again associated with better recall (see Hällgren, Larsby, Lyxell, & Arlinger, 2001; Kjellberg et al., 2008; Ljung, Sörqvist, Kjellberg, & Green, 2009; Marsh, Ljung, Nöstl, Threadgold, & Campbell, 2015).

Chapter 5 presents Experiment 3, which deals with the effects of attitude on lexical processing and recall. Participants from the North East of England completed the experiment and only stimuli recorded in TE were included. Importantly, there was an attitude manipulation such that participants were induced to hold positive attitudes towards one of the TE speakers and negative attitudes towards the other or vice versa. The results showed decreased lexical decision accuracy for the negatively presented speaker (see Figueira et al., 2017). Incongruent with the results from the previous experiments, semantic proximity did not

generally seem to aid recall. Instead, recall for semantically related changes was only better if the speaker was presented negatively rather than positively, potentially because negative attitudes correlate with items-specific processing (see Kensinger, 2009; Levine & Bluck, 2004; Storbeck, 2013; Storbeck & Clore, 2005).

Chapter 6 discusses the results from the three experimental chapters in relation to one another and to previous research. The current results are considered in relation to working memory, benefits of accent familiarity, semantic priming (McNamara, 2005; Perea & Rosa, 2002), task difficulty and item-specific versus relational processing. It is suggested that there is a familiarity benefit in speech processing (see Floccia et al., 2009, 2006; Impe et al., 2009), which might only emerge under adverse listening conditions (see Adank & McQueen, 2007; Stringer & Iverson, 2019). Additionally, recall benefits from item-specific processing, which might be induced by unfamiliar accents (in quiet), adverse listening conditions and negative attitudes (see Clopper et al., 2016; Grohe & Weber, 2018; Hällgren et al., 2001; Kensinger, 2009; Kjellberg et al., 2008; Levine & Bluck, 2004; Ljung et al., 2009; Marsh et al., 2015; Storbeck, 2013; Storbeck & Clore, 2005). Negative attitudes might decrease working memory capacity (see Figueira et al., 2017), which results in less accurate lexical processing. Potential interactions between familiarity, intelligibility and attitude are discussed and the importance of individual voice characteristics is reviewed, before directions for future research are provided.

Chapter 7 concludes this thesis by summarising the main results from the lexical decision and recall task as well as reiterating the potential main cognitive mechanisms behind these results, namely a familiarity benefit in speech processing and item-specific versus relational processing.

Chapter 2

Literature Review

2.1 Introduction

The aim of this literature review is to provide a summary of previous research into the effects of familiarity, intelligibility and attitude on the processing of different accents. The focus will be in particular on lexical processing and recall. The chapter starts off with an overview of models of speech processing and lexical access, thus explaining how the incoming speech stream is segmented and how lexical representations are accessed and stored. Abstractionist as well as exemplar-based models are addressed to provide an insight into current discussions in the field. However, the aim of this thesis is not necessarily to argue for or against one school of thought. Instead, the research questions centre around lexical access and recall performance as mediated by familiarity, intelligibility and attitude. Therefore, each factor will be addressed individually in the second part of this review in two steps. First, different operationalisations of each factor will be discussed as the approaches to familiarity, intelligibility and attitude varied in past research. Second, results from studies on lexical access and recall will be provided for each factor.

2.2 Models of Speech Processing

2.2.1 Abstractionist Models

Speech processing must be abstract to a certain extent as the acoustic signal produced by the speaker is not identical to the neural signal that is processed in the listener's brain. There are several stages, for example when the variations in air pressure oscillate through

the ear canal, during which "aspects of the signal may be enhanced, lost, or transformed" (Zsiga, 2022, 216). However, this kind of abstraction is not the one that abstractionist models typically refer to. Abstraction here refers to the units of representation, which are stored without surface-level information, and to the process of mapping the incoming signal on to them. Most abstractionist models assume two stages of speech processing. First, important acoustic cues are detected in the signal. In a second step, these cues are "matched to featural or phonemic representations that are stored in memory" (Zsiga, 2022, 230). Ultimately, the features or phonemes are combined to word forms, which allow for meaning retrieval in the mental lexicon and subsequent linguistic processing (see section 2.3).

Abstractionist models argue that the representations of words are "minimized in the sense of being stored using the fewest featural specifications possible" (Pierrehumbert, 2016, 34). Therefore, the number of stored representations of words is much smaller than the number of tokens found in the acoustic reality. This is facilitated by the phonological principle, according to which the word forms in the mental lexicon are combinations of a finite set of units, which are "meaningless in themselves, but in legal combinations they are associated with meanings" (Pierrehumbert, 2016, 37). To illustrate these ideas, different productions of the word *portal* can be considered. These productions might vary due to a number of reasons, including anatomical differences between speakers (thickness of vocal folds), accent differences (rhoticity) and speech rate (fast versus slow). Within an abstractionist framework, the aim is to, first, clean the signal from this surface-level variation and, second, map it on to stored abstract units (Johnson, 2008, 364). In doing so, different productions can be mapped on to the same representation. The representation, in turn, is comprised of abstract units, which can also be used for other words but, in this combination, carry a specific meaning. A central characteristic of this system is that it generalises over similar instances of the same word and phoneme. Such generalisations are necessary to process instances of identical words or phonemes that differ slightly in their acoustic characteristics from previously encountered ones. This avoids an over-trained system, which would only be able to detect input that is identical to stored tokens (see Pierrehumbert, 2016, 37-39 for a basic computation of this effect).

Drager and Kirtley (2016, 3) and Pierrehumbert (2016, 34, 40) summarise several findings that support the need for abstraction in speech processing.[1] A good starting point to argue for mechanisms of abstraction is the processing of artificial languages. As it is highly

[1] Many of these arguments deserve a more detailed discussion that is beyond the scope of this thesis. The intention here is provide a short summary rather than in-depth analysis of the need for abstraction in speech processing.

unlikely that subjects have encountered these languages before, they can offer insights into the building blocks of speech processing. Indeed, research by Richtsmeier (2011) has shown that the establishment of phonotactic rules in adults for an artificial language is driven by type rather than token frequency. This strongly suggests that the system operates beyond surface frequency and identifies more general patterns in the input that are retained in memory. Another point in favour of abstraction is that new perceptual categories are established very slowly while existing ones are extended quickly to include new tokens. This is evident from negative transfer effects in second language acquisition in both perception and production. For example, L1 Japanese learners of English frequently struggle to differentiate the English liquids /l, ɹ/ perceptually (e.g. Aoyama, Flege, Guion, Akahane-Yamada, & Yamada, 2004) and L1 German and Dutch speakers often have problems with producing the English dental fricatives /θ, ð/ (e.g. Hanulíková & Weber, 2010). Both of these findings can be explained by assuming that established phonemic categories are used to incorrectly subsume the L2 sounds. Abstraction is further supported by sound changes that occurred relatively independently of lexical frequency (Bybee, 2017). This phenomenon can be accounted for by abstraction and generalisation over a phonemic category. Finally, from a production point of view, speakers can pronounce and learn new (non)words quickly. This is sensible for abstractionist models since new (non)words are a combination of already existing units of representation.

The question remains how variation in the incoming signal due to indexical factors, which is considered noise in abstractionist models, is filtered out to allow for a mapping of variation on the acoustic surface to stored abstract units. One attempt to model this normalisation process for vowels has been made with formant ratio theories (e.g. Lehiste & Peterson, 1961), primarily with the aim to account for anatomical differences between speakers. These theories calculate the relative distances between vowel formants rather than consulting raw formant frequencies (Johnson, 2008, 366). In doing so, they can explain, for example, why the vowel [ʊ] typically has higher formant values in female than in male speech and yet listeners identify the same vowel. However, Johnson, Strand, and D'Imperio (1999) showed that formant ratios cannot be the only input for vowel identification. Listening to the same signals, their participants located the perceptual boundary between [u] and [ʊ] differently when they imagined a male versus female speaker.

Theories of vocal tract normalisation argue that vowels are normalised based on the estimated vocal tract length of the speaker during articulation. There is a number of cues that can be used by the listener to gauge the length of the vocal tract, such as F0 and F3 and,

if available, visual information (Johnson, 2008, 373-374). Vocal tract normalisation can also account for Johnson et al.'s (1999) finding, granted that differences between male versus female speakers are only due to (imagined) vocal tract length. In fact, vocal tract length theories are the foundation for normalisation procedures in sociophonetics (e.g. Lobanov, 1971 and Watt & Fabricius, 2003), which are used to compare the formants of male versus female speakers. These theories also predict that speech processing becomes more accurate if listeners are more familiar with the speaker and their vocal tract (Johnson, 2008, 374). Therefore, they can account for the finding that listeners can identify vowels better in single-talker than in mixed-talker conditions (Verbrugge, Strange, Shankweiler, & Edman, 1976). However, there are several sources of variation that are not captured by the vocal tract alone. For instance, male speakers usually have "a proportionally longer pharynx than women" and speakers may "adopt different [...] articulatory strategies to produce the 'same' sounds" (Johnson, 2008, 375; see also Samuel, 2011, 3). Furthermore, evidence that will be presented in detail below suggests that differences between male and female speakers are, to a certain extent, performed rather than a mere artifact of anatomical differences in vocal tract length.

2.2.2 Exemplar-Based Models

While abstractionist theories for speech processing have been developed since at least the 1950s (e.g. Potter & Steinberg, 1950), exemplar theory has "only gained wide acceptance in the phonetics/phonology community from the late 1990s" (Zsiga, 2022, 234). Originating in psychology, exemplar theory assumes that experiences are stored in the mind "as episodic memories, known as exemplars" (Drager & Kirtley, 2016, 2). Episodic memories were originally conceived as rich and detailed representations that individuals could remember consciously, for example "[t]elling an autobiographical story [...] or recalling an incident as if it were a movie in one's mind" (Pierrehumbert, 2016, 36). This notion of episodic memories or exemplars was altered in models of speech processing in that processing usually happens without conscious awareness. However, the idea of rich representations and a large memory space was maintained (Pierrehumbert, 2016, 36-37).

Accordingly, in contrast to the reasoning of abstractionist models, exemplars are not abstract units designed to account for a variety of surface forms. Instead, exemplars are rich representations of specific productions and contain indexical information. Indexical information "includes any number of markers of an individual's identity" (Drager & Kirtley, 2016, 5), ranging from their linguistic behaviour and physical appearance to their group membership

and one's relationship with them and so on. Social and linguistic information are inextricably linked in episodic models and the process of social indexing of linguistic information is assumed to be effortless (Foulkes & Docherty, 2006, 426-427). This makes exemplar theory a reasonable solution to many of the findings that challenge traditional abstractionist assumptions (see subsection 2.3.1).

Nosofsky's (2011) generalised context model will illustrate how incoming stimuli are compared against stored exemplars and classified into categories. Although the model was not developed for speech processing specifically, it demonstrates the classification process of stimuli in exemplar-based as opposed to abstractionist models well. Figure 2.1, which has been taken from Nosofsky (2011, 20), contains different exemplars. There are exemplars of categories A and B, which could easily be conceived of as two different sounds in a language. For each category, there are five exemplars. In the graph, the two categories vary along a horizontal and a vertical dimension. However, this is just for the sake of an easy illustration as it is assumed that memory is multidimensional. Exemplars that are more strongly encoded are printed in a larger font size. The incoming stimulus *i* needs to be classified into one of the categories. Within an abstractionist framework, which would only allow two abstract memory traces for A and B in the first place, the stimulus *i* would be stripped off any variation and then classified into one of the categories. In Nosofsky's (2011) generalised context model, on the other hand, the similarity between *i* and all stored exemplars is computed. For example, *i* is more similar to *B4* than *B1* on the horizontal axis. On the vertical axis, it is more similar to *B1* than *B4*. The stimulus is then categorised based on the overall similarity to the exemplars of categories A and B.[2] Importantly, exemplars contain rich memory representations and are not mere prototypes. In the context of speech processing, the exemplars are not abstract phonemes but contain fine-grained phonetic, social and contextual information. The categories emerge from exemplars that cluster together in the multidimensional mental space. As a result, abstraction is effectively the result of an online computation of similarities between exemplars.

[2] This is a simplified version of Nosofsky's (2011) model for illustration purposes. The model also includes a selective-attention mechanism that can stretch and condense axes. In Figure 2.1, categories A and B are mostly distinguished horizontally. Therefore, experience in categorisation of stimuli can stretch the horizontal axis to allow for an easier decision. Put differently, the main dimension along which exemplars of A and B vary becomes more fine-grained (Nosofsky, 2011, 19-20). The model further includes specific formulae for the categorisations (see Nosofsky, 2011, 21-36).

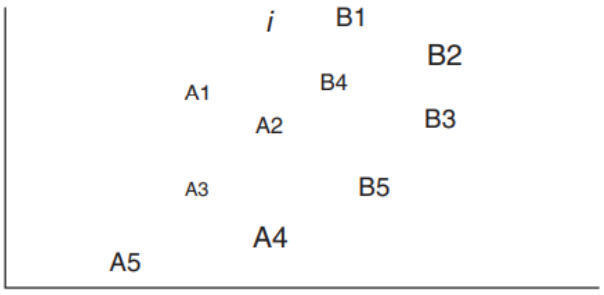


Figure 2.1: Categorisation in Nosofsky's (2011) generalised context model

Following the considerations above, models of speech processing usually assume that similar exemplars form clusters or exemplar clouds, which can be assigned a label (Drager & Kirtley, 2016, 3). Similarities can occur on a "semantic, phonological, visual, emotional, talker-specific [...] [and/or] event-specific" (Zsiga, 2022, 234) level. For example, different productions of the sound [s] would form an exemplar cloud (phonological similarity) but an exemplar cloud could also consist of representations of different sounds by the same speaker (talker-specific similarity). While this concept of labels suggests a hierarchical organisation of the network, it is essential to stress that these labels exist "'in addition to, not [as] a replacement of [...] the specific instances" (Zsiga, 2022, 234). Hence, exemplars and labels are not hierarchically organised through an abstraction from the former to the latter but exist side by side, which constitutes a decisive difference between exemplars and the traditional notion of the phoneme. Furthermore, the variety of exemplar clouds on different levels of the grammar require a multidimensional storage space.

Since the network is continuously updated, the centre of gravity of a cluster can change, which allows exemplar theory to account for linguistic change. A case in point for such a change is GOOSE[3] fronting, whereby the rounded high-back vowel moves towards a more centralised position in the vowel space (e.g. Baranowski, 2017, 301). At the initial stage of the sound change, there will be comparably few instances of centralised GOOSE tokens, which would be located in the periphery of the exemplar clouds. If an exemplar cloud is conceived of as a normal distribution, centralised [ʉ] would occupy the tail(s) of the distribution while back [u] would cover the larger part of the distribution around the mean. As centralised GOOSE tokens become more common, the centre of gravity of the exemplar clouds shifts towards exemplars with centralised GOOSE. Thus, exemplar theory accounts very well for the gradient nature of

[3] The keywords written in small caps throughout this thesis represent Wells's (1982) lexical sets. These lexical sets are "groups of words which share the occurrence of certain vowel categories in stressed syllables in citation form productions by speakers of [...] RP and [...] GA" (Kraus, 2017, 195). They are frequently used in sociolinguistic research to describe vocalic variation across varieties.

many sound changes and potential frequency effects (Bybee, 2017) as more frequent words are activated more easily.

Since linguistic exemplars also index social information, language change can not only be modelled in terms of lexical frequency but also with regards to constraining social factors. As will be shown in more detail below, the merger of NEAR and SQUARE and the merger of KIT and DRESS before nasals is more common in the speech of young New Zealanders and old Anglo Houstonians, respectively. Discrimination between the vowels in question is mediated by perceived speaker age (Hay, Warren, & Drager, 2006; Koops, Gentry, & Pantos, 2008). For instance, exemplars of NEAR and SQUARE with similar vowel qualities are indexed with the social information 'young'. Concurrently, frequency effects can be at work in that merged tokens form the centre of gravity in younger speakers' exemplar clouds while they occupy the periphery in older speakers. Indexing also allows exemplar-based models to explain sound changes that occur relatively independently of frequency. Thus, both linguistic and social cues can alter the distribution of variants within exemplar clouds.

While an abstractionist model frequently involves the extraction of features from the acoustic signal (see subsection 2.3.1 for specific models), speech processing within episodic accounts is usually conceptualised as a matching process of incoming with stored utterances. The generic term *utterances* is chosen here as several levels of the grammar (segments, words, phrases, etc.) might be used (simultaneously) in the matching process. An illustration for this matching process has been provided above for Figure 2.1. Those exemplars are activated that the "incoming speech [...] is most similar to, and perception is biased towards the activated exemplars" (Drager & Kirtley, 2016, 5-6). Thus, recently activated exemplars are more likely to be activated again if they are (partly) congruent with the new input. Within this matching process, linguistic and indexical pieces of information are again closely linked. For example, if an exemplar is activated that is associated with a particular group of speakers, social information associated with this group is also activated and can, in turn, bias the processing of further incoming signal.

Now that the building blocks of exemplar-based models of speech processing have been established, experimental evidence that challenges the assumptions of abstractionist models and supports exemplar-based models will be presented. The evidence will mostly argue for the close link between linguistic and indexical information. Put differently, these findings suggest that indexical information is not filtered out but used as an integral part of speech processing. For now, the focus will remain on individual segments. However, similar findings regarding

lexical access will be summarised in subsection 2.3.3, where the focus is on models of lexical processing.

One seminal study was conducted by Niedzielski (1999), who asked participants from Detroit to focus on a specific word in an auditorily presented sentence and to then identify the vowel in that word by selecting it from a list of six synthesised vowel tokens. The words were recorded with the same female speaker from Detroit, who raised the onset of MOUTH to [əʊ], a feature that is part of Canadian raising (e.g. Boberg, 2008). Canadian raising is stereotypically associated with Canadian English but is common in many accents of North America. Alleged speaker origin was manipulated through the labels *Michigan* or *Canadian* on the test pages. Thus, the incoming speech signal was identical but the listeners thought that the speaker was either Canadian or a fellow Michigander. The choice of vowel tokens varied systematically with alleged speaker origin. Participants chose synthesised vowels that displayed Canadian Raising when they thought the speaker was from Canada. However, synthesised vowels that followed General American (GA) pronunciation norms were chosen when the speaker was presented as coming from Michigan. Since the acoustic signal was the same for all listeners, the information of speaker origin must have influenced the perception of the vowel and, more importantly, the association of raised [əʊ] with Canadian English must be cognitively stored. If the vowel [əʊ] were simply mapped on to the phoneme /əʊ/ through formant ratio or estimated vocal tract length, the participants should have chosen the same synthesised tokens, regardless of speaker origin.

In Hay and Drager (2010), New Zealanders performed the experimental tasks either with toy kangaroos and koalas or toy kiwis present in the room. These stuffed toys were supposed to prime an Australian or a New Zealand context, respectively. All participants listened to a male speaker from New Zealand and identified his vowels on a continuum, which ranged from productions typically found in Australia to typical New Zealand English (NZE) productions. Hay and Drager's (2010) study made use of the differences in front vowels between Australian English and NZE. For instance, KIT typically has a lower and more centralised position in the vowel space in NZE. Compared to Niedzielski (1999), the priming of the alleged origin of the speaker was much more subtle here. Nevertheless, participants chose synthesised tokens that were in line with the national variety primed by the stuffed toys. Again, this finding suggests that speaker information mediates the perception of speech segments and that indexical information is stored in memory (see Strand, 1999 for a shift in the perception of /s/ versus /ʃ/ mediated by alleged speaker sex). It further shows that the activation of indexical information

feeds forward into linguistic processing. In both Niedzielski (1999) and Hay and Drager (2010), indexical information is not primarily manipulated via the auditory signal but through the visual paradigm (label on test pages and stuffed animals). The activation of the corresponding social exemplars then impacts segmental perception, which can be modelled through a lowered activation level of linguistic exemplars that are associated with the social ones (e.g. raising of /aʊ/ to [əʊ] in Canadian English).

The influence of indexical information on speech processing has also been investigated for accents where the vowels of two lexical sets are merging in specific age groups. New Zealanders are better at discriminating the NEAR and SQUARE vowels when tokens are presented with an image of an older speaker (Hay et al., 2006). Since the merger is primarily found in younger speakers, this finding suggests that speaker-group specific information is stored and used in speech processing. Specifically, listeners can distinguish the vowels more easily when primed with the image of an older speaker because older speakers distinguish NEAR and SQUARE more clearly.

Similarly, Houstonians discriminate the KIT and DRESS vowels before nasals more accurately when they see the picture of a younger Anglo Houstonian, who is less likely to merge these vowels than an older Anglo Houstonian (Koops et al., 2008). Finally, the McGurk effect is reduced if listeners are familiar with the speaker's voice (S. Walker, Bruce, & O'Malley, 1995). All of these studies suggest that the perception of segments is influenced by information that is available about the speaker, and by extension the speaker group that they belong to (older/younger speakers, dialect, etc.).

Research into speech production further highlights the importance of indexical information. As shown in subsection 2.2.1, speaker normalisation mainly focuses on eliminating variation that is due to anatomical differences between, for example, male and female speakers. However, research suggests that "anatomical differences are not the exclusive source of differences between men's and women's vowel spaces [...] [and] that talkers differ from each other in [...] ways that can not be predicted from vocal tract anatomy differences alone" (Johnson, 2008, 376). Perry, Ohde, and Ashmead (2001) demonstrated that age and body size, which correlate with vocal tract length, predicted formant frequency differences in children's speech before puberty well. However, gender also emerged as a significant factor, independently of age and body size, which indicates that there is a behavioural or stylistic component to male versus female speech before the development of major anatomical differences during puberty (Johnson, 2008, 378-379). Furthermore, there are cross-linguistic gender differences

in formant frequencies, which cannot fully be accounted for by anatomical differences between the speakers (Bladon, Henton, & Pickering, 1984; Johnson, 2008, 379-383). For example, the F2 difference between female and male after speaker normalisation was much higher in Russian than it was in Danish (Bladon et al., 1984). This suggests that speakers use certain speech patterns to create gender identities beyond anatomical differences. The patterns that encode gender identities varied cross-linguistically. This renders speaker normalisation based on anatomical differences in vocal tract length insufficient. Perceived vocal tract length is still a valuable cue in speech processing but the performance of gender must also be accounted for.

Finally, there is an argument to be made for the informativeness of variation within and between speakers, which has been evidenced in decades of sociolinguistic research. As Pierrehumbert (2016, 40-42) summarises, phonetic variation correlates with stress and the phrasal context, the frequency of a word and its predictability in context, the dialect of a speaker, speaker sex, age, ethnicity and the social network or community of practice that a speaker is part of, to name but a few examples. Speech patterns can, for example, encode group membership and different levels of formality, which, in turn, are highly relevant for non-linguistic behaviour. Therefore, it is sensible to assume that information on this variation is retained in long-term memory and retrieved during speech perception and production. Episodic or exemplar-based models show how indexical information can be stored alongside referential information.

While exemplar-based models account well for the influence of indexical information on speech processing, they are not without shortcomings. Important limitations of exemplar-based models are summarised by Docherty and Foulkes (2014, 48-53). One limitation has to do with the hypervariability of the individual experience as opposed to the relative cohesion of speech community norms. Individual members of a speech community can be expected to receive different input and perform different identities. Exemplar-based models would then predict different configurations of exemplar clouds in these individuals. However, despite these different configurations, successful communication between individuals is possible, phonotactic intuitions are shared and novel words are produced similarly (Docherty & Foulkes, 2014, 49). How do differential individual experiences result in the shared phonological system of a speech community? Further research that focuses on the individual is needed to shed light on this question. Three other limitations refer to underspecifications in exemplar-based models of speech processing. First, it remains to be shown how exactly social-indexical and lexical-

propositional information are integrated in memory. What is the amount of input necessary to link a phonetic pattern with a social or contextual cue? How does the system deal with socially/contextually-conditioned phonetic patterns that are more or less variable (Docherty & Foulkes, 2014, 48)? Second, exemplars have been modelled at different levels, for example the segmental and the word level. Which unit(s) of representation should be used for exemplar-based models? Related to this issue is the question which data type an exemplar corresponds to. Is it a waveform of the signal or a spectrogram? How is the transformation of the raw signal during the auditory process taken into account (Docherty & Foulkes, 2014, 50)? The final limitation concerns the long-term storage of exemplars, which is referred to as the "head-filling-up problem" in Johnson (1997, 146). One of the advantages of abstractionist theories is that abstract units are very efficient with regards to memory space. A finite set of stored phonemes, for instance, can be used to process an infinite number of phonetic events. If memory representations are detailed, however, as predicted by exemplar-based models, how are these representations stored efficiently and how do they change and/or fade over time? This question ties in with the previous concern regarding the units of representation.

2.3 Models of Lexical Access

2.3.1 Abstractionist Models

As described in subsection 2.2.2, exemplar theory has found its way into phonetics and phonology and, by extension, models of lexical access only within the last thirty years or so. As a result, the traditional models of lexical access, which will be reviewed in this section, follow the abstractionist agenda of stripping away from the signal any information that is not strictly linguistic. When reviewing these models, it is essential to keep the historical perspective in mind. The models aimed to provide an initial framework for lexical processing. Many of the studies that argue for the consideration of exemplar theory (see subsection 2.3.3) were conducted later on. Here, TRACE (McClelland & Elman, 1986) and COHORT (Marslen-Wilson, 1987) are reviewed first because they constitute classic models of lexical access. Next, an overview of the model by Stevens (2008) demonstrates the ongoing importance of the abstractionist agenda in models of lexical access.

2.3.1.1 TRACE (McClelland & Elman, 1986)

TRACE aims to explain both spoken and written word recognition and, as shown in Figure 2.2, is based on features and phonemes.[4] The focus of this review will be on spoken word recognition, for which the perceptual system extracts phonetic features from the acoustic signal, which "activate all compatible phonemes, which activate all compatible words" (Zsiga, 2022, 232). Several phonemes and words may be activated concurrently but their level of activation depends on their match with the available information and will decrease if new incoming information, for example a following vowel, makes the word form less plausible (Traxler, 2012, 105). Thus, activation is a limited resource which the potential candidates compete for.

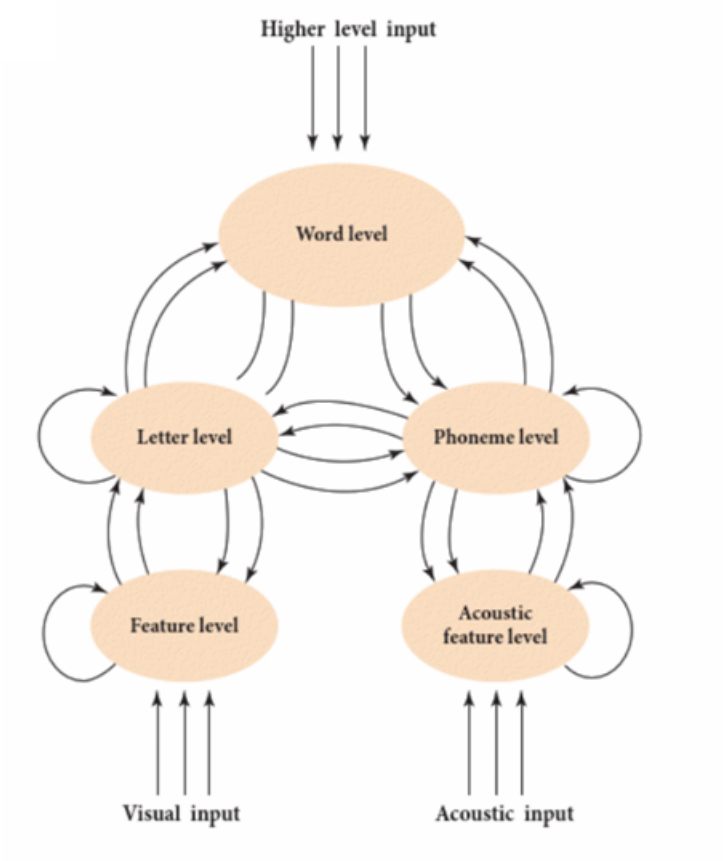


Figure 2.2: TRACE model of lexical access (Traxler, 2012, 106)

Importantly, vertically and horizontally adjacent tiers (except for the feature and the acoustic feature levels) are interconnected. Consequently, top-down processing can be used to aid word recognition. A good example for this is the Ganong (1980) effect. If listeners are presented with a spoken version of *desk*, where the initial plosive is an ambiguous token between

[4] See Kazanina, Bowers, and Idsardi (2018) for an in-depth discussion on the relevance of the phoneme for speech processing.

[d] and [t], they reliably identify the word as *desk* rather than **tesk*. In this case, lexical information is used in a top-down manner to aid the identification of the initial consonant. Models that only involve bottom-up processing would struggle to explain this effect as they do not allow for the flow of information from higher to lower levels of processing. In TRACE, the lexical or word level cascades downwards to guide the identification of phonemes and the extraction of acoustic features directly (see Traxler, 2012, 106-107 for an example of a similar effect in visual recognition). In general, its interactionist structure allows TRACE to account for many effects of top-down processing (e.g. phonemic restoration, R. M. Warren, 1970; see also Samuel, 2011, 6-7). Its shortcomings are assessed in comparison with COHORT, which will be presented next.

2.3.1.2 COHORT (Marslen-Wilson, 1987)

COHORT focuses exclusively on spoken word recognition and distinguishes three processes that are used to achieve this goal: activation/contact, selection and integration (Traxler, 2012, 108). Activation or contact denotes a bottom-up process, during which stored word forms are activated and entered into the cohort if their phonemes match the acoustic input, independently of the degree to which they fit the semantic context. During selection, that word form is selected which constitutes the best match to the input signal and the context. At this stage, the model becomes interactive because top-down processing is employed alongside bottom-up processing to select the best candidate among the initially activated cohort (Traxler, 2012, 108). Finally, the selected word from is merged into the ongoing utterance representation during integration.

COHORT emphasises the incremental nature of lexical access in that each word has a recognition point at which it becomes unique and which influences lexical decision latencies. Strings of segments can be identified as words or nonwords more quickly if they have an earlier recognition point (Marslen-Wilson, 1987, 80-84). Likewise, the semantic meaning of a word is activated faster if it has an earlier recognition point. Experimental evidence for this incrementality and the corresponding pattern of semantic activation comes from priming studies in that onset-matching nonword primes generate stronger priming effects than offset-matching nonword primes. Consider, for example, the target *stripes* and the hypothetical prime word *tiger*. However, instead of *tiger*, the two nonwords **tighem* and **prismer* are presented as primes. Stronger priming effects are observed for **tighem* because it has the same onset as *tiger*, which indicates that (non)words are processed incrementally (Marslen-Wilson &

Zwitserlood, 1989). This effect is not predicted by TRACE, where similarity alone is relevant, regardless of where the similarity occurs (Traxler, 2012, 110). Thus, according to TRACE, the priming effects should be equally strong, regardless of whether the nonwords are congruent with *tiger* in terms of their onset or coda. In addition, context cannot suppress the initial activation of words with phonologically identical beginnings. Thus, when *captain* was used in a semantic context that was strongly biased towards it, cross-modal priming experiments by Marslen-Wilson and Zwitserlood (1989) showed that *captive* was also activated due to the identical *capt-*. However, the priming of *captive* faded when the time interval between the presentation of the prime and the target was increased.

Contrary to TRACE, COHORT does not predict competition for activation between the candidates in the cohort. Accordingly, the size of the cohort does not influence lexical decision latencies (Marslen-Wilson, 1987, 86-87). COHORT further predicts that words can be accessed before they are fully perceived if their unique recognition point is located before the end of the word. Context can also help to speed up word identification. Later revisions of COHORT altered the model by incorporating word frequency and disregarding phonemes. It was theorised that more frequent words are activated more easily and that phonetic features are directly linked to the stored word forms, which made the intermediate identification of phonemes redundant. The latter allowed the model to account for the effects of subphonemic variation. For example, the sequence [pɒt] is generally produced faster in *pottery* versus *pot* (Traxler, 2012, 112) because of its use within a multisyllabic word in the first instance. Such durational differences are used to identify the right word faster.

2.3.1.3 Stevens (2008)

Stevens's (2008) model demonstrates how the formant ratio and vocal tract length theories presented in subsection 2.2.1 are integrated into models of lexical access. Stevens's (2008) model is primarily based on distinctive features, such as [±continuant]. As shown in Figure 2.3, it consists of two general procedures, a bottom-up and a top-down operation. The bottom-up operation involves the identification of acoustic landmarks and parameters in the incoming signal. These cues are used to extract distinctive feature bundles, which can then be compared to the word forms stored in the mental lexicon. Other cues, such as "syntactic and semantic knowledge and visual cues from the speaker's face" (Stevens, 2008, 143), facilitate this process. Importantly, these visual cues heavily focus on the identification of articulatory movement and do not refer to social-indexical cues.

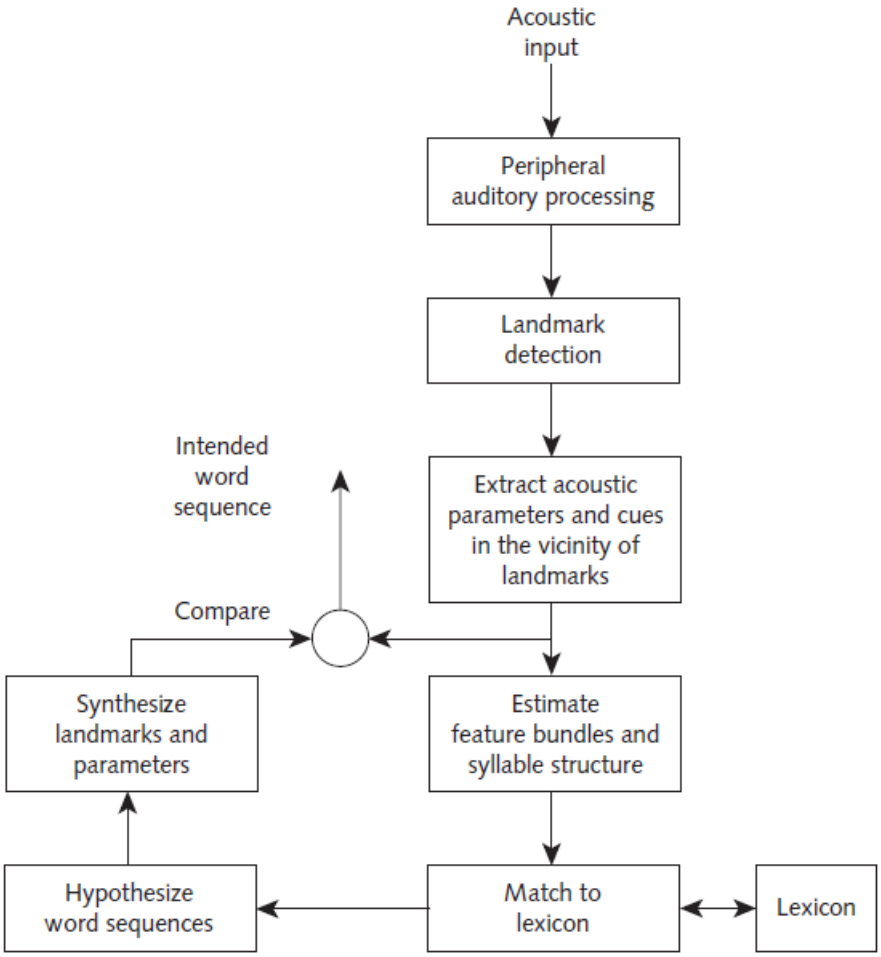


Figure 2.3: Feature-based model of lexical access (Stevens, 2008, 144)

The top-down operation starts at a later stage with the hypothesised word sequence that was extracted from the signal. Based on this word sequence, the perceptual system performs "an internal synthesis of landmarks and parameters" (Stevens, 2008, 143) that are associated with it. Since the word sequence consists of more than just individual sounds, a precise pattern of landmarks and parameters, including transitions between the segments, can be calculated based on stored representations. The output of this synthesis is then compared with the landmarks and parameters extracted from the signal to corroborate the word sequence intended by the speaker.

Tracing the individual steps of Stevens's (2008, 145) model shows that the perceptual process starts with peripheral auditory processing, which is identical for speech and non-speech sounds and involves the transformation of "the sound wave [...] into mechanical action and then patterns of electrical activity". During this step, those acoustic characteristics of the signal are enhanced which allow for an easier subsequent recognition of distinctive features.

Next, acoustic landmarks are identified to demarcate vowels, glides and different types of consonants. During the following step, these landmarks are used as guiding points for the extraction of acoustic parameters and cues. Stevens (2008, 146) assumes that "there is a universal set of such parameters" and that "[t]he selection of parameters [...] is motivated by the need to provide information about the articulatory gestures that generated the speech pattern [...]". The latter assumption is central to motor theories of speech processing (Samuel, 2011, 2-4).

Which parameters are extracted depends on the nature of the landmarks. For stops followed by vowels, for example, formant transitions are a decisive cue. The primary articulatory gestures to be reconstructed from the parameters are the shape of the vocal tract and the presence or absence of vocal fold vibration. In addition to segmental information, "[t]he parameters and the cues derived from the parameters also provide information about the syllable structure and other prosodic aspects of an utterance" (Stevens, 2008, 149). Next, the distinctive features of the segments are identified, primarily based on the extracted acoustic parameters and supported by contextual, durational, syllabic and prosodic information. Each feature is accompanied by a "measure of confidence" (Stevens, 2008, 150), which is influenced, for example, by the level of ambient noise in the speech signal. Bundles of distinctive features, along with "additional information concerning syllable structure, possible word boundaries, phrase boundaries, and syllable prominence" (Stevens, 2008, 151), are, finally, matched with the lexical representations stored in the mental lexicon. If available, this matching process is facilitated by visual information and the syntactic and semantic context. Importantly, Stevens's (2008, 151-152) model does not require the phoneme as a unit of representation. Entries in the mental lexicon consist of a syllable structure, in which each slot is filled with a combination of distinctive features.

The main focus of Stevens's (2008) model is the bottom-up operation, which yields a cohort of words or word sequences, which, in turn, are used as the input for the top-down operation. Based on the words or word sequences, the model first reconstructs the articulatory gestures needed to produce them and then synthesises the acoustic patterns associated with these gestures. The synthesised acoustic patterns of each word or word sequence are compared to the acoustic parameters and landmarks extracted from the signal (Stevens, 2008, 152; see circle in the centre of Figure 2.3). That word or word sequence is selected which provides the best fit between the synthesised and the actual parameters and landmarks. As touched on above, the internal synthesis benefits from using an entire word or word sequence as its input.

During the bottom-up operation, certain information, for example the location of phrasal stress, might not yet be available when landmarks are identified and parameters are extracted because it occurs later on in the utterance (Stevens, 2008, 152). The hypothesised word sequence, on the other hand, can account for the preceding and following context for each segment when synthesizing acoustic landmarks and parameters.

Notably, the top-down operation does not affect the bottom-up one directly in that the initial extraction of parameters is entirely guided by the signal. The internally synthesised acoustic patterns are matched with those identified in the signal but this matching process occurs towards the end when bottom-up processing is practically finished. Therefore, Stevens's (2008) model is interactive in that both bottom-up and top-down processing are used but not to the extent that stored lexical information directly influences the processing of the signal. This aspect of interactivity is incorporated in TRACE and COHORT. Note, however, that Stevens's (2008) measure of confidence during the estimation of features allows the model to explain, for example, the Ganong (1980) effect. Considering again the example of an ambiguous plosive between [t] and [d] before the sequence [ɛsk], the distinctive feature [±voicing] would be coded with a lower confidence measure and the internal synthesis of *desk* would enable the identification of the word. Predictions of Stevens's (2008) and TRACE could differ in terms of the speed of lexical access though as the lexical match occurs at an earlier stage in TRACE.

2.3.2 Abstraction in TRACE, COHORT and Stevens (2008)

Before more evidence is presented that argues for the integral role of indexical information on lexical access, this subsection summarises the two ways in which the models from the previous subsections are considered abstract. First, the processing of the acoustic signal is designed to generalise over intra- and inter-speaker variation. In all three models, the ultimate goal is to match aspects of the incoming signal with stored abstract forms. While the models differ in terms of which acoustic aspects are extracted from the signal and whether or not phonemes are assumed as an intermediate unit of processing, they all predict that surface variation must be filtered at some point. This is not to say that contextual factors are not considered by the model. The selection of candidates in COHORT is partly driven by context and, in Stevens's (2008) model, visual information can become relevant when the lexicon is accessed. However, these steps are preceded in all three models by a filtering of the signal. In other words, all models assume that the signal is mapped on to an abstract unit first, be it a

feature or a phoneme, before context becomes relevant. This implies that indexical information is, at least partly, lost and, more importantly, not retained in memory for speech processing.

The second aspect of abstraction is the storage of words in the mental lexicon. TRACE and the early version of COHORT explicitly include phonemes as units of processing, which group together "speech events that occurred at different times and places, and whose physical properties are objectively different" (Pierrehumbert, 2016, 37). As these abstract phonemes are combined into word forms, variation on the phonetic surface is not assumed to be retained in long-term memory. Stevens's (2008) model and the revised COHORT do not require phonemes and focus on distinctive features and word forms instead. As a result, in each model the entries in the mental lexicon include at least one level of abstraction away from the original phonetic surface. Representations in long-term memory consist of abstract units that generalise over the acoustic variation in the speech signal.

2.3.3 Exemplar-Based Models

Contrary to abstractionist models, exemplar theory argues that information from the phonetic surface is retained in long-term memory and facilitates the identification of segments and lexical processing. Lexical access in exemplar-based models is based on the basic matching and categorisation mechanism shown in Figure 2.1. If exemplars are modelled at the word level, the incoming signal activates different stored words/exemplars and the exemplar with the overall highest activation is accessed. As pointed out in subsection 2.2.2, there is still uncertainty as to whether exemplars are stored as words and/or other units. Regardless of this underspecification, the following studies suggest that the stored lexical information should contain more than abstract units of representation.

Goldinger's (1996) seminal study showed that memory performance is generally better if the words in the test phase and in the study phase are produced by the same speaker, an effect that is evident even when the two phases are one week apart. This suggests that speaker-specific productions of a word are retained in memory, which contradicts the idea of a speaker normalisation before any information reaches the mental lexicon. The importance of speaker familiarity was demonstrated under adverse listening conditions by Nygaard and Pisoni (1998). They familiarised their participants with different voices during several training sessions by having them allocate word-long utterances to the respective speaker. During test, the participants were better at transcribing new words in noise than the control group, the members of which had not been familiarised with the voices beforehand. Sentence-long

utterances during familiarisation did not improve word transcription in noise. However, it did increase accuracy when participants transcribed sentence-long utterances in the test phase instead. This suggests that exemplars could (additionally) be stored at the sentence level, encompassing information that is not available at the word level (e.g. phrasal and sentence stress). Such information could then facilitate the transciption of sentences in noise.

Focusing again on group characteristics rather than specific speakers, A. Walker and Hay (2011) had their participants listen to older and younger speakers producing words that were more frequent either in the speech of younger New Zealanders (e.g. *lifestyle*) or in the speech of older New Zealanders (e.g. *confectionary*) or age-neutral (e.g. *driver*). Responses in lexical decision tasks were faster if voice and age-specific word frequency matched, that is if words that are more common in young speech were presented in the voice of a younger speaker and vice versa. In addition to the categorical difference between speaker groups (e.g. old versus young or male versus female), there appears to be a gradient differentiation within groups in terms of prototypicality. Evidence for this comes from Strand (2000), who found that participants are faster at shadowing a word if the speakers' voice is more stereotypically male or female. Notably and in contrast with A. Walker and Hay (2011), the words in her study were not more frequent in male or female speech. Instead, the effect appears to be due to gender-prototypical pronunciations, which again suggests the retention of surface-level information in memory.

An important observation to make here is that exemplar-based models of lexical access provide a less definite framework for lexical processing than the abstractionist models reviewed in subsection 2.3.1. Stevens's (2008) model, for example, can be broken down into several steps and visualised accordingly (see Figure 2.3). Exemplar-based models of lexical access, on the other hand, seem to rely on general guiding principles (e.g. the storage of detailed representations) without strong claims as to how lexical processing is actually implemented. As mentioned above, this was raised as one of the shortcomings of exemplar theory by Docherty and Foulkes (2014). Many aspects of lexical processing remain underspecified and there is a lack of a consistent framework.

2.3.4 Summary: Abstractionist versus Exemplar-based Models

Abstractionist and exemplar-based models differ in terms of how the incoming signal is processed and how the processed information is retained in long-term memory. Strong abstractionist accounts argue that intra- and inter-speaker variation must be filtered out so that the signal can be mapped on to abstract units of representation. While abstractionist models

of lexical access, such as TRACE (McClelland & Elman, 1986), COHORT (Marslen-Wilson, 1987) and the model proposed by Stevens (2008), might differ with regards to specific units of representation, they share the assumption that these units are abstract and that the mapping of the signal on to the units occurs by filtering surface-level phonetic variation. Strong versions of exemplar theory, on the other hand, stress that any generalisation must be processed online and is not retained in long-term memory (Zsiga, 2022, 234). Both types of theories have explanatory power. Abstraction offers a good explanation for why (non)words can be perceived and produced relatively easily although they have not been encountered before. Exemplar theory offers a framework to justify the effects of indexical information on speech perception and production, such as the differential perception of segments based on speaker identity (e.g. Hay et al., 2006; Koops et al., 2008; Niedzielski, 1999) or the linguistic performance of gender beyond anatomical differences (Bladon et al., 1984).

Importantly, abstraction and episodic memory are not mutually exclusive. In fact, Pierrehumbert (2016, 33) explicitly states that "a hybrid model of phonological representation is needed". Although it is not part of the abstractionist models of lexical access presented so far, it is possible to assume a lexicon-external speaker model that informs meaning retrieval (see e.g. Cai et al., 2017). The entries in the mental lexicon would still be abstract but indexical information could be used to disambiguate homophones and/or polysemous words. Thus, from an abstractionist point of view, episodic information might facilitate lexical access but is still stored outside the mental lexicon in a separate storage unit. The question then becomes if speech processing is possible in an exclusively abstractionist manner, without the activation of this episodic information. Conversely, abstraction is included in many exemplar-based models through the conceptualisation of labels for exemplar clouds. The main difference remains though that abstraction in exemplar-based models is not a hierarchical process. The labels for exemplar clouds exist alongside the individual exemplars on the same level of representation (see subsection 2.2.2).

The idea of a hybrid model is further explored below by reviewing Sumner et al.'s (2013) account on socially weighted encoding and speech processing. There are several hybrid models of speech processing, some of which argue for a lexicon-external storage of indexical information (e.g. Cai et al., 2017), while others state that linguistic and indexical units are stored together. What they have in common is that they include different pathways or processing routes, to which linguistic and indexical information are subjected. Sumner et al. (2013) is an example of a model that suggests the common storage of linguistic and indexical information. It

was chosen here to illustrate hybrid models because it can account well for the experimental evidence presented so far.

2.4 Dual-Route Models Exemplified by Sumner et al. (2013)

Sumner et al. (2013) propose a dual-route model which argues that the encoding and perception of exemplars are socially weighted. The model has a linguistic and a socioacoustic pathway. Along the linguistic pathway, the incoming signal activates congruent exemplars and, in line with the predictions from subsection 2.2.2, the exemplar with the highest level of activation is selected (Sumner et al., 2013, 2, 7). Exemplars are assumed on both the word and the sub-lexical level, which allows the model to account for differences in the perception of individual segments (e.g. Niedzielski, 1999). The socioacoustic pathway of the model is used to analyse indexical information and social features embedded in the speech signal. Sumner et al. (2013, 7) argue that the encoding of exemplars along the linguistic pathway is constrained by these social features, such as the dialect of the speaker. As will be shown in more detail below, stronger representations will be built within the high-dimensional lexical space if the speaker has, for example, a standard dialect. Socio-indexical and linguistic information is stored together.

Sumner et al.'s (2013, 3) framework is based on two sets of findings, which they refer to as "[r]ecognition [e]quivalence and [m]emory [i]nequality". Recognition equivalence concerns findings that suggest that phonetically different exemplars of the same word are recognised and accessed equally fast although they vary substantially in token frequency. Sumner (2013), for example, conducted a semantic priming task, during which experimental tokens with word-medial *-nt-* were used. In AE, this sequence is frequently reduced from [nt] to [n] in words such as *splinter*, *winter* or *centre* in casual speech. Three different types of *-nt-* tokens were used: carefully produced [nt], carefully produced [n] and casually produced [n]. The latter two types differed in duration, with [n] in careful speech being, on average, 220 ms longer than [n] in casual speech (Sumner, 2013, 28). The American participants first heard a probe with word-medial *-nt-*, then saw an image of that probe and finally had to decide as quickly and accurately as possible if the following target word on their monitor was a real word or a pseudoword. Importantly, the target for the experimentally relevant trials was either semantically related to the probe (e.g. *centre-town*) or not. Latencies for that final decision show how quickly the lexical meaning of the probe was accessed since the probe primed the target semantically.

Crucially, participants' lexical decision latencies did not differ significantly between the probes *spli*[nt]*er* and (casual) *spli*[n]*er* (Sumner, 2013, 29-30). This suggests that careful *spli*[nt]*er* and casual *spli*[n]*er* result in similar semantic priming effects although the former occurs less frequently than the latter. There was a processing cost for careful *spli*[n]*er*, which, as Sumner (2013, 30), argues could be due to a mismatch between the careful speech style and the use of a variant typically found in casual speech. As shown above, this has durational repercussions in that careful productions of the [n] variant were longer in duration than casual productions of [n]. In fact, Sumner et al. (2013, 3-4) contend that the latter is problematic in other studies which did not find recognition equivalence between casually and carefully produced tokens but a recognition advantage for the carefully produced or standard token (e.g. Andruski, Blumstein, & Burton, 1994). These studies employ splicing procedures which are intended to maintain ecological validity but actually result in a mismatch between word frame (careful) and the phonetic variant that is inserted into it (casual). The decisive question here is why lexical decision latencies in Sumner (2013) were equivalent for casually and carefully produced probes despite the much higher frequency of casually produced tokens.

Memory inequality refers to findings indicating "that words with infrequent, but idealized variants are remembered better than words with frequent, non-idealized variants" (Sumner et al., 2013, 4). A case in point for memory inequality is Experiment 3 in Sumner and Samuel (2009). Sumner and Samuel (2009, 495) used a "long-term priming paradigm [...], [during which] participants hear critical items and fillers presented in two blocks, with critical items in the first block ultimately acting as primes for targets presented in the second block". The primes and targets used in the experiment varied with regards to rhoticity, which is a common feature in many varieties of American English (AE) but typically absent in New York City (NYC) English (Sumner & Samuel, 2009, 489). Sumner and Samuel (2009) investigated three groups of participants: first, listeners who were rhotic and less familiar with the non-rhotic NYC accent (AE); second, listeners who were rhotic but born and raised in NYC (Covert-NY) and; third, listeners who were non-rhotic and born and raised in NYC (Overt-NY).

The participants performed speeded lexical decisions in both blocks, with critical items in the first block either produced by a non-rhotic female speaker from NYC (e.g. *baker* [beɪkə]) or a female speaker with a GA accent (e.g. *baker* [beɪkɚ]). These critical items were repeated 20 to 30 minutes later in the second block, during which they were produced by a male speaker who was either from NYC (non-rhotic) or a GA speaker (rhotic). With this design, Sumner and Samuel (2009) intended to find out if the (non-)rhotic tokens in the first block facilitated

lexical decisions for the targets in the second block and how this was mediated by the listeners' own accent and origin. In all three listener groups, they found stronger facilitation effects for words that were presented with r-full pronunciations ([beɪkɚ] instead of [beɪkə]) in both experimental blocks. This is sensible for the AE and Covert-NYC group because they used rhotic pronunciations themselves. However, participants in the Overt-NYC group were non-rhotic and yet their decisions were faster and more accurate for rhotic rather than non-rhotic primes and targets (Sumner & Samuel, 2009, 496-498). Sumner and Samuel (2009) further found no priming effect of non-rhotic tokens for the GA listeners.

The finding for the Overt-NY group is difficult to account for within exemplar-based accounts that focus on frequency. Sumner and Samuel's (2009) experiment included both rhotic and non-rhotic tokens for all participants. Therefore, it cannot be argued that social information on speaker origin ('from NYC' or 'speaker of GA') was activated more frequently. Token frequency does not offer a viable solution either. Overt-NYC listeners are non-rhotic and, as such, very likely to have encountered non-rhotic productions of words very often. Thus, their exemplar clouds should be strongly centred around non-rhotic exemplars. However, stronger memory effects were observed over a latency of 20 to 30 minutes for the GA rhotic rather than the non-rhotic tokens. This is what Sumner (2013) refers to as memory inequality. Standard tokens, despite being less frequent, are more strongly encoded due to their social status and, thus, lead to the observed memory effects.[5]

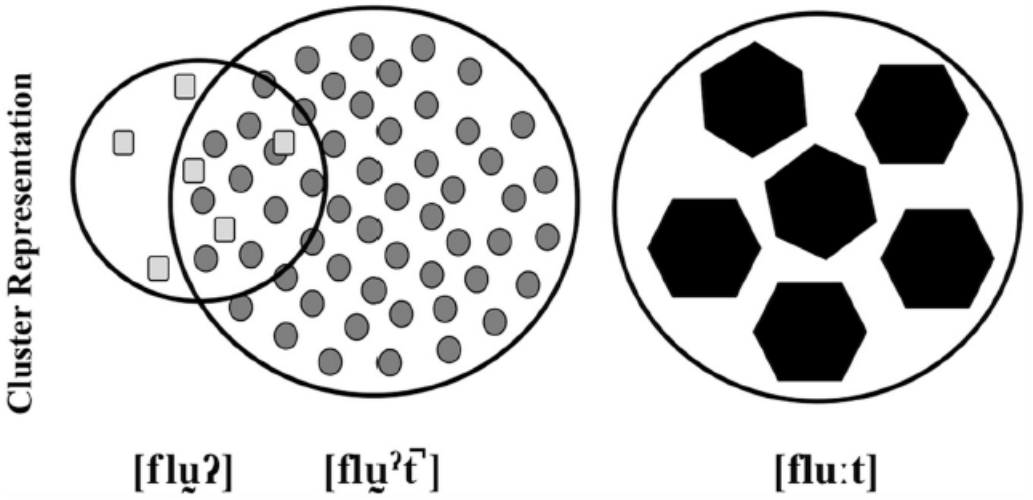


Figure 2.4: Socially-weighted encoding of different variants (Sumner et al., 2013, 6) The left clusters overlap because their exemplars are articulatorily similar.

[5] In addition to this socially weighted encoding, the relatively formal context of a linguistic experiment might make the standard tokens more readily available.

Based on these findings, Sumner et al. (2013, 5-6) theorise that the encoding of tokens as exemplars in the mental lexicon is mediated by frequency and social weight. Thus, three types of tokens can be distinguished: typical tokens normally found in casual speech, atypical idealised tokens found in the standard and atypical non-idealised tokens, which are found infrequently in casual speech. Sumner et al. (2013, 6) present variants of word-final /t/, when it is not part of a consonant cluster, as an example. In AE, the /t/ in *flute* can be realised as an unreleased and coarticulated [flu̚ʔt̚] (typical), as a fully released [fluːt] (atypical idealised), and as a glottal stop [flu̚ʔ] (atypical non-idealised). The [ʔ] variant would be encoded weakly because it cannot profit from high frequency or social status. [ʔt̚] profits from the frequent usage of this variant, which leads to a higher resolution in the exemplar cloud. However, it is relatively weakly encoded because it will, in most situations, lack overt social status (see smaller size of exemplars in Figure 2.4).[6] Fully released [t] might not be frequent but, due to the fact that it occurs in more formal contexts, it receives a lot of social weight. While the encoding resolution is low, the social status of this variant ultimately increases its encoding strength, which can account for the memory inequality effect described above.

[6] It must be acknowledged here that the concept of *status* is very much dynamic and depending on the situation as variants that are not associated with overt prestige might very well receive covert prestige within speech communities. The notion of *standard* or *idealised* tokens in Sumner et al.'s (2013) model, as will be shown below, depends on the overt prestige of the standard language (standard language ideology; see e.g. Lippi-Green, 2011).

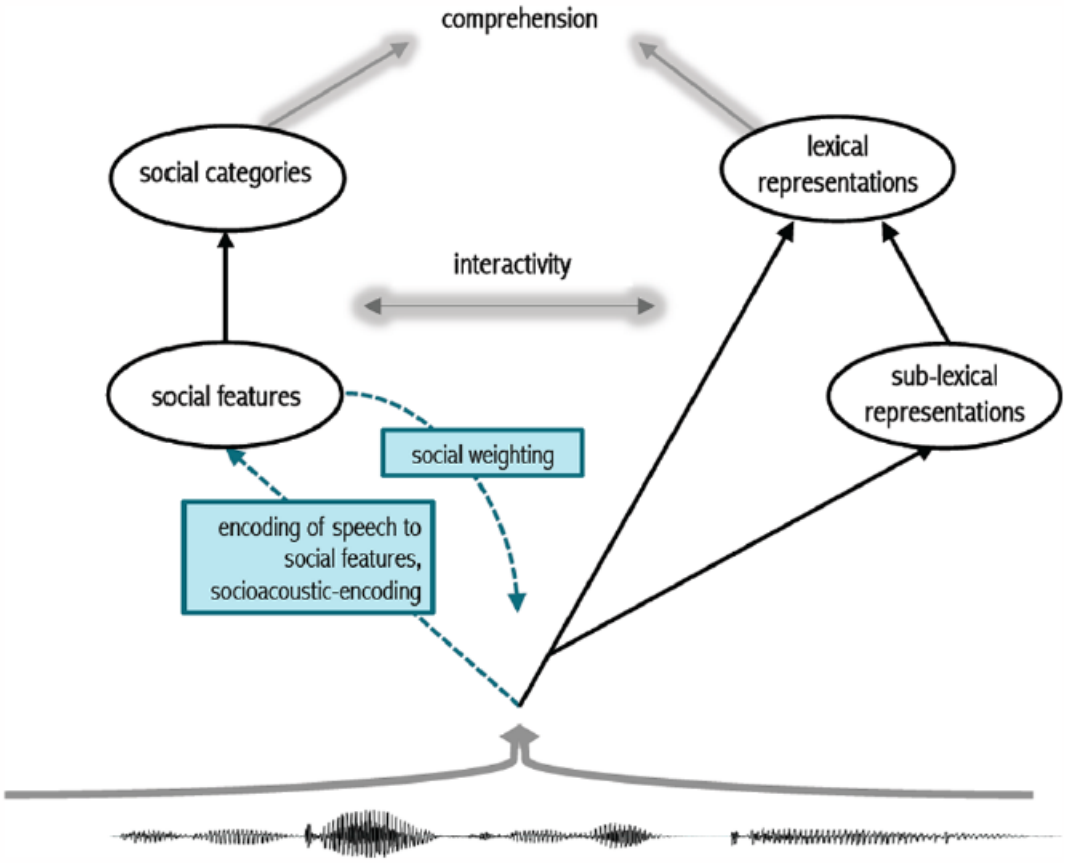


Figure 2.5: Model of socially-weighted encoding (Sumner et al., 2013, 7)

Sumner et al.'s (2013) complete model of socially-weighted encoding and speech processing is shown in Figure 2.5. Within this dual-route model, the acoustic signal is searched for social and linguistic information along the left and the right pathway, respectively. Along the linguistic route, lexical representations are activated either directly or via sub-lexical components. Simultaneously, social features are identified based on "[l]earned, and subsequently stored, acoustic patterns" (e.g. non-rhotic productions as a marker of NYC English; Sumner et al., 2013, 8). The individual social features combine to social categories and have repercussions on the strength of encoding. Through social weighting, standard or ideal tokens are encoded more strongly and, thus, recognised as well as and remembered better than casual tokens although the latter might be much more frequent. As Sumner et al. (2013, 8) point out, "[o]ver the years, orthography, meta-linguistic commentary about standards, and other types of experience [...] compound to contribute to socially salient patterns". Listeners become more adept at identifying socially salient variants and encode them more strongly if they represent the standard. The socioacoustic and the linguistic routes are connected to one another. This suggests that if there is overt information on the speaker available, social categories can be activated, which can guide linguistic processing.

As a result, Sumner et al.'s (2013) model can account well for the experimental evidence reviewed so far. For example, it allows for the differential perception of sub-lexical units based on available indexical information. A case in point is Niedzielski's (1999) research on the perception of MOUTH onsets by listeners from Michigan. If a speaker is presented as Canadian (Niedzielski, 1999), this social feature is fed back into the linguistic pathway and influences the perception of the vowel. The framework of socially weighted encoding is also suitable for A. Walker and Hay's (2011) finding that lexical decisions are faster if words are presented in voices that match the age-specific frequency of these words (e.g. *confectionary* in the voice of an older speaker versus *lifestyle* in the voice of a younger speaker). If a word is heard more often from speakers of a certain age group, the social feature 'age' is identified along the socioacoustic pathway. Since linguistic and indexical information is stored together in Sumner et al.'s (2013) model, it makes sense that lexical access occurs more quickly if the voice that a word is presented in matches the social features associated with this word in the multidimensional mental space.

One shortcoming of the model proposed by Sumner et al. (2013) is that it does not seem to consider the influence of stored lexical representations on the perception of segments. While the influence of indexical information on segmental perception can be traced well, the model does not include top-down processing along the linguistic pathway. The latter has been exemplified by the Ganong effect in subsection 2.3.1.

2.5 Familiarity, Intelligibility and Attitude in Speech Processing

The previous subsections have provided an overview of different approaches to speech processing, ranging from abstractionist models (see subsection 2.2.1) to exemplar-based models, which call for the common storage of detailed linguistic and indexical information in long-term memory (see subsection 2.2.2). Hybrid models have also been exemplified through Sumner et al.'s (2013) model. As mentioned in section 2.1, the agenda of this research is not to argue for a specific model. While hybrid models that allow for abstraction, retention of surface-level information and the close link between linguistic and socio-indexical information are preferred, the main aim is to see how familiarity, intelligibility and attitude affect the processing of different accents. Therefore, previous studies into the effects of these three factors on different levels of processing will now be reviewed. For each factor, an overview of how it has been operationalised will be given first, followed by relevant research into its effect on speech processing.

2.5.1 Familiarity

2.5.1.1 Operationalisation

There are two aspects that make a clear definition of familiarity difficult. Firstly, it has been operationalised differently across past studies. Secondly, and more importantly, it is often combined with manipulations of intelligibility (e.g. Babel & Russell, 2015; Clopper & Bradlow, 2008), which makes it difficult to isolate the effect of familiarity. What the majority of the studies presented here have in common is that they treat familiarity categorically, that is accents are, overall, either 'familiar' or 'unfamiliar' to groups of participants. However, different approaches have been taken to decide if an accent is one or the other. Most studies combine a regional criterion with a language background and demographic questionnaire (Adank & McQueen, 2007; Clopper, 2017; Floccia et al., 2009, 2006; Impe et al., 2009; Martin, Garcia, Potter, Melinger, & Costa, 2016; Scott & Cutler, 1984). In these studies, familiarity is based on the idea that participants are more familiar with local as opposed to non-local accents, which is supported by their responses in the questionnaire. In a different set of studies (e.g. Maye, Aslin, & Tanenhaus, 2008), the participants' unfamiliarity with an accent is guaranteed by using artificial accents. While these studies provide more control for familiarity, they might lack ecological validity in that it is uncertain how results drawn from artificial accents can be extended to natural ones. Finally, there are studies that question the concept of familiarity altogether and, instead, highlight the importance of how similar or distinct two accents are in, for example, their spectral and durational characteristics (e.g. Stringer & Iverson, 2019).

The next subsection provides an overview of studies that consider the effects of familiarity mostly on lexical processing. These studies have been chosen because they consider familiarity without combining it with intelligibility and because they inform the design of the current research.

2.5.1.2 Familiarity and Lexical Processing

Adank and McQueen (2007) used an animacy-decision task to measure processing differences between Dutch accents. They included the Dutch accent spoken in Nijmegen as the familiar accent and East Flemish as the unfamiliar one (Adank & McQueen, 2007, 1926). Their experiment consisted of two test phases with an intermediate exposure phase. During test, participants heard individual words in both accents and had to decide as quickly and accurately as possible if these words were animate or inanimate. During exposure (approxima-

tely 23 minutes), the participants decided if the subject of the auditorily presented sentence was in singular or plural. In a between-participant design, half of the participants listened to the familiar accent while the other half heard the unfamiliar one during exposure. No target word was repeated in the exposure and the two test phases. Adank and McQueen (2007, 1927-1928) found that decisions were more accurate during the second test phase, regardless of which accent they were presented in and which accent the participants were exposed to. Reactions were generally faster for targets produced in the familiar accent even when participants encountered the unfamiliar accent during exposure. The latencies suggest a processing cost for the unfamiliar accent, which cannot be overcome by 23 minutes of exposure to this accent.

Impe et al. (2009) conducted a speeded lexical decision task with students from secondary schools. The targets were presented in eight different regional accents of Dutch from Belgium and the Netherlands as well as the standard variety of both countries. Impe et al. (2009) found an asymmetry in lexical processing between participants from the Netherlands and those from Belgium. Belgian Dutch participants' responses to tokens presented in Netherlandic accents were faster than vice versa. Additionally, reactions were faster for target pronunciations that were closer to the national standard varieties. The results again suggest a processing cost for unfamiliar accents but also highlight the importance of the prestige of certain accents. In this case, Belgian Dutch speakers have "for many years [...] been exposed to and familiarised with the Netherlandic Dutch accent and lexicon, instead of the other way around" (Impe et al., 2009, 112). This exposure increases familiarity and the speed of lexical access. Conversely, the Dutch participants were less familiar with the Belgian standard.

Benefits in terms of lexical access for the standard accent have been demonstrated by Sumner and Samuel (2009) in the American context. As shown in section 2.4, they distinguished rhotic AE participants from participants who grew up in New York and displayed either non-rhotic productions (overt-NYC) or rhotic productions (covert-NYC). They investigated how form and semantic priming for lexical access were mediated by the rhoticity of the targets and the different listener populations. Sumner and Samuel (2009) found the strongest priming effects for r-full tokens, regardless of whether participants were AE, overt-NYC or covert-NYC. In addition, non-rhotic tokens primed targets for the overt- and covert-NYC participants more strongly than for the GA participants. These results again suggest that familiarity with an accent facilitates lexical processing and that there is an asymmetry between regional and standard accents in that rhotic tokens are successful primes even for overt-NYC participants, who

were non-rhotic. For AE participants, on the other hand, non-rhotic tokens were not equally successful primes.

Floccia et al. (2006) and Floccia et al. (2009) investigated the lexical processing of different accents in L1 speakers of French and English, respectively. The series of experiments in both studies used similar stimuli: Disyllabic real or pseudowords were placed as targets at the end of a set of carrier sentences (Floccia et al., 2006, 1279; Floccia et al., 2009, 383). The participants were asked to determine the lexical status of the targets as quickly and accurately as possible.

In Floccia et al.'s (2006) experiments, the stimuli were produced in six different accents (Parisian French, North Eastern French, two Southern French accents, Swiss French and an L1 English accent), which were categorised into three groups ('home', 'familiar', 'unfamiliar'), depending on the participants' origin. The six experiments presented in Floccia et al. (2006) varied with regards to which accents were used, the participants' origin (North East France or Switzerland), the length of the carrier sentences and the presentation of accents in the lexical decision task (blocked or random). When the accents were presented in random order, Floccia et al. (2006, 1288) found that the unfamiliar regional accents resulted in "an initial perturbation in speech processing, which becomes evident after only a certain amount of accented signal has been processed". The processing delay for the unfamiliar accent varied with the length of the carrier sentence and became more pronounced if the critical items were preceded by longer sentences. Importantly, the initial delay associated with the unfamiliar regional accent could be mitigated by sufficient input of that accent. This is evident from the finding that latencies were not significantly different for home and unfamiliar regional accents if the sentences were blocked by accent. In other words, if the accent of the sentences did not change randomly throughout the experiment, latencies converged for the home accent of the participants and the unfamiliar regional accent (Floccia et al., 2006, 1284-1286). In this blocked design, there was also no evidence of an initial processing cost when the accent changed, which Floccia et al. (2006, 1289) attributed to the potentially insufficient power of the design and an insufficient number of observations per condition.

Latencies for the L2 accent remained significantly longer than those for the home and the unfamiliar regional accent, which suggests that lexical processing does not return to baseline levels for the L2 accent even in a blocked design. Based on these findings, Floccia et al. (2006) suggest that the adaptation to unfamiliar regional accents involves an initial processing delay, which can decrease to baseline levels with sufficient exposure. This adaptation effect

was not found in the studies presented so far although Adank and McQueen's (2007) included an exposure phase of more than 20 minutes.

Floccia et al. (2009) further investigated the time course of this adaptation effect and potential surprise effects induced by the task. Their experiments included four different accents: Plymouth English (the home accent of most participants), Irish English, Received Pronunciation (RP) and L2 English (L1 French). Floccia et al. (2009, 389-390) found that latencies increased if a block of sentences in Plymouth English was followed by another block in Irish English or French-accented English. Again, latencies were longest for the L2 accent. Floccia et al. (2009) did not find evidence for an adaptation to Irish English or L2 English back to baseline levels of processing (i.e. the latencies for Plymouth English).[7]

In order to test if the initial processing delay is due to surprise because of a change in accent, Floccia et al. (2009) conducted another experiment, during which they manipulated participants' expectations to as well as their familiarity with the task. The participants first heard a block of sentences in Plymouth English followed by a block in French-accented English. Participants who were unfamiliar with the task displayed a stronger initial processing delay when the accent changed. Importantly, this effect could also be brought about in experienced participants when the surprise effect was renewed "by focussing participants' attention upon accent characteristics" (Floccia et al., 2009, 394).

In their final experiment, Floccia et al. (2009, 395-396) presented the different accents in the experiment randomly with intermittent blocks of four sentences in the same accent. Their intention was to, first, test accent adaptation when changes in accents did not come as a surprise any more and; second, analyse if latencies decreased throughout the blocks of four sentences. They found that latencies were shortest for the home accent, with no significant difference between Irish and French-accented English. There was no evidence for adaptation over the intermittent blocks of four sentences. The difference between the home and the other accents suggest that the processing delay was not solely due to a surprise effect but partly generated by the participants' level of familiarity with an accent. However, it is noteworthy that latencies for the Irish English as opposed to the L1 French speakers were only different in the blocked design of Floccia et al.'s (2009) first experiment.

[7] Floccia et al. (2009, 390) further had a separate group of participants rate the speakers from the study with regards to how 'Plymothian', 'Irish' and 'French' they sounded. They found that these accentedness ratings were lower for the Plymothian versus the Irish/L1 French speakers. Based on this, they tentatively suggest that there could be different processing modes for the local versus non-local accents.

Taken together, the results suggest that lexical processing for unfamiliar accents does not resume to baseline levels, "at least not within the timeframe [. . .] [of] up to fifteen consecutive sentences" (Floccia et al., 2009, 401). Floccia et al. (2006) found a stable difference in the processing of a regional versus L2 accent in French, which could not be replicated for English here. Floccia et al. (2009, 401-402) argue that this is due to the higher proficiency of the L2 speakers used in the 2009 study. Importantly, Floccia et al. (2009) identified a surprise effect that influences speech processing when a new accent is perceived. Methodologically, such a surprise effect could be mitigated by a random presentation of accents throughout the experiment.

Clopper (2017) investigated cross-speaker and cross-dialectal interference effects on the processing of local versus non-local accents of AE. Clopper's (2017) participants, who were all from the Midland region (south of Midwest of United Staes of America), performed a speeded lexical classification task, during which they were asked to identify the acoustic stimulus as one word or another. Throughout the first series of experiments, the word pairs *bed/bad* and *head/had* were chosen because the Northern Cities Vowel shift makes it more likely to confuse these tokens qualitatively,[8] which, in turn, increases the likelihood of interference effects (Clopper, 2017, 28). The stimuli for the first series of experiments were produced by Northern and Midland speakers. This enabled Clopper (2017) to systemically investigate how latencies and the accuracy for each semantic classification were affected by single- versus mixed-talker conditions, the latter of which could include speakers of the same dialect (e.g. two Midland dialect speakers) or speakers of different dialects.

Clopper (2017) found that reactions were generally slower and less accurate in mixed- as opposed to single-talker conditions, especially when the mixed-talker conditions included speakers with different dialect backgrounds. Specifically, reactions to Midland-accented tokens became slower when presented with Northern-accented tokens in the same block (Clopper, 2017, 35, 41). Reactions to Northern-accented tokens in mixed-dialect blocks did not become slower but less accurate. Thus, there is evidence for a cross-dialect interference effect on speech processing when the local and the non-local dialect are presented in one condition.

Clopper's (2017) second series of experiments investigated if such an effect also emerged when two non-local dialects with different levels of enregisterment are combined. While the dialect spoken in the American South is associated with many linguistic stereotypes and, thus,

[8] Due to the Northern Cities Vowel Shift, the /ɛ/ in *bed* and *head* can easily be confused with the /æ/ in *bad* and *bad*.

highly enregistered, the Northern dialect and its chain vowel shift are far less salient. In fact, speakers from the North self-report that they speak standard AE (Niedzielski, 1999; Preston, 2013). Again, word pairs were chosen that would likely result in the cross-dialectal confusion of tokens (*side/sod*, *ride/rod*).[9] All listeners in this series of experiments came from the Midland region. The results showed decreased accuracy for mixed-talker conditions. Both Northern- and Southern-accented tokens were classified less accurately when presented with the other dialect in the same block. Furthermore, reactions were slower for Northern-accented tokens in the mixed-dialect blocks (Clopper, 2017, 45-47, 50-51).

Clopper (2017, 57) concluded that processing costs were highest for non-local and non-enregistered dialects (Northern), lowest for the local dialect (Midland) and intermediate for non-local but enregistered dialects (Southern). Notably, these processing costs were independent of the correct overt identification of these regional dialects, which was poor in Clopper's (2017) experiments. With regards to familiarity, these results suggest processing costs should be lower for the dialect that participants are most familiar with and comparatively low for dialects that they are aware of in terms of stereotypes. Importantly, Clopper's (2017) results are mostly based on mixed-dialect blocks. Put differently, some of the processing costs did not emerge from blocks that contained a single speaker.

In the studies presented so far, familiarity with a dialect is the result of (long-term) exposure to a certain variety. In Floccia et al. (2009), for example, the Plymothian accent was considered familiar for the participants from Plymouth. Studies with artificially controlled accents can shed further light on how listeners adapt to artificial accents, which they cannot have encountered before due to their artificial nature. Maye et al. (2008, 547), who synthesised an artificial accent of AE with lowered front vowels, is a case in point. In their artificial accent, /i/ was realised as [ɪ], /ɪ/ was realised as [ɛ], /ɛ/ was realised as [æ] and, finally, /æ/ was realised as [a]. The participants first listened to a twenty-minute extract from *The Wizard of Oz* in GA before doing a lexical decision task. One to three days later, the participants heard the same extract in the artificial accent and, again, did a lexical decision task. The results showed a significant effect of exposure on the second lexical decision task. Items with lowered vowels (e.g. *witch* [wɛtʃ]) were judged significantly more often as words during the second task. This effect was also found for words that were not part of the extract, which demonstrates that participants did not only remember specific words from the text but generalised the lowered vowels to new words.

[9] For this experiment, Clopper (2017) took advantage of the monophthongal realisation of /aɪ/ in the American South.

A follow-up experiment, for which front vowels were systematically raised, showed that this adaptation is "specific to the direction of the shift [...] and not to a general relaxation of the criterion for what constitutes a good exemplar of the accented vowel category" (Maye et al., 2008, 543). Importantly, in both experiments, reactions were generally slower to tokens with altered vowels. Thus, Maye et al.'s (2008) results strongly indicate an adaptation effect but are still evident of a processing cost in terms of latencies associated with the unfamiliar accent. This suggests that lexical access for a newly encountered accent was successful in that the right stored representation could be retrieved. However, the process of retrieval might take longer than for linguistic input presented in a familiar accent. Recall in this context Adank and McQueen's (2007) finding that latencies for the East Flemish accent were not faster even after an exposure phase that was similar in length to the one used by Maye et al. (2008). Much longer exposure seems to be necessary so that latencies for local and non-local accents assimilate.

Maye et al.'s (2008) findings suggest that exposure leads to changes in the perceptual system. In this regard, Brunellière, Dufour, Nguyen, and Frauenfelder (2009) found that the contrast between /e/ and /ɛ/, which is absent in Southern French (Fagyal, Kibbee, & Jenkins, 2006) and positionally limited in Parisian French (Fagyal, Hassa, & Ngom, 2002) but stable in Swiss French productions, was perceptually less salient to their participants from Geneva. Brunellière et al. (2009, 391) argue that the Swiss French "are often exposed to variable pronunciations of words containing /e/ and /ɛ/ via the media or their interactions with speakers from other French-speaking regions", which results in a reduced salience of the phonological contrast. Thus, exposure to certain varieties influences the perceptual system. This, in turn, can facilitate lexical access, as the above studies have shown.

Importantly, as demonstrated by Kraljic, Brennan, and Samuel (2007), listeners take into account the source of the variation for adaptations to the perceptual system. Kraljic et al. (2007) researched if perceptual learning, specifically the shift of the boundary between /s/ and /ʃ/, was influenced by the source of variation. Their experiment consisted of an exposure and a test phase. During exposure, participants did a lexical decision task. Participants in the dialectal condition heard /s/ as [~sʃ] (an ambiguous production between the two fricatives) only before /tɹ/ (e.g. *string*), which is a feature of the Long Island and wider New York accent (Kraljic et al., 2007, 57, 59). In the idiolectal condition, participants heard /s/ as [~sʃ] not only before /tɹ/ but in all positions. During test, the perceptual boundary between /s/ and /ʃ/ was measured by category identification along a six-step continuum from [asi] to [aʃi]. The results

showed that, first, lexical decisions were more accurate and faster in the dialectal condition and, second, that perceptual learning only took place for participants in the idiolectal group (Kraljic et al., 2007, 62-65). The latter finding was independent of whether the participants' own dialect included the realisation of /s/ as [ʃ] before /tɹ/. This strongly suggests that the source of variation in the signal is informative for speech processing in that perceptual learning only took place in the idiolectal condition. Put differently, there was no adaptation in perception if the production of /s/ as [~sʃ] was governed by complementary distribution. However, this complementary distribution facilitated accuracy and speed of lexical access.

Taken together, the research in this subsection strongly suggests a processing cost associated with unfamiliar accents although the exact operationalisation of an '(un)familiar' accent varies between studies. Furthermore, listeners are highly sensitive to the source of variation in the incoming signal and their perceptual system is malleable. As a result, (long-term) exposure to an accent can facilitate lexical processing. This has been demonstrated by the processing benefit associated with many standard accents (Adank & McQueen, 2007; Impe et al., 2009; Sumner & Samuel, 2009). What remains to be seen is the amount of exposure needed so that the lexical processing of an accent approximates the baseline level for a local accent. While the processing of standard accents shows that life-long exposure does result in equally (if not more) successful lexical processing, short-term exposure in experiments does not necessarily seem to be sufficient. The question also remains when indexical information, such as the regional origin of a speaker, is used exactly during lexical processing. This is addressed in the following subsection.

2.5.1.3 Integration of Indexical Information during Lexical Access

One central question in lexical processing is whether stored lexical representations can be accessed without the concurrent processing and activation of indexical information. Furthermore, if indexical information is activated and used during lexical access, as suggested by exemplar theory, at what point during lexical access does this occur? A starting point to answer these questions is a seminal study by Van Berkum, Brink, Tesink, Kos, and Hagoort (2008), who measured the N400 effect, a measure of the detection of semantic anomalies, in 24 L1 Dutch speakers. The participants listened to sentences that were either semantically accurate, inconsistent with the speaker identity (e.g. male voice: *I recently had a check-up at the gynaecologist in the hospital*) or incongruent with world knowledge. Which category a sentence belonged to was determined by a single target word, which followed a generic intro-

ductory sentence and either did or did not breach the norms of sex, age, social class or world knowledge (Van Berkum et al., 2008, 582-583, 589). Van Berkum et al.'s (2008, 584-587) results showed that speaker-inconsistent sentences evoke an N400 effect that is similar to that of sentences that violate world knowledge, despite being smaller in size. Importantly, this N400 effect was measured within 200-300 ms of the onset of the critical word and before it was finished, which suggests that speaker information is taken into account at the very early stages of lexical processing. This finding corroborates Sumner et al.'s (2013) proposition that the speech signal is processed for social and linguistic information simultaneously. It is also congruent with the immediate link between social and linguistic information in exemplar-based models.

Martin et al. (2016) found that listeners use lexical frequency distributions of a dialect during lexical access, provided that they are sufficiently familiar with the dialect in question. Martin et al. (2016, 377-378) worked with lexical pairs that refer to roughly the same concept but are not as frequently used in British English (BE) versus AE (e.g. *holiday/vacation*). The N400 effect was measured in L1 BE participants. A questionnaire was used to determine that the participants' "self-rated exposure to spoken American on a 7-point scale (1: never; 7: everyday) was 4.8 (SD = 1.6)" (Martin et al., 2016, 378). The participants heard sentences in British and American voices. The critical items (e.g. *holiday* or *vacation*) did not occur at the very beginning or end of the sentences and were not strongly constrained by the sentential context. Martin et al. (2016, 381-382) found an interaction between the accent context and the critical word in that there was a stronger N400 effect if there was a mismatch between the voice and word (e.g. *lift* in an American voice). This suggests that stored lexical frequency distributions are not monolithic but take into consideration differences between dialects. Hence, there was a disruption in processing when British listeners hear an American speaker using *lift* because this word is less common in AE. Such a disruption could not be accounted for if British listeners' processing were solely based on the lexical frequency distributions in their own dialect. As a result, familiarity with a dialect appears to enclose the acquisition of lexical frequency distributions.

One indexical factor that has been discussed extensively in the literature is whether a speaker is a first or second language speaker. While the processing of L2 speech is not the focus of the current thesis, it can shed light on when and how indexical information is used during lexical processing. Goslin, Duffy, and Floccia (2012), for example, measured event-related potentials (ERPs) to find out if the same cognitive processes are used for all accents or

if the processing of regional versus L2 accents is fundamentally different and requires different types of cognition. Their research is situated within an abstractionist framework and they formulate two hypotheses that differ with regards to the locus of processing. According to the Perceptual Difference Hypothesis (PDH), all types of accents are normalised pre-lexically, with more effort required for L2 accents (Goslin et al., 2012, 93). In contrast, the Different Processes Hypothesis (DH) predicts that pre-lexical processing is sufficient only for the normalisation of regional accents as they "predominantly involve language coherent phonological variations" (Goslin et al., 2012, 93). Because of the less predictable nature of L2 accents, however, the DH states that more feedback from the lexical level would be necessary to eventually adjust the pre-lexical processing (see also Norris, McQueen, & Cutler, 2003).[10]

The Phonological Mapping Negativity (PMN) and N400 were measured in participants from Southwest England while they listened to short sentences in different accents. PMN and N400 were used as proxies of pre-lexical and lexical processing, respectively. Three groups of accents were used: the home accent (Southwest England), two regional accents (South Wales and Yorkshire) and two L2 accents (L1s Italian and Polish) (Goslin et al., 2012, 95-96).[11] Pre-tests with different participants from Southwest England found that intelligibility, as measured by transcription performance, was high for all accents (home 94% > regional 91% > L2 81 %). Furthermore, accent identification was best for the home accent while no significant difference was found for the regional versus L2 accents (Goslin et al., 2012, 95).

Goslin et al.'s (2012, 97-99) results showed that, if ERPs for the home accent were used as a baseline, regional accents elicited the strongest PMN while L2 accents were associated with the weakest PMN. For N400, the highest amplitude was found for L2 accents while no significant differences were detected between the sentences presented in the home and regional accents. These results can be better explained by the DH since they suggest more pre-lexical processing for regional accents. The locus of processing for L2 accents, on the other hand, seems to be at the lexical level (higher N400 amplitude). This suggests that pre-lexical processing is "sufficient to normalise unfamiliar regional accents, as by the 350-600ms epoch [where N400 is typically found] there [is] no longer any significant difference between the ERP for the home

[10] The claim that L2 accents are more variable than regional accents or that L2 accents vary in less structured patterns is also found in other published work (e.g. Floccia et al., 2006, 1277). This claim appears to be based on contrastive accounts between the phoneme inventories of two languages and should be treated with caution for two reasons. Firstly, even the speech of L1 speakers can vary considerably, especially when they are geographically mobile (e.g. Nycz, 2015, 2018). Secondly, such arguments can easily feed into the perception of 'native' accents as more correct and only obtainable by L1 speakers of a language.

[11] Goslin et al.'s (2012) subsuming of two different accents into one category (e.g. the South Welsh and the Yorkshire accent into the regional category) might be problematic due to the differences between these accents. This will be discussed further in subsection 2.5.2.2, when research by Stringer and Iverson (2019) is presented.

and regional accents" (Goslin et al., 2012, 101). Feedback from the lexical level is necessary for the processing of L2 accents, as evidenced by the high N400 amplitude. The results from the transcription task in the pre-test demonstrate that the amount of pre-lexical processing is not correlated with intelligibility as PMN was highest for the regional rather than the L2 accents, which constitutes further evidence for DH (Goslin et al., 2012, 100). In summary, Goslin et al. (2012) argue that pre-lexical processing is more successful for unfamiliar regional than L2 accents, the latter of which require feedback from the lexical level. What is especially important for the current discussion is that these differences in the processing of regional versus L2 accents suggest that this piece of indexical information is used to inform the (pre-)lexical processing of the incoming signal. Again, this suggests an early integration of indexical cues.

While Hanulíková, Van Alphen, Van Goch, and Weber's (2012) study focuses on syntactic rather than lexical processing, it is another good example of how processing is influenced by indexical information. They analysed ERPs of 30 native Dutch participants, who listened to a total of 344 sentences recorded by first and second language (L1 Turkish) speakers of Dutch. These sentences were either correct, grammatically incorrect (e.g. incorrect inflection of the adjective) or contained a semantic anomaly (Hanulíková et al., 2012, 880-881). The results show that a P600 effect, which is commonly associated with the processing of syntactic errors, was only observed when the participants listened to an L1 speaker while it was absent when they processed the L2 accent. Furthermore, the P600 effect for L1 speech was absent during the second half of the experiment, which suggests a short-term adaption to L1 speakers' grammatical errors (Hanulíková et al., 2012, 882-883, 886).

Importantly, Hanulíková et al. (2012, 883-884) found practically identical N400 responses for semantically anomalous sentences, regardless of whether Dutch was the individual speaker's L1 or L2. These results suggest that L1 listeners pay less attention to frequent grammatical errors while not disregarding the semantic content of sentences produced by L2 speakers entirely. L1 speakers' ignorance of the grammatical errors is likely a result of "accumulated experience with L2 speakers, leading to changes in estimates of the probability of correct formal features" (Hanulíková et al., 2012, 885). A first indication of how L1 listeners are desensitised to grammatical errors is given by the decrease of P600 during the second half of the experiment in the L1 condition. When errors occur frequently, they are first corrected effortfully by the system but then the neural response decreases. As a result, comprehension becomes more efficient as cognitive resources are not used to correct these mistakes, which is

interpreted by Hanulíková et al. (2012, 886) as part of "a natural and automatic attempt to ensure successful communication between native and non-native speakers".

In summary, there are two common themes that emerge from the neurolinguistic studies in this subsection. First, the integration of indexical information (e.g. gender, age, L1/L2 speaker, etc.) into lexical processing occurs during the early stages of processing (Van Berkum et al., 2008). Furthermore, the processing of the incoming signal itself is affected by indexical cues, which provides further evidence for its early integration. Specifically, indexical cues trigger different lexical frequency distributions (Martin et al., 2016), loci of processing (Goslin et al., 2012) and amplitudes of responses (Hanulíková et al., 2012). For the processing of regional (and L2) accents, it is sensible to assume that the degree of familiarity affects the size of these effects. For example, the differing processing modes in Hanulíková et al. (2012) are the result of long-term exposure to L1/L2 speech. Research into how familiarity affects recall is presented next. Recall performance, which is closely related to lexical processing, constitutes the second measure of interest for the research in this thesis.

2.5.1.4 Recall Performance

The scope of the research in this thesis goes beyond (sub-)lexical processing. One of the major aims is to determine how familiarity, intelligibility as well as attitude affect the recall of speech in different accents. This subsection provides an overview of the research into recall performance as mediated by different accents. While the majority of studies focus on L2 speech, their methodology and findings remain informative for the current discussion.

Lev-Ari and colleagues have conducted several studies to investigate if there are differences in the recall of L1 versus L2 speech. Lev-Ari and Keysar (2012) and Lev-Ari (2015a) used a change detection paradigm, in which participants first listened to a story and were then asked to correct the twelve lexical changes in a written transcript of said story. The participants listened either to an L1 or an L2 speaker, who only differed in their accent but not, for instance, in lexical or discourse-pragmatic choices. Participants in the L1 condition detected significantly more changes than those in the L2 condition (Lev-Ari & Keysar, 2012, 529-530). Importantly, this difference became non-significant when participants were told about the memory task before listening.

This suggests tentatively that participants are able to represent L2 speech to the same level of semantic detail when asked to listen for memorisation rather than comprehension. Lev-Ari

and Keysar (2012) argue that the identification of a speaker as a second-language speaker results in a different processing mode. For L2 speech, L1 listeners employ a less-detailed processing mode, which negatively "influence[s] the level of detail in the representation of [...] [a] speaker's language" (Lev-Ari & Keysar, 2012, 536). Less-detailed processing or "good-enough representations" (Ferreira, Bailey, & Ferraro, 2002) state that the representation of linguistic input is frequently not robust and underspecified so that stored schematic information is used to aid comprehension. Good-enough representations have been used to explain, for example, the effect of linguistic focus or the use of active versus passive voice on memory performance (Ferreira et al., 2002; Sturt, Sanford, Stewart, & Dawydiak, 2004). Lev-Ari and Keysar (2012) were the first to assume that social information mediates the level of processing and the extent to which good-enough representations are used in comprehension. The speech of L2 speakers might be processed in less detail because they are assumed to have lower linguistic proficiency.

Interestingly, as Lev-Ari and Keysar's (2012) second experiment demonstrated, participants with higher working memory performed more poorly in the L2 condition, that is to say they detected fewer changes in the transcript. This might seem surprising because higher working memory is usually associated with better performance in memory tasks (e.g. Cowan, 2008). However, Lev-Ari and Keysar (2012, 535) argue that adjusting the processing mechanisms to a less-detailed mode for L2 speech is effortful and only occurs in individuals with higher working memory, which then results in less-detailed processing and poorer recall.

The use of good-enough representations and schematic information in the processing of L2 speech is further supported by Lev-Ari (2015b), who tracked listeners' eye movements while they selected images. The American participants listened to either an L1 or an L2 speaker of English, who instructed them which pictures to select. The participants selected four pictures per trial: The first three tokens (e.g. witch, man on magic carpet, Santa) built up a dominant context (*imaginary creatures*) and a less dominant context (*modes of transport*). In the decisive forth trial, the participants saw four images: the competitor, which fits the dominant context (a mermaid); the target, which fits the less dominant context (a ferry); and two semantically unrelated images. Importantly, they were instructed to pick the [fɛɻi], a homophone with one meaning that most accurately fits the target from the less dominant context (*ferry*) and another meaning that was semantically related to the dominant context (*fairy*). However, for the dominant context, the word form (*fairy*) and the image (mermaid) did not match as well as the word form and image for the less dominant context (ferry).

The results showed that, first, participants selected the competitor significantly more often when listening to an L2 speaker and, second, that the competitor received significantly more fixations in the L2 condition from listeners with higher working memory (Lev-Ari, 2015b, 6-8). Conversely, listeners with lower working memory displayed no such preference in fixations for the context-induced competitor. This finding again suggests that the integration of indexical information, here the L1/L2 status of the speaker, requires cognitive resources. The findings are further in line with Goslin et al.'s (2012) results that there are structural differences between the processing of L1 and L2 accents (see, however, Stringer & Iverson, 2019).

Further evidence on this matter comes from Lev-Ari (2015a, 190) and Lev-Ari, Ho, and Keysar (2018, 836-837). Their participants first performed the change detection task from Lev-Ari and Keysar (2012) to calibrate their processing mode to a more general and context-dependent state for L2 speech and the opposite for L1 speech. The participants then performed a retrieval-induced inhibition task[12] that used 18 lexical items: six items per three lexical categories (animals, fruit and occupation) (Lev-Ari et al., 2018, 838). During the memorisation phase, the first out of three phases of this task, the participants saw all items and their corresponding category on a screen. During the practice phase, participants saw the category name and the first letter of six tokens (from two categories), which they were asked to identify. As a result, the test phase included three types of tokens, with "a third of the tokens [...] practiced, a third of the items [...] not practiced but belong[ing] to the practiced categories ('inhibited'), and a third of the items [...] not practiced and belonging to the non-practiced category ('control')" (Lev-Ari et al., 2018, 840). In the test phase, accuracy and latencies were measured for participants' decisions on whether or not they had seen the presented items in any of the previous phases.

The results in Lev-Ari (2015a, 191) and Lev-Ari et al. (2018, 840-841) showed that, regardless of the previous perception of L1 or L2 speech, participants' latencies were significantly lower for practised than for control items, as was to be expected. The difference in latencies between control and inhibited items, however, was only significant in the L2 condition, with inhibited items requiring longer latencies than control items. Lev-Ari (2015a, 191) and Lev-Ari et al. (2018, 841) argue that latencies were longer for inhibited items in the L2 condition because the induced less-detailed processing mode increased the lexical competition between these items and the inhibition required to make the decision whether the item had been seen before.

[12] In retrieval-induced inhibition, the activation of a unit in memory results in the inhibition or forgetting of another one (e.g. Verde, 2021).

One question that remains is why the less-detailed processing mode that was activated during the change detection paradigm would be extended to the inhibition task. Since the inhibition task presented the tokens visually on a screen without audio input, the previously activated processing mode would not necessarily be expected to be relevant for the processing of these words. Nevertheless, the spill-over effect in processing mode from the change detection to the inhibition task is also evident in representations of one's own speech when interacting with L2 speakers. In another experiment, Lev-Ari et al. (2018, 841-842) had participants interact with a confederate who was either an L1 or an L2 speaker of English. The participants read a short story and were then asked questions about it by the confederate. After some distractor tasks, the participants should select their previous answers from a list of pre-constructed answers to the questions. Lev-Ari et al. (2018, 843) report that participants interacting with L2 speakers remembered fewer of their own answers and were significantly more prone to false positives in the selection of answers. Somewhat similar observations were made by Baxter, Khattab, Krug, and Du (2022), who had their participants listen to either an L1 speaker, an advanced L2 speaker or an intermediate L2 speaker. Participants' recall of their own responses was worse for the advanced L2 as compared to the L1 speaker. For the intermediate L2 speaker, higher working memory was associated with poorer recall.

Recall performance for different regional L1 accents of Spanish was measured by Frances et al. (2018). Their experiments were based on Lev-Ari and Keysar's (2012) change detection paradigm (see above) and Lev-Ari and Keysar (2010), who found that L1 speakers of AE judged trivia statements as less credible when they were presented in an L2 voice. Frances et al. (2018) presented participants from Barcelona with trivia statements that were recorded with speakers from Barcelona (local accent) or Latin America (non-local accents). In their three experiments, which revolved around memory performance and credibility judgments as mediated by accent, they found that "listeners remember and believe speakers' messages similarly regardless of their accent, both when explored directly and indirectly" (Frances et al., 2018, 68). While the participants did rate the non-local accents as more accented and less intelligible in most experiments (Frances et al., 2018, 66-68), these ratings did not affect the memory and credibility measures. Frances et al. (2018, 69) used Latin American accents that are associated with relatively high prestige and speculate that differences might emerge for accents with more diverging social evaluations.

An insight into how these recall differences might come about is given by Clopper et al. (2016, 87), who distinguished between lexical recognition, which "involves mapping the

incoming acoustic signal to the target lexical category", and lexical encoding, which involves "updating the cognitive lexical representation to reflect the current token". Clopper et al. (2016) analysed recognition and encoding differences between Midland and Northern AE. The Midland dialect is more similar to GA than the Northern dialect, which is characterised by the Northern Cities Vowel Shift (Clopper et al., 2016, 88-89). Participants from both dialect backgrounds were tested in the Midland region. Clopper et al.'s (2016, 89-92) first experiment tested encoding strength: During familiarisation, the participants transcribed words spoken in both dialects. During test, they had to decide whether the auditorily presented word had been part of the transcription task as quickly and as accurately as possible.

Since different speakers were used to create the stimuli, the memory test included three different types of tokens: first, tokens produced by the same speaker during familiarisation and test (same talker); second, tokens produced by a different talker during test but with the same dialect background as the speaker of the token during familiarisation (different talker); and, third, tokens produced by a talker during test with the other dialect background compared to the speaker of the token during familiarisation (different dialect). The results showed that both dialects were highly intelligible, that is transcription performance was very good for participants of both Midland and Northern origins. In terms of memory, responses were overall faster and more accurate for same-talker tokens, which is in line with the assumptions of exemplar-based models presented in subsection 2.2.2. Interestingly, for the other tokens, responses were faster for the different-dialect as opposed to the different-talker group, which indicates "that identifying a target word as 'old' is easier when the test and study tokens are produced by different talkers from different dialects" (Clopper et al., 2016, 94). This suggests that encoding strength is not necessarily higher for the local or native accent since Midland participants, for example, encode Northern tokens more strongly.[13]

Clopper et al. (2016, 95-99)'s (2016, 95-99) second experiment tested implicit rather than explicit lexical memory. Participants performed two non-speeded transcription tasks in noise. In line with the first experiment, tokens in the second task were either part of the same-talker, different-talker or different-dialect group. Implicit lexical memory was operationalised as "a difference score calculated from the mean accuracy of the old words in each of the two blocks [...] of the experiment" (Clopper et al., 2016, 96). Put differently, transcription accuracy for words that were included in both the first and the second task was used to measure if

[13] It might be the case that retrieval-induced inhibition or forgetting contributes to this pattern of results (see footnote 12). Here, the fact that the accent was the same but the speaker was different (different talker) might have contributed to the worse performance for this type of tokens.

the repetition of words resulted in better transcription performance and, thus, implicit lexical memory. Clopper et al. (2016) found a stronger repetition effect for tokens produced in the Midland dialect. This suggests, first, that the Midland dialect, which is closer to the standard GA, is encoded more strongly in noise, and, second, that the Northern dialect results in a greater processing cost for both Midland and Northern participants (Clopper et al., 2016, 100). Taken together, Clopper et al.'s (2016) experiments suggest that processing differences between familiar/local versus unfamiliar/non-local dialects surface more visibly in noise. Clopper et al.'s (2016, 100) results further suggest that Northern participants did not encode Northern-accented tokens more strongly than Midland-accented ones. The Midland dialect is closer to GA than the Northern one and it is also more relevant locally as the experiments were conducted in the Midland region. Thus, the experimental evidence put forward by Clopper et al. (2016) suggests that proximity to the standard and relevance in the local context influence encoding strength in addition to whether or not a dialect is the participant's native dialect (see Evans & Iverson, 2007; Impe et al., 2009).

For German listeners, Grohe and Weber (2018) found an effect of accent familiarity on recall performance. Their experiments revolved around the production effect in memory, which denotes better recall of words that have previously been read out aloud (versus silently). Grohe and Weber's (2018, 574, 578, 580) participants were all highly familiar with the Swabian dialect found in South West Germany. During the familiarisation phase of the experiments, participants heard and read aloud individual words. Targets included a sequence of /st/, which, in Standard German is realised as [ʃt] syllable-initially (e.g. *Stein.statue* 'statue made of rock') and as [st] in all other positions (e.g. *Christ.baum* 'Christmas tree' and *Lis.te* 'list'). Importantly, the Swabian accent uses [ʃt] in all positions while [st] is used in many northern varieties.

Participants read aloud and/or heard targets in either the highly familiar Swabian accent [ʃt] or the less familiar Northern accent [st]. The test phase included a free-recall task, during which participants listed all words they could remember, and a visual recognition task, during which participants had to decide if the visually presented words had been part of the familiarisation phase (Grohe & Weber, 2018, 576). Grohe and Weber (2018, 581) found main effects for familiarity with the accent and presentation mode on memory performance. Memory was better when participants read aloud the tokens and when the tokens were heard or read in the familiar Swabian accent. These main effects remained stable when participants could not hear themselves (because of white noise presented over headphones) when they read aloud the

tokens. If familiarisation and test were one week apart, the main effect for presentation mode (read aloud versus heard) remained while familiarity with the accent no longer emerged as significant. Thus, Grohe and Weber's (2018) results suggest that processing a familiar accent does facilitate short-term recall performance but in addition to and not in interaction with the production effect.

While research on the lexical processing of regional accents has generally shown a processing cost for unfamiliar accents, the effects of familiarity on recall performance are varied. Less-detailed processing has been found for L2 speech (Lev-Ari, 2015a; Lev-Ari et al., 2018; Lev-Ari & Keysar, 2012) but Frances et al. (2018) could not replicate Lev-Ari and Keysar's (2012) results with different regional accents of Spanish. In Clopper et al. (2016), there is evidence of a recall advantage for more familiar and locally relevant accents but their study focused mostly on individual words, at least for the explicit recall measure. Lastly, Grohe and Weber (2018) found a short-term recall advantage for the familiar accent. Based on these findings, more research is needed to approximate more closely the role of familiarity on recall. The next subsection is concerned with intelligibility and its effects on lexical processing and recall.

2.5.2 Intelligibility

2.5.2.1 Operationalisation

Similar to familiarity, there is not one definition of intelligibility and this factor has been operationalised differently. Moreover, intelligibility is used as the dependent variable in some studies while it is integrated as an independent variable into other studies. For example, the stimuli in a lexical decision task might be presented at different noise levels, which would constitute the manipulation of intelligibility as an independent variable. As a dependent variable, intelligibility could be measured as the participants' transcription performance for different accents. The use of intelligibility as a dependent and independent variable could also be combined by implementing a transcription task at different noise levels. The non-uniform treatment of intelligibility is further complicated by its co-existence with the concept of comprehensibility. Before findings regarding the role of intelligibility in speech processing are presented, both concepts are defined and contrasted from one another to allow for terminological clarity.

One operationalisation of intelligibility and comprehensibility can be found in Munro and Derwing (1995) and Derwing and Munro (1997) in the context of the processing of L2 speech.[14] Here, intelligibility is referred to as "the extent to which the native speaker understands the intended message" (Derwing & Munro, 1997, 2) and has been operationalised in terms of transcription performance. Thus, L2 speech is considered more intelligible if L1 listeners are better at transcribing it. Comprehensibility, on the other hand, refers to the metacognitive cue "of how difficult or easy an utterance is to understand" (Derwing & Munro, 1997, 2) and is measured through ratings on Likert scales. As such, comprehensibility is one type of what Oppenheimer (2008, 2) refers to as *fluency*: "the subjective experience of ease or difficulty associated with a mental process".

Derwing and Munro (1997) showed in their experiments that transcription performance and ratings of comprehensibility and accentedness for L2 speech were related to one another. However, they remained separate elements to consider in that participants might judge a speaker as heavily accented and difficult to understand but their performance in the transcription task was still high. Therefore, participants' experience of how easily they could understand a speaker did not fully predict their actual performance, which is an important insight for the development of language attitudes (see Dragojevic et al., 2017 and Dragojevic, 2020).

Floccia et al. (2009) explained intelligibility and comprehensibility in terms of abstractionist models of lexical access. Within this framework, comprehensibility denotes "the processes necessary to retrieve a possible lexical candidate, involving mainly bottom-up activation from phonetic representations up to the lexicon" (Floccia et al., 2009, 380). Intelligibility refers to the processes that are necessary to select the most suitable candidate, which "entail[s] top-down mechanisms involving lexical and pragmatic knowledge" (Floccia et al., 2009, 380). Effectively, comprehensibility and intelligibility refer to pre-lexical and lexical levels of processing, respectively. According to Floccia et al. (2009, 379), intelligibility has been operationalised through accuracy measures (e.g. transcription) while, for comprehensibility, lexical decision latencies or ratings would be used (see Derwing & Munro's, 1997, use of Likert scales).

For the current thesis, the main aspect to consider here is the metacognitive nature of comprehensibility. Contrary to Floccia et al.'s (2009) definition, *comprehensibility* will not be used here to refer to latencies but will be reserved for metacognitive ratings by the participants.

[14] In fact, as Edwards, Zampini, and Cunningham (2018, 538) point out, a lot of research on intelligibility and comprehensibility (as well as accentedness) has been conducted on L2 speech and within the field of second language acquisition.

The term *intelligibility* will mainly be used to refer to manipulations of the stimuli (intelligibility as an independent variable, e.g. adding noise to the signal) and, in one instance, to an accuracy measure in a task (intelligibility as a dependent variable, e.g. performance in a transcription task; see subsection 4.2.3).

As has been mentioned in subsection 2.5.1.1, the factors familiarity and intelligibility are considered concurrently in many studies. The following subsection presents such studies, with a focus on lexical processing.

2.5.2.2 Intelligibility, (Familiarity) and Lexical Processing

Although Clarke and Garrett's (2004) study focuses mostly on the processing of L2 speech, it is included here because it is seminal in the field and the basis for many other important studies (e.g. Floccia et al., 2006, 2009). Clarke and Garrett (2004) conducted three experiments with a cross-modal matching paradigm[15] and found that L1 listeners of English adapted very quickly to Spanish- and Chinese-accented English. Clarke and Garrett (2004) divided the task into blocks that differed with regards to the accent(s) presented (L1 English, Spanish- or Chinese-accented English), the number of accents presented (single-accent versus mixed-accent condition) and the listening conditions (regular or added noise). In all experiments, the experimental sentences were followed by a block of L1 English. The latencies from these "*baseline* trials" (Clarke and Garrett, 2004, 3649, italics in the original) were used to normalise the experimental trials.

Clarke and Garrett (2004) found for all three experiments that the participants' latencies generally decreased, which resulted in reactions that were equally fast for the L1 and L2 speakers. Through the manipulation of listening conditions, Clarke and Garrett (2004) further found that the adaptation to L2 speech is accent-specific and not due to general learning mechanisms (see also Goslin et al., 2012 in subsection 2.5.1.3). The integration of Spanish- and Chinese-accented English allowed Clarke and Garrett (2004) to investigate the effect of long-term familiarity since Chinese-accented English is generally less common in the location of the experiment (Tucson, Arizona). For both L2 accents, Clarke and Garrett's (2004) results suggested a very quick adaptation by the participants. Thus, long-term familiarity did not provide a benefit for the cross-modal matching task. Adaptation, that is speeding up of

[15] In a cross-modal matching task, participants first hear a target in isolation or a sentential context. Next, they see a visual probe and have to decide if it matches the target.

reactions in the task, was stronger when noise was added to the stimuli, potentially due to increased task difficulty.

In summary, Clarke and Garrett's (2004) participants needed only one minute of exposure or less to Spanish- and Chinese-accented speech to react equally fast to stimuli presented in these voices as compared to an L1 voice. Their finding that long-term familiarity does not increase the speed of adaptation significantly supports "the hypothesis that the listener learned the characteristics of the [L2] accented speech on-line" (Clarke & Garrett, 2004, 3656). Clarke and Garrett (2004, 3657) acknowledge that adaptation in their experiments is to a specific L2 voice and might not necessarily generalise to other voices of the same accent, at least not with such minimal exposure. Clarke and Garrett (2004) follow an abstractionist framework to account for their finding by arguing that speech that diverges from the L1 norm can be processed more quickly if it has been encountered recently because the patterns of variation are available to the processing system. This is congruent with their finding that long-term familiarity did not significantly improve performance in the cross-modal matching task. Therefore, familiarity seems to play a minor role here. Manipulations of intelligibility, however, appear to induce stronger adaptation effects. Clarke and Garrett's (2004) results might have also been affected by a surprise effect due to a sudden accent change between blocks for some of the experiments (see Floccia et al., 2009 in subsection 2.5.1.2).

Evans and Iverson (2007) found only limited evidence for adaptation in their student participants, who left their northern hometown Ashby to study at different universities throughout Britain. Evans and Iverson (2007, 3815) conducted a longitudinal study with these students over the course of two years and hypothesised that "northern listeners who are regularly exposed to SSBE [Standard Southern British English] might [...] adjust their perceptual processes to better adapt to SSBE". Although such exposure is likely at university and participants did change their production of FOOT/STRUT and BATH towards the norms of SSBE, there was no general change in perception over time. Accordingly, participants' performance on the recognition of sentences spoken in SSBE in noise did not generally improve over the two years (Evans & Iverson, 2007, 3823-3824) although performance for this accent was generally higher than for the northern accent included in the experiment. There was further a between-participants effect. Those students who displayed more severe changes in production were also better at sentence recognition in noise than students with smaller shifts in production.

These results suggest that more exposure to SSBE does not necessarily result in better performance over time. The findings do, on the other hand, suggest that life-long exposure

to SSBE, which is also associated with higher ratings of prestige (e.g. Sharma, Levon, & Ye, 2022), facilitate the processing of this accent to the point that, under experimental conditions, sentence recognition in noise was more successful for SSBE than the participants' home accent (Evans & Iverson, 2007, 3823-3824). In fact, there might have been a ceiling effect in Evans and Iverson (2007) such that there was little potential for sentence recognition in noise to improve for the SSBE accent after the participants left Ashby. Importantly, this study did not include an unfamiliar regional accent, for which sentence recognition in noise might have been worse.

A processing asymmetry similar to the one described by Impe et al. (2009) for accents of Dutch and Sumner and Samuel (2009) for accents of AE was also found by Adank, Evans, Stuart-Smith, and Scott (2009). In their study, they asked participants, who were either L1 speakers of SSBE or Glaswegian English (GE) to judge the sensibility of statements in quiet and three different noise conditions. Self-reports on familiarity demonstrated that the GE participants were familiar with both GE and SSBE while the SSBE participants were less familiar with GE. The difference in familiarity resulted in a processing imbalance under adverse listening conditions: SSBE listeners' judgments were less accurate and slower when hearing GE as opposed to SSBE in noise. When the signal was free of noise, no difference was found between the accents. Importantly, GE participants processed both accents equally well, regardless of the listening conditions. In Adank et al.'s (2009) second experiment, SSBE listeners performed the same task with sentences spoken in a SSBE, GE and an L2 accent (L1 Spanish). Again, there was a processing delay for GE and an even stronger delay for the L2 accent. These results suggest that effects of familiarity might not necessarily surface in quiet and that standard varieties function as a centre of gravity, which ties into the discussion of the influence of attitude and prestige on accent processing (see subsection 2.5.3).

As discussed in subsection 2.5.1.2, artificially created accents, while raising questions of ecological validity, are a good way of ensuring participants' unfamiliarity with an accent. Adank and Janse (2010) is an example of how an artificial accent can be combined with manipulations of intelligibility in an experimental design. Adank and Janse (2010, 737), systematically changed the vowels of Dutch in their experimental stimuli: Standard Dutch /ɑ/ was realised as [aː] and vice versa, /ɔ/ was realised as [oː] and vice versa and so on. Young and old participants heard sentences in both Standard Dutch and the newly created accent. Speech perception thresholds were measured by adding different levels of noise to the signal and assessing at which levels listeners were still able to repeat at least half of the sentence back.

The results of the experiment showed that listeners were more tolerant to the added noise for Standard Dutch as compared to the artificial accent (Adank & Janse, 2010, 738). The processing cost for the artificial accent decreased over the course of the experiment, which indicates adaptation to this unfamiliar accent. Moreover, an effect of participant age emerged in that "older listeners [overall] had considerably more difficulty understanding the novel [i.e. artificial] accent" (Adank & Janse, 2010, 739). The mode of adaptation to the unfamiliar accent also differed between the two age groups. Older participants' initial adaptation to the artificial accent was more successful but their performance declined in the later blocks. Younger participants' performance, on the other hand, improved with increased exposure. Similar to the studies presented in this subsection so far, Adank and Janse's (2010) results suggest that listeners are more tolerant to adverse listening conditions for familiar accents. Here, there is further evidence for a learning effect that decreases the processing cost associated with the artificial accent.

To conclude the summary of results relating to intelligibility and its interaction with familiarity, a study will be introduced that partly questions the importance of familiarity and focuses on accent similarity instead. Stringer and Iverson (2019) investigated how the spectral and durational similarity between the speaker's and the listener's accents affect processing. Stringer and Iverson's (2019, 2214) study is based on evidence which suggests that the processing of an accent is not only determined by the amount of one's exposure to it but also "the acoustic-phonetic similarity between talker-listener accents". For example, in Pinet, Iverson, and Huckvale (2011), accent similarity trumps familiarity in that low-level French learners of English recognised sentences in noise equally well for stimuli in SSBE and Korean-accented English. While an explanation based on familiarity would predict a better performance for stimuli in SSBE because it is the accent used for many learning materials, the spectral analyses revealed that the French learners' vowels are equally similar in quality and duration to the vowels in the SSBE and Korean-accented stimuli.

To test the effect of accent similarity, Stringer and Iverson (2019, 2216-2218) conducted two experiments: one that measured ERPs and one that measured sentence recognition in noise. Their participants were SSBE speakers and learners of English from the North East of Spain. In the sentence-recognition experiment, the participants heard a sentence and were asked to repeat it back to the experimenter. The stimuli were taken from the newly developed non-native speech recognition (NNSR) sentences (Stringer & Iverson, 2020), which are also used in the current research project (see subsection 3.2.2.1.2). The sentences were recorded in

SSBE, GE and L2 English (L1 Spanish) and each of these accents was presented in quiet and noise with three different signal-to-noise ratios. Intelligibility was operationalised as the number of key words in the sentences that the participants could repeat back to the experimenter. Lastly, recordings were made of the participants' speech to measure the mean durational and qualitative similarity between the participants' vowels and those produced by each speaker group (SSBE, GE and L2).

Stringer and Iverson (2019, 2218-2219) found that their English listeners demonstrated impaired intelligibility of SSBE stimuli only under the most adverse listening conditions. For light noise levels, GE and SSBE were equally intelligible for English listeners although higher noise levels resulted in a decreasing intelligibility of GE stimuli. The L2 stimuli consistently scored lowest in terms of the number of key words repeated. The group of Spanish participants performed equally well with the L2 and the SSBE stimuli in all signal-to-noise condition. Across all conditions, GE was the least intelligible for this group of listeners.

Importantly, these results were largely congruent with the spectral analysis in that, for English participants, performance in the task followed the pattern of accent similarity. On average, the English participants' vowels were, unsurprisingly, closest to the ones produced by the SSBE speakers for the stimuli, followed by the GE and then the L2 speakers. The Spanish listeners' vowels were closest to the L2 speakers' vowels, followed by the SSBE and then the GE speakers. This hierarchy is congruent with their task performance for GE stimuli but does not readily explain why they performed equally well for SpE and SSBE stimuli. Stringer and Iverson (2019, 2222) suggest that "[t]his could reflect greater difficulties in acquiring spectral properties when producing English vowels [...], or could suggest that exposure to SSBE has helped the accent to become more intelligible than predicted by accent similarity". The results from the behavioural task were complemented by Stringer and Iverson's (2019) second experiment.

For this experiment, the participants' PMN and N400 were measured while they listened to predictable and semantically incongruent sentences from NNSR. The sentences were presented in quiet in the three accents from the repetition task (SSBE, GE and L2 English). Stringer and Iverson's (2019, 2214-2215) motivation behind this procedure was to capture effects in quiet that might not be detected by the less sensitive behavioural measures. The results demonstrated "that a PMN was not reliably elicited in this study" (Stringer & Iverson, 2019, 2219), which rendered a further investigation into the potential relationship between PMN and accent intelligibility unnecessary. Measures of N400, on the other hand, were partly

congruent with the findings from the sentence-recognition task in that stronger responses were measured in Spanish listeners for GE sentences (Stringer & Iverson, 2019, 2219). However, English listeners did not display a significantly stronger N400 for L2 English, which indicates "that the influence of talker-listener pairing at the lexical integration stage of word recognition is not solely due to accent intelligibility" (Stringer & Iverson, 2019, 2222). These results differ from Goslin et al.'s (2012) study on the processing of regional and L2 accents (see subsection 2.5.1.3). Contrary to Goslin et al. (2012), Stringer and Iverson (2019) neither found the strongest PMN for GE nor did they detect the highest N400 amplitude for L2 English. However, Goslin et al. (2012) subsumed two different accents in both their 'regional' and 'L2' accent groups. According to Stringer and Iverson's (2019, 2223) account on the importance of accent similarity, "combining accents in this way [...] may have obscured processing differences relating to the particular talker-listening pairing".

In summary, Stringer and Iverson (2019) demonstrated that, to a certain extent, familiarity effects can be accounted for by the similarity between the specific accents of the listener and speaker. Furthermore, these speaker-listener accent pairings appear to influence lexical processing. However, as noted by Stringer and Iverson (2019, 2213, 2222), accent similarity alone cannot account for the processing asymmetry found between the speakers of a standard versus non-standard accent (see Adank et al., 2009; Impe et al., 2009). In these cases, an accent-similarity account would predict equal levels of intelligibility for both types of listeners since their accents are equidistant. Put differently, the standard speaker would find the non-standard speaker as intelligible as vice versa because the spectral distance between the standard and the non-standard accent is the same as vice versa. Because of the evident processing asymmetry, Stringer and Iverson (2019, 2223) suggest that accent similarity "may provide a baseline level of intelligibility that exposure can improve on".

Taken together, studies that focus on intelligibility (and familiarity) demonstrate that differences in the processing of accents might not emerge in quiet, at least when behavioural measures are used. When noise is added to the stimuli, there is an evident processing cost for unfamiliar accents (e.g. Adank et al., 2009; Adank & Janse, 2010). Standard or more prestigious accents play a special role here as Evans and Iverson (2007) found that sentence recognition in noise is best for SSBE rather than the accent that is closest to their participants' native accent. In addition to familiarity, accent similarity can have an effect on the processing of accents although accent similarity alone is not sufficient to explain the processing imbalance

between standard and regional accents (Adank et al., 2009; Impe et al., 2009; Stringer & Iverson, 2019; Sumner & Samuel, 2009).

2.5.2.3 Recall under Adverse Listening Conditions

The current research project revolves around the recall of narrations, following up on the studies by Lev-Ari and Keysar (2012), Lev-Ari (2015a) and Lev-Ari et al. (2018) presented in subsection 2.5.1.4. There has been comparatively little research on the recall of narrations in noise, especially in terms of the recall of different accents. Within the few relevant studies, many of which investigate clinical participant populations with hearing loss conditions, recall performance has been operationalised differently. For example, Schneider, Daneman, Murphy, and See (2000) and Tye-Murray et al. (2008) used a questionnaire, while Piquado, Benichov, Brownell, and Wingfield (2012) and Ward, Rogers, Van Engen, and Peelle (2016) asked participants to freely retell the story. In Wasiuk, Radvansky, Greene, and Calandruccio (2021), the recall was measured by means of a visual recognition task. What these studies have in common is that they generally found poorer recall when listening conditions were more challenging.

A case in point are the findings from Ward et al. (2016), whose participants heard 60- to 80-word long stories under different listening conditions before being asked to recall them freely. Young (18 - 29 years) and old (64 - 76 years) participants were recruited for the experiment. The stories were presented in quiet and in an acoustically degraded format, which was achieved through vocoding. In vocoding, the speech signal is split into several frequency channels, the amplitude curve of which is extracted and then applied to a carrier signal in the same frequency channel (A. Warren & Gibson, 2023). Once the different frequency components of the carrier signal have been created, they are combined into the vocoded signal. The carrier signal in Ward et al. (2016) was broadband noise. While vocoding might not decrease the intelligibility of speech as much as noise masking, especially when more frequency channels are used, it does degrade the acoustic signal and it makes speech processing more challenging "[b]ecause spectral detail within a channel is lost" (Ward et al., 2016, 100). Recall performance was scored by segmenting the semantic content of the stories into idea units, which could be matched against the participants' responses. The idea units were organised hierarchically, with higher levels reserved for more detailed information such as descriptive adjectives (Ward et al., 2016, 101, 111). Besides the recall task, the participants' verbal working memory, general verbal

ability and hearing thresholds were measured. A sentence repetition task was used to confirm that the stimuli were intelligible.

Ward et al.'s (2016) results showed main effects of age, with younger participants performing better, and semantic level, with worse recall of more detailed information. Importantly, there was an interaction between listening conditions and semantic detail in that first-level information was recalled worse for the vocoded signal. Higher verbal working memory was generally associated with better recall performance. Taken together, these results suggest that recall is poorer if the intelligibility of the speech signal is decreased. Ward et al. (2016, 105-106) discuss that the size of this effect in their study was quite small in that recall in the vocoded conditions was degraded by no more than four percentage points compared to recall of the signal presented in quiet. However, this might be due to the generally high intelligibility of vocoded speech in contrast with, for example, noise-masked speech.

As mentioned above, studies that looked at narrative recall generally found that performance decreased for challenging listening conditions. These challenging listening conditions could be due to noise added to the speech signal, the vocoding of it, and/or because of degraded listening capacities in the participants. Adverse listening conditions appear to increase task difficulty, which results in poorer recall (see the effect of verbal working memory in Ward et al., 2016).

A larger body of research is available on the recall of word lists under adverse listening conditions. A case in point is Kjellberg et al. (2008), who further developed work by Rabbitt (1966, 1990) and Surprenant (1999). Their participants completed sentence recognition and recall tasks in quiet and in noise. Their working memory was also tested via a reading span test by Hällgren et al. (2001). The sentence recognition task consisted of an encoding and a recognition phase. During encoding, the participants heard 20 sentences, split between a noise (signal-to-noise ratio (SNR) of 4dB) and a quiet condition. After each sentence, the participants were instructed to repeat the sentence out aloud. During recognition, they saw sentences on a screen in front of them and had to decide if they heard these sentences during encoding. For the recall task, the participants were instructed to memorise two lists of auditorily presented words. During test, the participants wrote down the words they could recall. The lists contained 50 words each and one was played in quiet while the recording for the other contained background noise (SNR of 4dB). There was a pause in the recording after each word and participants were asked to repeat the word during this pause.

In terms of accuracy, the results showed that sentence recognition was equally good in noise and in quiet. Free recall of words was significantly worse under adverse listening conditions, especially for the first and final 10 items on the 50-word lists. Higher working memory capacity aided the recall of list-final items in noise but did not correlate with sentence recognition. Kjellberg et al. (2008) argue that word recognition requires more cognitive resources in noise, which leaves less capacity for subsequent encoding and recall, similar to the findings presented for narrative recall above. With regards to the absence of an effect of noise on sentence recognition, they argue that the task difficulty of the go/no-go design might not have been high enough to elicit an effect of condition (see also Frances et al., 2018 in subsection 2.5.1.4). In sum, Kjellberg et al.'s (2008) results suggest that the poorer recall of words presented in noise is due to the effortful recognition of words under adverse listening conditions.

In a semi-replication of the above study, Ljung and Kjellberg (2008) found that reduced free recall could also be induced by a prolonged reverberation time in the speech signal. Reverberation time is mediated by the acoustics of a space and "depends on how much of the sound [...] is reflected and how much is absorbed by the surfaces in the room" (Ljung & Kjellberg, 2008, 1). It refers to the time difference between the original and the reflected sound arriving at the listener, with longer reverberation times impeding on comprehension more strongly. Similar to Kjellberg et al. (2008), the participants completed a working memory reading span test, a sentence recognition task and a free recall task. The stimuli of the latter two were presented with either a short or a long reverberation time (0.53 s and 1.17 s, respectively). While the results showed no effect of longer reverberation time on sentence recognition, recall of words from this condition was significantly worse, especially for the initial parts of the word list. This again suggests that more resources are used during word recognition, which leads to less successful encoding and recall thereafter. However, working memory capacity did not mediate recall in Ljung and Kjellberg (2008), which, according to the authors, might be due to a lack of sensitivity of the working memory reading span test. Taken together, the results from Kjellberg et al. (2008) and Ljung and Kjellberg (2008) show that the recall of word lists is worse under adverse listening conditions. Importantly, this does not seem to be due to a failure in terms of identifying words but due the increased use of cognitive resources at that stage of decoding.

Marsh et al. (2015) also investigated the recall of word lists in quiet versus in noise. They applied an adapted version of the Deese-Roediger-McDermott (DRM) paradigm (Roediger

& McDermott, 1995)[16] and had their participants listen to ten word lists during encoding. Each word list consisted of 36 items that were clustered into three themes. The words under one theme were presented in a blocked manner and they all primed a critical lure, which, importantly, was not part of the list. After listening to each list, the participants were asked to recall the words from it. Marsh et al. (2015) calculated several recall measures, including the number of words recalled, the order in which they were recalled and the presence or absence of the critical lure in the participants' responses. They hypothesised that "listening to spoken words in noise reduces the semantic processing of those words" (Marsh et al., 2015, 4). In noise, more cognitive resources would need to be spent on the decoding of the words, which would result in reduced capacities for higher-order processing, encoding and recall. This should result in the recall of overall fewer words in noise but also the recall of fewer critical lures, which critically depends on semantic processing. These patterns were indeed present in the results. In noise, fewer words and fewer critical lures were recalled. Additionally, the semantic clustering in the order of the recalled words was less pronounced for the noise condition. Marsh et al.'s (2015) discussion of these results contains two relevant aspects. First, as mentioned before, lower-lewel processing such as decoding the word requires more cognitive resources in noise than in quiet, which leaves fewer capacity for higher-level processes. Second, the processing mode in noise is more verbatim and item-specific rather than gist-based and relational: "Item-specific processing involves encoding items by their features, elements, and distinctive qualities. Relational processing involves encoding items in relation to other concepts in memory" (Storbeck & Clore, 2005, 786). For the recall of longer stretches of speech in noise, these findings suggest that semantic relations cannot be established as reliably. This is in line with Ljung et al.'s (2009) finding that the recall of lectures by university and secondary school students was worse under adverse listening conditions even though the students could decode the speech signal effectively.

In summary, the findings of studies that have been conducted on the recall of narrations and word lists in noise suggest that recall is poorer and, for word lists, more item-specific under adverse listening conditions. This might be caused by a shift in how the available cognitive resources are allocated. In noise, more effort is required to decode the incoming speech signal

[16] In the DRM paradigm, participants are first presented with a list of words that share strong semantic ties with a word that is not on the list. For example, *hot, snow, warm, winter, ice, wet, frigid, chilly, heat, weather, freeze, air, shiver, Arctic* and *frost* are all strongly related to *cold* although the latter is not part of the list. When participants are asked to freely recall the words from the list, they often mention the critical lure *cold*, which is interpreted as an effect of false memory (Roediger & McDermott, 1995, 809, 814).

successfully, which leaves fewer resources for encoding and, finally, recall. The current research project will investigate how this interacts with the listeners' familiarity with an accent.

2.5.3 Attitude

2.5.3.1 Operationalisation

As demonstrated above, speech processing is not only affected by factors that are inherent to the incoming signal or the speaker. Individual differences between listeners, such as their familiarity with an accent, also affect the processing of stimuli. One individual difference central to the current research project is language attitude. Broadly, an attitude is defined in social psychology as "[a] mental representation that summarises an individual's evaluation of a particular person, group, thing, action, or idea" (Smith, Mackie, & Claypool, 2014, 230). The entity that is evaluated is referred to as the "attitude object" (Smith et al., 2014, 231). Accordingly, language attitude can be narrowed down as an individual's representation of how they evaluate a particular variety and the speakers of said variety. The attitude object is the variety and, by extension, its speakers.

Attitudes can be divided into explicit and implicit attitudes, with different research tools being used to measure the two types of attitudes. Explicit attitudes denote evaluations of attitude objects that "people openly and deliberately express [...] in self-report or by behaviour" (Smith et al., 2014, 232). Thus, explicit attitudes have been assessed by direct questions, observations, behavioural responses, Likert scales, etc. As will be shown in subsection 2.5.3.2, research into language attitudes frequently uses Likert scales to measure subjects' evaluations of varieties in terms of status, solidarity and, sometimes, dynamism (Giles & Billings, 2004, 190). Evaluations of attitude objects that are not subject to cognitive control are referred to as implicit attitudes. Different measures of implicit attitudes have been developed, such as attitude priming and, famously, the implicit association test (IAT; e.g. Greenwald, McGhee, and Schwartz, 1998). Put simply, the IAT measures and compares latencies based on the assumption that latencies are shorter for combinations of mentally more closely related concepts (e.g. *spider* and *bad* or *holiday* and *good*; see Hogg and Vaughan, 2018, 189). Research has shown that explicit and implicit attitudes do not necessarily converge. When measuring language attitudes, for example, it might be the case that participants are hesitant to evaluate the speakers of a variety overtly as 'uneducated' or 'unintelligent' while implicit measures might reveal exactly these associations. Importantly, implicit attitudes should not be mistaken

as "pure measures of what people 'really' think about attitude objects, while explicit attitudes are designed to dissemble and distort" (Smith et al., 2014, 233). On the contrary, the two types of attitudes simply measure different things and can complement each other to achieve a more complete picture of language attitudes.

This subsection now addresses the potential importance of explicit and implicit language attitudes for speech processing. First, it will provide a more general overview of current research into language attitudes held by people in the United Kingdom (UK). Then it will present studies that looked specifically at the effect of attitudes and affect on lexical processing and recall.

2.5.3.2 Language Attitudes in the UK

Montgomery (2018, 131-136) summarises language attitude research by Bishop, Coupland, and Garrett (2005); Coupland and Bishop (2007); Giles (1970); ITV News (2013); Strongman and Woosley (1967) and Trudgill (1982), which has been conducted over approximately the last 50 years. The individual studies differ as to which attitude object the participants responded to (conceptual terms for different varieties or specific speech samples of accents) and the scales that were used to measure attitudes. Rather than going into too much detail about each of the studies,[17] the intention here is to demonstrate how language attitudes have developed over time, especially those held towards Tyneside English (TE) and NZE as these accents will be part of the experiments in the following chapters. The overview from Montgomery (2018) is complemented by Levon et al. (2023) and Sharma et al. (2022), which provide the most recent insight into language attitudes in the UK.

There is an obvious caveat in comparing studies with different participant samples and rating scales for the attitudes measured. However, studies since the 1970s show a remarkable consistency in the evaluation of different varieties both in terms of prestige and solidarity/social attractiveness (Levon et al., 2023; Montgomery, 2018, 134-135). One interesting finding is that prestige and social attractiveness ratings for RP have dropped considerably from 1970 to 2022. Nonetheless, RP still ranks very highly among the different varieties included. Similar consistency can be found at the other end of the scale. Birmingham, Liverpool and Cockney English, for example, consistently scored low in terms of prestige. Social attractiveness ratings for TE, on the other hand, were comparatively high across the studies.

[17] See Montgomery (2018, 131-136) for further details on each study.

Specifically, TE ranked tenth and twenty-third out of 34 accents for social attractiveness and prestige, respectively, in Coupland and Bishop (2007, 79). In this study, participants did not listen to specific speakers but responded to accent labels. Coupland and Bishop's (2007) results also included ratings for NZE, which scored higher than TE on both scales. It ranked sixth for social attractiveness and seventh for prestige. In the ITV News (2013) study, TE ranked second out of 10 varieties in terms of friendliness. For the intelligence and trustworthiness ratings, TE occupied the seventh and fifth place respectively. NZE was not included in this study but there is information on this variety in Sharma et al. (2022), where it ranked fifth (out of 38) in terms of prestige and fourth in terms of social attractiveness. TE occupied rank 32 and 26 for prestige and social attractiveness, respectively, in Sharma et al.'s (2022) study.

Taken together, the studies presented so far demonstrate that language attitudes in the UK have changed relatively little over the last 50 years. TE has consistently scored higher in terms of solidarity than prestige and NZE has received favourable ratings on both scales. One additional finding from the ITV News (2013) study was that 28% of the participants reported having been discriminated against based on their accent in either a social or professional situation. This makes research into language attitudes and its potential effects on speech processing and recall even more important. Before such research is summarised, the complexity of language attitudes will be addressed.

2.5.3.3 Complexity of Language Attitudes

Language attitudes are not monolithic entities and vary with regards to the attitude object (own accent versus someone else's accent), personality traits, education and social network, to name but a few examples. Dewaele and McCloskey's (2015) research provides a good illustration of how complex language attitudes can be. Dewaele and McCloskey (2015) measured the explicit attitudes of more than 2000 multilinguals towards an L2 accent in their own and others' language use with an online questionnaire. They found that participants held less negative attitudes to the L2 accent of others as compared to their own. They further found more positive attitudes to an L2 accent in general for participants who lived abroad, were older and had grown up and/or lived in an ethnically diverse community (Dewaele & McCloskey, 2015, 230-232). The analysis of personality traits revealed more positive attitudes towards others' L2 accent in participants with higher extraversion, emotional stability and tolerance of ambiguity. Participants with higher educational attainment displayed more negative attitudes

towards others' L2 accent. Lastly, Dewaele and McCloskey (2015, 230) found that participants with higher linguistic proficiency in several languages held more negative attitudes towards their own L2 accent.

Dewaele and McCloskey's (2015) finding that language attitude ratings were related to the Big Five personality traits (see O'Connor, 2002) suggests that these ratings are at least partly beyond the individual's conscious control, despite the potential mitigating effect of the effort to come across as politically correct. The multitude of factors that affected language attitudes demonstrate how complex these evaluations of one's own accent and others' accents are. Similar observations were made by Kraut and Wulff (2013), who measured language attitudes to recorded voices. Again, several factors, such as the recorded speakers' gender, their proficiency and L2, influenced the attitude ratings. Additionally, several interactions between these factors were found. Thus, complexity arises not only on the listener's or evaluator's side, as shown by Dewaele and McCloskey (2015) but also on the speaker's side. While these results mainly focus on L2 accents, a similar degree of complexity can be expected for different regional L1 accents, both in terms of the listener and the speaker.

Besides the many factors that influence attitudes, it is essential to recall that the concept of language attitude itself can be divided into explicit and implicit attitudes, which need not converge. Pantos and Perkins (2012) showed that with L1 and L2 accents. Their L1 speaker had "a predominantly mid-Atlantic accent from the Philadelphia region", while the L2 speaker was Korean because research by Lindemann (2003) demonstrated that listeners from the United States perceive this accent as L2 but cannot reliably identify the L1 of its speakers. The participants completed three tasks in total. The first task was an IAT with auditory stimuli produced by the AE and Korean-accented speaker. This allowed Pantos and Perkins (2012, 8-9) to measure implicit evaluations of the two voices, in particular the strength of association between the categories 'good/bad' and 'U.S./foreign'.

Next, the participants' explicit attitudes were tested with a fictional medical trial. The participants heard two testimonies, one in favour of the defendant, followed by a testimony from the plaintiff's side. Each testimony was recorded in both voices and a 2x2 design was used, which resulted in four treatment conditions (American-American, American-Korean, Korean-American and Korean-Korean). The participants rated each speaker on 11-point Likert scales for dimensions including "believability, credibility, trustworthiness [...], and clarity" (Pantos & Perkins, 2012, 10). They were further asked to identify the speakers' nationalities and which testimony they would rather support. For the final task, the partici-

pants were presented with two different verdicts and judged how fair they found each one (Pantos & Perkins, 2012, 10-11). Hence, the second and the third task both measured the participants' explicit attitudes, which could be compared to the results from the IAT.

Pantos and Perkins (2012, 11) found no significant differences between the L1 and L2 voice for the dimensions used in the second task and fairness judgments from the third task. However, when asked at the end of the second task which testimony they would support, the participants generally preferred the L2 speaker, "regardless of whether he testified for the defendant or the plaintiff" (Pantos & Perkins, 2012, 11). Interestingly, the results from the IAT showed an implicit bias in favour of the American-accented speaker. On average, the participants made their decisions in the IAT before they heard the entire word, which "suggests that for implicit attitude formation, *foreign* is a salient and meaningful out-group category for listeners" (Pantos & Perkins, 2012, 12, italics in the original). This observation also resonates with the early integration of socio-indexical cues into lexical processing (see subsection 2.5.1.3). Pantos and Perkins (2012, 11-12) further found that participants' more positive implicit attitude towards the L1 speaker positively correlated with their preference of the L2 speaker at the end of the second task. This correlation was only found for participants who were exposed to both voices during the second task.

In summary, language attitudes need to be considered multifactorial in three respects. First, characteristics of the listener affect attitudes (e.g. Dewaele & McCloskey, 2015). Second, characteristics of the attitude object/the speaker mitigate attitudes (e.g. Kraut & Wulff, 2013). Finally, language attitudes themselves can further be split into explicit and implicit attitudes, which require separate investigation (e.g. Pantos & Perkins, 2012). The next subsection presents research that investigates the potential link between language attitudes and the lexical processing of someone's speech.

2.5.3.4 Attitudes and Lexical Processing

Studies that look into the influence of language attitudes on lexical processing are relatively sparse and partly show contrasting results. Derwing, Rossiter, and Munro (2002), for instance, demonstrated that more positive attitudes towards L2 English do not result in improved comprehension. Derwing et al. (2002, 248-249) divided their participants into three different groups: an accent instruction group, a familiarity group and a control group. All groups took part in a pre-and a post-test, during which they did a comprehension task and a transcription task. The comprehension task consisted of listening to a short passage of speech

and then answering questions on it in bullet points (Derwing et al., 2002, 249). The stimuli for these tasks were recorded with L1 speakers of (Canadian) English and Vietnamese learners of English. The participants further completed a questionnaire with Likert-scaled and open-ended questions to measure their explicit attitudes towards L2 English during the pre- and the post-test.

Importantly, in the eight weeks between the tests, the three groups of participants received different interventions (Derwing et al., 2002, 249-251): Based on different tape recordings, the accent instruction group first discussed issues in cross-cultural communication that emerged from the recordings. They then learned about the characteristics of the English spoken by Vietnamese learners of English, such as differences in the phoneme inventories of the two languages. The familiarity group discussed the issues in cross-cultural communication but received no accent training. The control group, finally, did neither exercise in the eight weeks between pre- and post-test.

For both the comprehension task and the transcription task, Derwing et al. (2002, 251-252) found a main effect of time such that performance was generally better in the post-test. However, this was regardless of the intervention that the participants did or did not receive. Significant differences did emerge for the attitudinal questionnaire. Specifically, participants in the accent instruction group were more confident about their comprehension of speakers with an (L2) accent (Derwing et al., 2002, 252-253). As for the comprehension and transcription tasks, the measured attitudes generally became more positive, that is there was a main effect of time. However, the accent instruction group displayed the strongest improvements, which are mirrored by their responses to the open-ended questions from the questionnaire. Derwing et al. (2002, 253) report that these responses bear evidence of the training that the group received and "can be categorised [...] [in terms of] phonological features, patience/empathy, success, and confidence", which shows that explicit instruction is conducive for the improvement of language attitudes.

For the current discussion, it is important to note that Derwing et al.'s (2002) interventions did not affect the participants' performance during the linguistic tasks. This suggests that the changes in attitude do not directly affect the comprehension of a short passage and transcription accuracy. It might also be the case that the tasks used by Derwing et al. (2002) were not sensitive enough to capture the potentially small effects of attitude. In any case, attitude is still highly relevant in that, as argued by Derwing et al. (2002, 256), the training and the more positive attitudes resulting from it increase the participants' willingness to engage in

conversations with L2 speakers. This could increase familiarity, which is conducive to lexical processing (see subsection 2.5.1.2).

Evidence for the direct influence of attitude on speech processing is brought forward by Lev-Ari, Dodsworth, Mielke, and Peperkamp (2019), who argue that implicit attitudes are irrelevant for phonetic processing but become significant for lexical access. Their participants from North Carolina performed two types of tasks. For the phonetic task, they listened to 15 monosyllabic words in a Southern accent and then mapped the vowel in the word on to a vowel continuum ranging from a GA to a Southern version of the vowel. Importantly, half of the participants were made to believe that the speaker is from Ohio while the other half thought the speaker was from South Carolina (Lev-Ari et al., 2019, 2307). For the lexical task, participants listened to Southern productions of sentences and answered questions about the meaning of these sentences, the answer to which critically depended on a target word (e.g. *Jill was* **sad/said** *to be away from her hometown*). A Southern production of this target word mapped on to a different meaning in GA. In the example, *sad* is [sæd] in GA but [sɛd] in a Southern accent, which maps on to *said* in GA.[18] Again, participants thought the speaker was either from the North (Ohio) or the South of the United States (South Carolina) (Lev-Ari et al., 2019, 2306-2307). The measured individual differences were working memory, the quotient on the autism spectrum and implicit attitudes.

The results for the phonetic task showed a significant interaction between alleged speaker origin and working memory. If the speaker was presented as Southern, the vowels were classified more towards that end of the spectrum by listeners with higher working memory (Lev-Ari et al., 2019, 2308). This suggests again that the integration of indexical information into speech processing is effortful. If the speaker was presented as coming from Ohio, no such influence of working memory was found, possibly because a Southern accent is more salient. For the lexical task, a main effect of speaker origin emerged. Listeners accessed the accent-congruent meaning of the target, depending on where they thought the speaker was from (*sad* for GA and *said* for Southern accent). The speaker origin effect interacted with implicit attitudes in that participants were more likely to access the Southern-accent meaning of the word (*said*) when they had a negative implicit bias against the South and thought that the speaker was from there. This suggests that implicit attitudes can mediate the effect of indexical information on lexical processing. However, more research is needed to substantiate this initial claim.

[18] For this particular example, it must be noted that the Southern [sɛd] (*sad*) that the participants heard and the GA [sɛd] (*said*) differ in vowel quantity. This cue might be used by the participants to identify the former meaning rather than the latter, which is usually associated with a shorter vowel.

2.5.3.5 Affect, Working Memory and Recall

While there is little research focused on the effect of attitudes on speech processing specifically, there is a larger number of studies that investigated how affect or emotional state influence cognitive processes in general. Although affect is not one of the primary factors of interest for the current research, it is plausible that negative attitudes result in negative affect, which makes the research in this subsection relevant. Two aspects of the effect of attitude on cognition will be discussed in more detail in the following: first, the finding that negative affect reduces working memory capacity; and second, that negative affect results in a processing mode that promotes item-specific rather than relational encoding.

Figueira et al. (2017) researched the effect of negative affect on the contralateral delay activity (CDA). CDA constitutes an electrophysiological measure that indicates how many items are held in working memory (Vogel & Machizawa, 2004). There is a positive correlation between CDA and the number of items held in working memory, with "an asymptote at three to four items" (Figueira et al., 2017, 985). Figueira et al.'s (2017) participant pool consisted of 26 healthy individuals who completed 240 trials in total. During each trial, they first saw either a neutral or an unpleasant picture from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 2008). Next, they performed a visual change detection task. The participants had to decide if two arrays of coloured squares were identical or different. Specifically, the arrays were split into two halves and each half contained either two or four coloured squares. Before the first array was presented, an arrow showed the participants which half to focus on during this trial. Finally, the participants completed questionnaires to measure anxiety levels and susceptibility to intrusive thoughts. Figueira et al. (2017, 985) hypothesised that participants who were more anxious and prone to intrusive thoughts would experience negative affect more strongly, which would reduce working memory capacity more.

Figueira et al. (2017) did not find an effect of affect on the participants' decision as to whether the two arrays were identical.[19] As expected, because of the increased task difficulty, participants performed worse for arrays with four rather than two coloured squares. Importantly, CDA was mediated by an interaction between the number of squares and affect in that its amplitude increased more from the two- to the four-square array when neutral pictures were shown compared to when unpleasant pictures were shown. This indicates reduced working memory capacity for the participants in the unpleasant condition (Figueira et al., 2017, 989).

[19] Figueira et al. (2017, 990) suggest that the absence of a behavioural result might be due their study being underpowered.

Fittingly, the increase in CDA amplitude was higher for participants who scored lower on the anxiety and intrusive thought questionnaires. These results demonstrate that affect can have an influence on working memory capacity, which, in turn, might influence participants' performance on tasks that require short-term encoding and storage of information. As Figueira et al. (2017, 990) comment, similar findings were made by Dolcos and McCarthy (2006), Cohen et al. (2016), Pereira et al. (2006) and Stout, Shackman, and Larson (2013).[20]

In addition to this effect on working memory capacity, research suggests that different affective states result in different types of processing, with negative and positive affect promoting item-specific and relational processing, respectively (Gasper & Clore, 2002; Isen & Daubman, 1984). The two types of processing compete with each other (Arndt & Reder, 2003). The encoding of incoming information is fundamentally different for these types of processing: "Item-specific processing involves encoding items by their features, elements, and distinctive qualities. Relational processing involves encoding items in relation to other concepts in memory" (Storbeck & Clore, 2005, 786). Thus, the research presented here is highly relevant for the recall element of the current thesis, especially with regards to the role of attitude. Relational processing also appears to correspond to the less-detailed processing account raised by Lev-Ari and Keysar (2012) in that representations of the incoming signal are less detailed and are encoded more in relation to existing representations.

The influence of affect on how tokens are encoded has often been researched via the DRM paradigm (see footnote 16). In relation to the types of processing introduced above, item-specific processing should decrease the likelihood of false recall in the DRM paradigm because it is less dependent on the semantic relations between the items on the list. Relational processing, on the other hand, by its very nature should result in more false memories. The results from Storbeck and Clore (2005) followed this pattern. Their first experiment included 100 participants, who were allocated to three affect groups in a between-participant design: positive versus negative versus neutral. Affect was manipulated by playing different music pieces for eight minutes.[21] Next, the participants completed a recall task in the DRM paradigm (36 visually presented lists with 15 words, ordered by semantic proximity to the critical lure) and filled in a mood and personality questionnaire (Storbeck & Clore, 2005, 786-787). The

[20] See also Xie, Ye, and Zhang (2022) for a meta-analysis of different studies that investigate the effect of negative affect on the processing of information and working memory.

[21] The participants in the positive condition heard *Eine Kleine Nachtmusik* by Mozart while those in the negative condition heard *Adagietto* by Mahler. Previous research verified the effect that these pieces have on affect (Niedenthal & Setterlund, 1994). No music was presented to the participants in the control group.

results showed that the number of critical lures listed by the participants during recall was significantly lower in the negative condition.

Storbeck and Clore's (2005) second experiment was designed to identify the cognitive process behind the results from the first experiment. Specifically, they investigated if the critical lures were less available in the negative condition due to more item-specific encoding or if, alternatively, the participants used cognitive monitoring strategies more successfully to inhibit the lures. To this end, the instructions during the recall task were changed in the second experiment such that "participants were asked to recall items they had seen, but they were also asked to list any additional related words that had come to mind during the study or recall phase" (Storbeck & Clore, 2005, 788). These instructions were supposed to cancel the effects of potential monitoring processes. Put differently, if the results from the second experiment matched those of the first one, it would be due to a more item-specific encoding and, thus, lower activation of the critical lures. The results showed again that significantly fewer critical lures were recalled by participants in the negative condition. In addition, the extended instructions did not yield a higher number of recalled lures (Storbeck & Clore, 2005, 789). Granted that such a null result needs to be treated with caution, this indicates that more item-specific encoding rather than more successful memory strategies generated the results in the negative condition. Interestingly, the positive group generally patterned with the neutral or control group, which was congruent with their responses in the post-recall questionnaire. Taken together, Storbeck and Clore's (2005) results suggest that items are encoded more specifically, as opposed to how they related to other concepts in memory, when the individual's affective state is negative.

To provide further evidence for the claim that negative affect induces item-specific processing, Storbeck (2013) conducted three additional experiments, which were designed to include two separate dependent variables, one each for item-specific and relational processing. In Experiment 1, the affect manipulation was achieved via music (see footnote 21) and the recall task consisted of 14 trials with three stages each. During encoding, the participants saw items from a DRM list in one of four quadrants on the screen. During free recall, the participants recalled the list items. The number of lures mentioned by the participants was used as a measure of relational encoding. During spatial recall, six words from the list were shown in the centre of the screen and the participants had to decide which quadrant these words had appeared in during encoding. The number of correct answers here was used to measure item-specific encoding. The results were congruent with Storbeck's (2013) predic-

tions in that, first, participants in the negative conditions recalled significantly fewer critical lures and; second, their spatial recall was significantly better.

The second experiment was conducted to rule out two potential confounds (Storbeck, 2013, 805-810). First, the results in Experiment 1 might have been due to enhanced spatial working memory induced by a negative emotional state. To mitigate this, Experiment 2 included the recall of the font style the words were presented in rather than their location. Second, emotional valence might have confounded previous results if the positive condition was more arousing than the negative condition or vice versa. In reaction to this, Experiment 2 used images from the IAPS with similar valence ratings. 10 DRM lists were presented to the participants. The recall test was slightly different from Experiment 1 and used a go/no go rather than a free recall design. The results were again in line with the predictions: Participants in the negative condition made fewer mistakes for the critical lures and their sensitivity for font recall was higher.

Storbeck's (2013) Experiment 3 was designed to investigate if false recall in the non-negative conditions could be reduced if the participants' attention was "directed away from semantic qualities to perceptual qualities by the inclusion of pictures" (Storbeck, 2013, 810). Pictures accompanied the words on half of the 14 lists in a within-participant design. No spatial or font recognition task was included. The results showed that the number of critical lures recalled was lowest in the negative condition. However, in the positive and in the neutral condition, the recall of these lures was lower when the lists included pictures of the words. Taken together, the results from Storbeck and Clore (2005) and Storbeck (2013) show that negative and neutral/positive affect promote item-specific and relational processing, respectively. Relational processing can be disrupted if the attention of the individual is shifted away successfully from a semantic focus by means of task demands. This is in agreement with Lev-Ari and Keysar's (2012) finding that participants recall L2 speech better when they were asked to listen for memory rather than comprehension.

Kensinger (2009) further reviewed evidence on how emotion influences memory. Her summary demonstrates how the findings presented so far extend to encoding into and retrieval from long-term memory. In line with the findings above, participants are generally better at remembering negative occurrences in more detail and with greater accuracy than positive events. For example, Levine and Bluck (2004) tested the memory accuracy around the televised O.J. Simpson murder trial. Their findings showed that participants in favour if the verdict were more prone to accepting memory lures than participants against it. Similar findings were

made by Bohn and Berntsen (2007) and Kensinger and Schacter (2006). As neuroimaging studies demonstrate, this behavioural difference in the retrieval of emotionally charged information from long-term memory is associated with the activation of different brain regions for positively versus negatively loaded information during both encoding and retrieval (Levine & Bluck, 2004, 9-12).

In summary, the experimental evidence from several studies with different methodologies strongly suggests that affect mediates the encoding and retrieval of information in short-term as well as long-term memory. While negative affect might result in reduced working memory capacity, it also induces an item-specific processing mode. This results in memory traces that are less dependent on the activation of schematic information during retrieval, which emerges in behavioural experiments as a memory benefit in conditions with negative affect. These findings are highly relevant for the current discussion of the effect of attitude on speech processing. Granted that negative attitudes result in negative affect, the reduced working memory capacity might impede on working memory and associated linguistic processes, such as lexical access (see Lev-Ari et al., 2019). With regards to recall, however, negative attitudes might contribute to more specific memory traces and, thus, better performance.

2.6 Outlook on Experimental Chapters

This literature review has demonstrated that the influence of familiarity, intelligibility and attitude on lexical processing and recall is complex. There is an evident need for further research into this area, in particular with regards to recall performance and the effect of attitude. The following chapters will consider the three factors of interest one at a time.

Chapter 3

Experiment 1: Familiarity

3.1 Introduction

The broad aim of the research project is to investigate how familiarity, intelligibility and attitude affect the processing of different accents, especially with regards to lexical access and recall performance. For each factor, lexical access and recall will be tested with a lexical decision task and a change detection task, respectively. The experimental evidence presented in this chapter focuses on familiarity, which is defined here as someone's long-term familiarity with an accent through lifelong exposure.

The purpose of this chapter is twofold. First, it introduces the lexical decision and change detection task as well as additional tasks, which were all tested by means of the pilot study. Second, the results of the main experiment on familiarity are presented. As will be shown in more detail below, the pilot study and the main experiment largely used the same stimuli and procedure but differed in terms of selection criteria for participants, which were stricter for the main experiment. The remainder of this section summarises relevant previous research and provides the reasoning behind the design used here. It concludes with the research questions for Experiment 1 and an overview of the following sections in this chapter.

3.1.1 Accent Familiarity and Lexical Processing

The lexical decision task of the pilot study was developed based on Clarke and Garrett (2004), Floccia et al. (2006), Floccia et al. (2009) and Maye et al. (2008). These studies were chosen because they tried to answer similar research questions, namely the effect of different

accents on lexical processing. The studies have already been covered in subsection 2.5.1.2 but the methods and main findings are repeated here to motivate the research questions in subsection 3.1.3.

Although Clarke and Garrett (2004) used a cross-modal matching[1] rather than a lexical decision task, their findings are relevant because they found that the processing of L2 speech, while initially slower, quickly returned to performance levels for L1 speech. Specifically, differences in participants' latencies for Spanish- and Chinese accented English as opposed to AE dissipated within less than sixty seconds of exposure to the L2 accents. This finding suggests that, within the cross-modal matching paradigm, listeners adapted to L2 speech very quickly.

In a series of experiments using different regional and L2 accents of French, Floccia et al. (2006) used a lexical decision task to investigate accent normalisation. Their participants heard sentences and had to decide if the final word of the sentence was a real word or a nonword. The participants indicated that they heard a real word with a button press and that they heard a nonword with the absence thereof (go/no-go design). When the presentation of accents was randomised, they found longer latencies for unfamiliar regional accents as opposed to the participants' home accents, especially when the sentence preceding the target was longer. When the presentation of accents was blocked, latencies were not significantly different for home versus regional accents but longer for the L2 accent. These findings suggest that the initial presentation of an unfamiliar regional accent results in a processing cost, which is overcome in the blocked design. Importantly, however, Floccia et al. (2006, 1289) did not find this initial processing cost when the accent changed in their blocked design, which they attributed to the potentially insufficient power of the design and an insufficient number of observations per condition.

Floccia et al. (2009) then conducted experiments with different regional and L2 accents of English to further investigate the initial processing cost and the following adaptation. In a series of experiments, they presented different regional and L2 accents in both a blocked (15 sentences of the same accent) and a randomised manner. In the blocked design, Floccia et al. (2009) found longer lexical decision latencies (LDLs) for the regional and the L2 accent as compared to the participants' Plymothian home accent. However, they did not find evidence for a return of lexical processing to home-accent levels throughout the block. Through a

[1] Participants listened to a sentence and then had to decide if the visual probe presented afterwards was the final word of the sentence.

manipulation of the participants' expectations, Floccia et al. (2009) further found that part of the initial delay could be due to a surprise effect when the accent changes from one sentence to another. In the randomised design, latencies were again longer for the non-home accents, with no indication of adaptation over the course of the experiment whatsoever.

Finally, Maye et al. (2008) investigated adaptation by using an artificially created accent of English, which ensured that participants were truly unfamiliar with it. The artificial accent had consistently lowered front vowels (e.g /i/ → [ɪ]). Both of Maye et al.'s (2008) experiments consisted of two sessions that were up to three days apart. During one of the sessions, the participants first heard a twenty-minute long passage in GA, followed by a lexical decision task with words. The second session followed the same procedure but used the artificial accent. For the lexical decision task, items were presented auditorily and included GA pronunciations and artificially changed vowel qualities. The latter were either in line with the consistent lowering of front vowels of the artificial accent or represented other modifications (e.g. raising of back vowels). Importantly, Maye et al. (2008) found that their participants were more likely to classify tokens with lowered vowels as words after being exposed to the artificial accent. This effect followed the direction of the shift and could not be attributed to "simply relaxing the criterion for what constitutes a good exemplar for the front vowel categories" (Maye et al., 2008, 555). This is strong evidence for adaptation to the new accent although, in contrast with the previous studies, the participants were exposed to the accent substantially longer.

As will be explained in more detail in subsection 3.2.2.1.1, the lexical decision task in this pilot study tested reactions to (non)words placed at the end of carrier sentences, similar to (Floccia et al., 2009, 2006). The presentation of speakers was blocked, which allowed for the observation of potential adaptation effects, which are defined here as an increase in lexical decision accuracy and/or the reduction of LDLs after exposure to an accent. To prevent surprise effects as best as possible, the participants were informed before the task that they would encounter four different speakers. Of course, the participants did not know when exactly the speakers would change, which means that the contribution of a surprise element to the potential initial perturbation when encountering a new speaker could not be fully excluded.

3.1.2 Accent Familiarity and Recall

The following studies investigated how recall performance in change detection paradigms is affected by different linguistic and paralinguistic factors. These studies informed the current

research questions (see subsection 3.1.3) and the design of the recall task (see subsection 3.2.2.1.2).

In Sturt et al. (2004), participants read pairs of stories and reported back to the experimenter whether or not the stories were identical and, if they thought they were different, what the difference was. The differences between the stories were either semantically related (*beer* → *cider*) or unrelated (*beer* → *music*). To investigate the depth of semantic processing, Sturt et al. (2004) further manipulated linguistic focus through the use of (pseudo-) cleft sentences or the preceding sentence.[2] They found in both experiments that the detection rate for semantically related words was influenced by focus. Specifically, the detection rate was lower for these words when they were out of focus. Sturt et al. (2004) used this finding as evidence for good-enough representations: The depth of semantic processing is only as high as demanded by the linguistic focus. Detection rates are lower for unfocused than for focused targets because the latter are processed in more semantic detail, which makes the detection of changes even for semantically related words possible. For unrelated words, linguistic focus did not play a role since, even with good-enough representations, the semantic distance was sufficient to allow for the detection of changes. Importantly, all participants were instructed from the beginning to find changes between the stories.

Lev-Ari and Keysar (2012) also used a change detection paradigm. However, their approach differed from Sturt et al.'s (2004) in the following respects. First, they did not consider the semantic distance between the changed words. Second, rather than using linguistic devices to manipulate focus, they investigated the effect of different speakers (L1 versus L2) on recall performance. Third, they used one story with eleven changes instead of individual stories with a single change. Fourth, they manipulated the task demands between participants. In

[2] An example for how the target *the cider* is put into focus through a pseudocleft sentence is provided in (1a) while (1b) shows a near-identical story with the target out of focus (Sturt et al., 2004, 884; emphasis in bold added).

(1a) *Everyone had a good time at the pub. A group of friends had met up there for a stag night. What Jamie really liked was* **the cider**, *apparently.*

(1b) *Everyone had a good time at the pub. A group of friends had met up there for a stag night. It was Jamie who really liked* **the cider**, *apparently.*

In (2a), the target *with the hat* receives focus through the semantic context of the preceding sentence. (2b) again presents a near-identical story without focus on the target (Sturt et al., 2004, 885; emphasis in bold added).

(2a) *Everybody was wondering which man got into trouble. In fact, the man* **with the hat** *was arrested.*

(2b) *Everybody was wondering what was going on that night. In fact, the man* **with the hat** *was arrested.*

their 'comprehension' group, the participants were told that the story would be followed by comprehension questions whereas participants in the 'memory' group knew about the change detection task beforehand and heard the sentences of the story in random order. Finally, the participants heard the story and did not read it. Lev-Ari and Keysar (2012) found an interaction between task demands and speaker in that more changes were detected for the L1 than the L2 guise in the comprehension condition but not in the memory condition. Lev-Ari and Keysar (2012) suggested that the processing of L2 speech is less-detailed, unless this mode is overridden by the task demands. In a second experiment, Lev-Ari and Keysar (2012) found that the switch to less-detailed processing is effortful since the interaction between task demands and speaker could only be found in participants above a certain working-memory threshold. Lev-Ari and Keysar's (2012) findings suggest that contextual factors, such as the explicit instructions that a memory task will follow, override the effects of the speaker manipulation. Lev-Ari and Keysar (2012, 534-535) further argue that their findings were not driven by intelligibility alone since performance for the L2 speaker should be equal across the comprehension and the memory conditions if intelligibility were the driving force behind their results.

However, one aspect to consider here is that they did not control the semantic distance between the original words and the altered words in their change detection paradigm. While some changes were between related words (e.g. *optimistic* → *happy*), others were more distant semantically (e.g. *crowded* → *noisy*). As a result, it might be the case that, even in the memory condition, there could be an effect of the speaker if the words are semantically related. So far, Lev-Ari and Keysar's (2012) results suggest that a paralinguistic cue, that is the speakers' accent (L1 versus L2), can induce good-enough representations if the participants' working memory is above a threshold and if they are instructed to listen for comprehension rather than memory.

Whether this also holds for regional rather than L2 accents was tested by Frances et al. (2018). In one of their experiments, they had participants from Spain listen to 60 trivia statements, which had been recorded with speakers who either shared the participants' accent ('local' condition) or were from Latin America ('regional' condition). Afterwards, the participants read 60 statements, half of which they had heard during encoding. For the other half, the final word was replaced with a different word. The participants had to decide if they had heard the sentences during encoding but did not have to recall the original sentences. Similar to Lev-Ari and Keysar (2012), there was a comprehension condition and a memory

condition. In contrast, however, the speakers were used within participants. Put differently, the participants did not hear only one speaker but both local and regional guises. After the task, the participants rated the comprehensibility and accent of the speakers (*Spanish* versus *clearly foreign/not Spanish*; Frances et al., 2018, 65) on nine-point Likert scales.

Their results showed no effect of the speakers' accent on task performance, neither in the comprehension nor in the memory condition. No significant correlations were found between intelligibility or accentedness ratings and task performance. Lev-Ari and Keysar (2012, 535) suggested that the differential processing of L2 speech is correlated with "expectations regarding non-native speakers' linguistic competence". In Frances et al. (2018, 66), the regional speakers were rated as somewhat less intelligible on average (not statistically significant) although this did not seem to influence recall performance. Importantly though, Frances et al.'s (2018) task differed from Lev-Ari and Keysar's (2012) in that the participants were not asked to identify the original word, which might have resulted in an effect of accent. On the other hand, their design allowed for testing different speakers within participants.

The current experiments aim to capitalise on the methodological advantages of the studies above and to further test recall performance for various sets of accents under different listening conditions. Here, the focus is on the recall of familiar versus unfamiliar accents. As will be described in detail in subsection 3.2.3.3, all participants were instructed to detect changes between short stories, which were presented auditorily. The stories differed in a single word and, following Sturt et al. (2004), the change was either semantically related or unrelated. In contrast with Lev-Ari and Keysar (2012), different shorter stories with one change each rather than a longer story with eleven changes were used. This approach made it possible to control the location of the change without moving away as far from Lev-Ari and Keysar's (2012) design as Frances et al. (2018) did. Since participants were asked to identify the change after each pair of stories, it was not possible to manipulate task demands, which effectively resulted in the absence of a comprehension condition.

While this approach might seem counterintuitive at first sight, as this is precisely the condition for which Lev-Ari and Keysar (2012) found significant differences, it is justified by the following reasons. First, as pointed out above, it might be the case that there are still effects in the memory condition if semantic distance is included as an independent variable. Second, Sturt et al. (2004) found effects for the semantically related changes and their study did not include a comprehension condition. Third, the current design allows for a within-participants manipulation of accent and for the presentation of several stories produced by the

same speaker, which would have been impossible following Lev-Ari and Keysar (2012) because the surprise effect of the comprehension condition could not be renewed for a second or third speaker. It further makes it possible to use more than one speaker per accent, which makes the conclusions for the effect of accent more robust as the effect of idiosyncratic differences is reduced. Finally, the current experiments aim to test the immediate recall of information presented in different L1 accents, which can be tested most reliably with the design presented in subsection 3.2.3.3.

3.1.3 Research Questions

Based on the previous research presented above, the following research questions will be addressed:

(1) How does accent familiarity affect lexical processing?

(2) Is there evidence for adaptation in terms of lexical processing?

(3) What is the effect of word frequency on lexical processing?[3]

(4) How do accent familiarity and semantic proximity of the change affect recall?

(5) Overall, is there evidence for less-detailed processing of the unfamiliar versus familiar accent?

Next, further information is given on the pilot study. The following sections then address the participants, stimuli, procedure and data analysis steps for the main experiment. Finally, the experimental results are presented and briefly summarised in anticipation of the following chapter on intelligibility.

3.2 Pilot Study

The aim of the pilot study was to test the lexical decision and recall tasks and to gain a first insight into the effects of familiarity, as defined by the research questions in the preceding subsection. The pilot study was conducted online via the PennController for IBEX (PCIbex) (Zehr & Schwarz, 2018). Throughout the study, L1 English participants completed a number of tasks that included stimuli recorded with two Tyneside English (TE) and two New Zealand

[3] This research question was of secondary interest. It was included to investigate the well-attested positive effect of word frequency on lexical processing (e.g. Grainger, 1990; Whaley, 1978).

English (NZE) speakers. The Tyneside English speakers are referred to as TE_1 and TE_2. The same reference system is used for the NZE speakers: NZE_1 and NZE_2. In the following, the participants, stimuli, procedure, data analysis steps and results of the pilot study will be considered.

3.2.1 Participants

3.2.1.1 Demographic Information

L1 English speakers, regardless of their regional origin, were recruited for the pilot study.[4] Altogether, 34 participants completed the pilot study. They were partly recruited from the Speech and Language Sciences student pool within the School of Education, Communication and Language Sciences (ECLS) at Newcastle University. The remaining participants were either sourced from Prolific (Prolific Team, 2023)) or the supervisory team of this project. The participants received SONA credits or payment for completing the study. Two participants were excluded because they did not pass the attention check (see subsection 3.2.3). Another two participants were excluded because of technical difficulties and because they reported strong fatigue. For two participants, no results were saved. Three further participants were excluded because English was not their L1.

Some participants came from the Tyneside region in North East England. Their data were not analysed here but for the main experiment because the selection criteria for the latter were stricter and, thus, participants were more difficult to recruit (see subsection 3.3.1). As a result, the data of 19 participants was analysed for the pilot study. All participants identified as female at the time of the pilot study. The mean age of the participants was 24.3 years (*sd* = 7.8 years).

3.2.1.2 Familiarity Measures

Two types of measures were used to assess the participants' familiarity with TE and NZE: first, their performance during the accent matching task and; second, their self-reports from the demographic and language background questionnaire at the end of the experiment. Detailed information on these parts of the pilot study is provided in subsection 3.2.3. Briefly, the accent matching task asked participants to select the regional origin of the four speakers included in

[4] See Appendix A for the participant information sheets, declarations of informed consent and debriefs used in the experiments.

the study. During the questionnaire, the participants indicated the size of their social networks in Newcastle and New Zealand and how familiar they were with the accents spoken there.

As can be seen in Figure 3.1, accent matching was better for TE than NZE. TE speakers were misidentified as coming from Birmingham (50.0% of incorrect responses), Liverpool (35.7%) and London (14.3%). For the NZE utterances, the choice of incorrect answers included Australia (78.3 %), London (17.4%) and the United States (4.3%). A simple logistic regression model with accent as the only predictor showed that accent matching accuracy was significantly worse for the NZE speakers, β = -0.97, SE = 0.47, p = .041.[5]

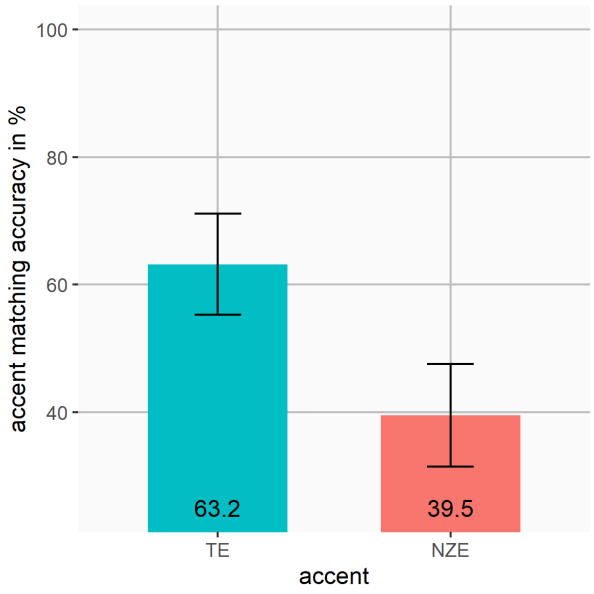


Figure 3.1: Accent matching accuracy as mediated by accent (bars: mean, error bars: ±one SE, colour: accent)

Information on the participants' self-reported familiarity with TE and NZE as well as their social network can be found in Table 3.1. All scales had seven steps and ranged from a minimum of 0 to a maximum of 6 (see subsection 3.2.3.6). As the reported means show, familiarity ratings were generally higher for TE than NZE. Social network ratings were more conservative but followed the same pattern in that, on average, participants indicated that more members of their social network were from Newcastle than New Zealand. In fact, 75% of the participants indicated that their social network consisted of no or very few New Zealanders and only one participant chose a value on the upper end of the scale for this question. For both

[5] The low accuracy rates might partly have been caused by the fact that the participants were not told explicitly that they could use each label more than once. The participants potentially found one of the TE and NZE speakers more representative of the corresponding accent, and, as a result, were apprehensive about reusing the label. This was rectified for the main experiment (see subsection 3.2.6).

measures simple linear regression models, which only included accent/location as a predictor, showed significantly lower values for NZE (familiarity: β = -2.42, SE = 0.55, $p < .001$; social network: β = -2.16, SE = 0.53, $p < .001$). These results were unsurprising since most participants were students at Newcastle University. Granted the caveats of self-reported familiarity on numerical scales, this shows that the participants in the pilot study were overall more familiar with TE as compared to NZE, which was also reflected in the accuracy rates for the accent matching task.

	Newcastle (English)[13]	**New Zealand (English)**
Familiarity with Accent	3.9 (1.7)	1.5 (1.6)
Social Network	2.7 (2.1)	0.5 (1.1)

Table 3.1: Self-reported familiarity and social network information: mean and standard deviation from seven-point Likert scales (0 to 6)

3.2.2 Stimuli

3.2.2.1 Recording Materials

3.2.2.1.1 Sentences for Lexical Decision Task

The sentences for the lexical decision task were taken from Stringer and Iverson (2020), who developed three different versions of 439 sentences in an effort to create a new set of materials for research into speech processing by L2 participants. Hence, their sentences complied with the B1 level of the Common European Framework of Reference for Languages both lexically and syntactically (Stringer & Iverson, 2020, 562).

While Stringer and Iverson (2020) developed materials primarily for L2 listeners, they provided a suitable resource for the lexical decision task and, as will be shown below, were also tested on L1 listeners. Stringer and Iverson's (2020, 563-566) three versions of each sentence differed primarily with regards to the predictability of the final word from the sentential context. It was either predictable from the semantic context of the sentence, neutral in terms of its predictability or semantically anomalous with the preceding sentence (see Table 3.2 for two example sets). Stringer and Iverson (2020) occasionally changed the wording of individual sentences to make them grammatical or semantically more plausible (see the neutral sentence from Example Set 2 in Table 3.2). The semantic context was construed by two to three content words in the sentence, referred to by Stringer and Iverson (2020, 563) as "pointer

words" (underlined in Table 3.2) . Predictability was determined through cloze tests with both L1 and L2 speakers of English. For the current pilot study, neutral sentences were chosen to keep the predictability of the target at the end of the sentence low. Specifically, only those neutral sentences were included with a target that exhibited a cloze probability of $\leq 30\%$. This means that the likelihood that L1 or proficient L2 speakers of English completed the sentence in question with that target after reading the sentence frame was not higher than 30%. Each sentence further contained exactly two pointer words and all target words were monosyllabic.

	Example Set 1	**Example Set 2**
sentence frame	You should put your rubbish in the	For breakfast children eat toast or
predictable	You should put your rubbish in the **bin**.	For breakfast children eat toast or **cereal**.
neutral	You should put your tickets in the **bin**.	For dinner students sometimes eat **cereal**.
anomalous	You should put your rubbish in the **rail**.	For breakfast children eat toast or **literature**.

Table 3.2: Example sets from Stringer and Iverson's (2020) materials (underlined: pointer words, bold: targets)

Of Stringer and Iverson's (2020) neutral sentences, 68 that fulfilled these criteria were chosen at random within the 'tidyverse' package (Wickham & RStudio, 2023) in the R programming environment (R Core Team, 2023). Stringer and Iverson's (2020) set of sentences provided information on the phoneme and letter counts as well as the phonological neighbourhood density of the targets. This information was fed into the ARC nonword database (Rastle, Harrington, & Coltheart, 2002) to generate a set of 68 nonwords. Only monosyllabic nonwords with orthographically permitted sequences of letters were chosen to closely match the words from the original sentences. The nonwords were then pasted to the sentence frames of the 68 neutral sentences so that the final word (real word or nonword) was the only difference within each pair of sentences. Once compiled, the nonword sentences were inspected to avoid nonwords that, given the context of the sentence, could be mistaken for real words easily. Ten nonwords had to be adapted for that reason, usually by replacing one of their sounds or by adding another sound. For example, the nonword *ciff* was altered to *ciffp* because the former could be mistaken as a cleaning product brand, given the preceding sentence *I went to the supermarket to buy a*. In two cases, altering the nonword resulted in bisyllabic nonwords (*ropu* and *waty*), which was deemed acceptable since only the responses to sentences with real words were evaluated and sentences with nonwords merely acted as fillers in the lexical decision task.

3.2.2.1.2 Short Stories for Recall Task

The stimuli for the recall task followed Sturt et al. (2004) in that they consisted of triplets of stories, two of which were presented to the participants during a given trial. Each story was three sentences long. In total, 54 story triplets were constructed: 24 for experimental trials, 24 for filler trials and 6 for practice trials. The three versions of each story differed in a single word. The word change was most constrained for the experimental triplets: It occurred in the second sentence of the story and the replacement word was either semantically related or unrelated to the original word. Examples are given below for the original (3a), related (3b) and unrelated (3c) version of an experimental triplet:

(3a) *Sam and Kate made an interesting discovery yesterday. They found a small chest full of* **rubies** *in their attic. They had no idea where it came from or who put it there.*

(3b) *Sam and Kate made an interesting discovery yesterday. They found a small chest full of* **diamonds** *in their attic. They had no idea where it came from or who put it there.*

(3c) *Sam and Kate made an interesting discovery yesterday. They found a small chest full of* **drugs** *in their attic. They had no idea where it came from or who put it there.*

The words were chosen so that the original and related words could be subsumed under one hyperonym (*gems* for (3a) and (3b)), which does not include the unrelated words. An overview of the original, related and unrelated words can be found in Table 3.3. The target words were intended to be concrete nouns, which is the case for practically all nouns listed although *technology* and *decoration* might be considered abstract. The unrelated word *phone* occurs twice (sets 6 and 21) and *toilet* is once used as the original word (set 11) and once as the unrelated word (set 19). However, as will be shown in subsection 3.2.3.3, the participants did not necessarily hear the unrelated versions of these sets and certainly did not hear them in the same accent. Table 3.3 further shows the accent that the tokens were presented in.

	Original Word	Related Word	Unrelated Word	Hyperonym	Accent
1	crocodiles	alligators	elephants	reptiles	TE
2	arena	stadium	pub	big venues	TE
3	magazines	newspapers	technology	periodical publications	TE
4	laptop	computer	window	computing machines	TE
5	beer	cider	decoration	drinks	TE
6	purse	handbag	phone	bags	TE
7	ginger	garlic	wine	spices	TE
8	jackets	coats	boots	upper body clothing	TE
9	pillows	blankets	drawers	bedding	TE
10	Facebook	Twitter	TV	social media	TE
11	toilet	bathroom	hall	rooms for personal hygiene	TE
12	dress	skirt	book	clothing items	TE
13	doctor	nurse	teacher	medical professionals	NZE
14	rubies	diamonds	drugs	gems	NZE
15	stool	chair	box	items for sitting	NZE
16	kitten	puppy	snake	common pets	NZE
17	pizza	pasta	sushi	Italian food	NZE
18	mango	banana	smoothie	fruits	NZE
19	printer	scanner	toilet	reduplication devices	NZE
20	basil	parsley	juice	herbs	NZE
21	hand	arm	phone	body parts	NZE
22	steel	iron	plastic	metals	NZE
23	football	basketball	music	sports	NZE
24	lighter	matches	water	ignition items	NZE

Table 3.3: Overview of changed words in experimental story triplets

The purpose of the filler triplets was, primarily, to generate an equal number of changed and unchanged stories and, secondarily, to distract participants from the fact that the change in the experimental items always occurred in the second sentence. Hence, the first or the third sentence was altered for the filler stories. As can be seen in the example filler stories below, no intent was made to control the semantic distance between 'original', 'related' and 'unrelated' words:

(4a) *This report is* **very** *relevant for the company's new project. They are currently finalising their first prototype. With this report, they might be able to speed up the process.*

(4b) *This report is* **highly** *relevant for the company's new project. They are currently finalising their first prototype. With this report, they might be able to speed up the process.*

(4c) *This report is* **extremely** *relevant for the company's new project. They are currently finalising their first prototype. With this report, they might be able to speed up the process.*

(5a) *I'm more than ready for a relaxing holiday. It's been a while since I last took some days off. The week in* **Ireland** *will recharge my batteries.*

(5b) *I'm more than ready for a relaxing holiday. It's been a while since I last took some days off. The week in* **Scotland** *will recharge my batteries.*

(5c) *I'm more than ready for a relaxing holiday. It's been a while since I last took some days off. The week in* **Wales** *will recharge my batteries.*

In addition, 6 story triplets were constructed for practice purposes in the experiment. Half of these 6 triplets were constructed according to the rules for experimental triplets while the other half emulated fillers.

3.2.2.1.3 Sentences for Transcription and Accent Matching Task

An additional 30 sentences from Stringer and Iverson (2020) were included for the transcription and the accent matching task. The accent matching task was used as a measure of familiarity (see subsection 3.2.1.2) while the transcription task was included to trial its format for further experiments on intelligibility. In line with the sentences for the lexical decision task, the cloze probability of these sentences was $\leq 30\%$. However, the final words of each sentence had more than one syllable and no sentences with nonwords were constructed.

3.2.2.2 Accent Characteristics and Recording Procedures

3.2.2.2.1 Tyneside English

Since this research was conducted at Newcastle University, TE was chosen as one of the accents because local participants could be assumed to be highly familiar with it. TE is spoken in the Tyneside conurbation, which is part of North East England. A key geographic feature of the conurbation is the name-giving River Tyne, which has an important geographic reference function in that it separates, for example, Newcastle and Gateshead (Pearce, 2009, 173). The Tyneside conurbation has a population of about 867,680 inhabitants, the largest part of which is contributed by Newcastle with approximately 303,200 inhabitants (Office for National Statistics, 2021). From a linguistic point of view, Newcastle and the *Geordie* accent play a strong representative role for the Tyneside conurbation, in particular for non-Tynesiders.

In-depth descriptions of the vowel and consonant system of TE can be found in Beal, Elizondo, and Llamas (2012, 26-37), Foulkes and Docherty (1999), Mearns (2015, 167-170), Warburton (2020, 51-53), Watt (1998, 203-251), Watt and Allen (2003), Watt and Milroy (1999, 28) and Wells (1982, 374-376). Briefly, important characteristics of TE are the absence of the FOOT-STRUT split and BATH lengthening and retraction. In these regards, TE patterns with other northern English varieties. FACE and GOAT exhibit a number of phonetic variants, with the monophthongal variants [eː, oː] being the most prevalent. With regards to consonants, the voiceless plosives /p, t, k/ in TE have been covered extensively in the literature as their phonetic realisation includes glottalised ([ʔ]) and glottally reinforced ([ʔ͡p, ʔ͡t, ʔ͡k]) variants.

The stimuli were recorded with two middle-aged female speakers of TE in October and November 2020. As mentioned in section 3.2, the two speakers will be referred to as TE_1 and TE_2 throughout this thesis. Both speakers were in their late forties when the recordings were conducted and had been living in or near Newcastle most of their lives. Their accents were deemed representative of TE by the supervisory team. The speakers were recruited via social media and were reimbursed for their time with a gift voucher. In order to guarantee high-quality recordings, the speakers were recorded on campus in ECLS. Altogether, the recording sessions lasted around 75 to 90 minutes.

Figure 3.2: Recording setup for TE speakers

The speakers read the sentences for the lexical decision and the transcription task as well as the short stories for the recall task from a computer screen. PCIbex was used to present the text and the speakers moved on to the next sentence or story via mouse click. They were encouraged to read the stimuli as many times as necessary to achieve a version without hesitations and any emphasis on a particular word, for example the final word in the sentences for the lexical decision task. Before sentences with nonwords were presented, the speakers saw a screen with the nonword and a pronunciation guide for the nonword. For almost all nonwords, this guide was a real English word that rhymed with the nonword (e.g. *French* for *crench*). The speakers could practise the nonword and then moved on to the full sentence with the nonword (e.g. *The stuff is on a white crench*). The sentences were split into two blocks (86 in the first block and 80 in the second block). The stories were split into three blocks of 54 stories each. The stories in the three blocks were identical, except for the changed word (see Table 3.3 for experimental triplets). The blocks were separated by breaks of at least two minutes.

For the recordings, an Edirol R09HR digital recorder and a Sennheiser radio microphone were used. The stereo recordings were made in the wav format at a sampling and digitisation rate of 44.1 kHz and 16 bit, respectively. The noise level of the microphone was set to -30 dB as background noise was not picked up at this level. As can be seen in Figure 3.2, the microphone was clipped to a microphone stand and slightly elevated to decrease the distance

between speaker and microphone. The researcher checked the recording quality throughout the session with headphones connected to the recorder.

Recordings in the sound-attenuated booth in ECLS were not possible at the time due to COVID-19 restrictions. Instead, a quiet and small room in ECLS was chosen that had also been used for research into parent children interactions. Except for the desk shown in Figure 3.2, the only other furniture consisted of a table with four chairs. This setup and the acoustics of the room were similar to the one that the NZE recordings were conducted in (see subsection 3.2.2.2.1). Next to the desk, there is a small booth with a one-way window. The researcher sat in said booth during the recording to make sure that the audio setup was functioning as planned. Another advantage of the researcher not being close to the speakers was that they felt less observed. The researcher also checked that the speakers produced the utterances as required. This was critical for the recall task since the stories should differ in a single word only. Any recordings that were problematic were re-recorded with the researcher present so that he could let the participants know what the problem was. For TE_2, all stimuli could be recorded within a single recording session. TE_1 came in for recordings twice. The second recording session for TE_1 lasted around 60 minutes.

3.2.2.2.2 New Zealand English

Following the same rationale that was used for including TE in this study, NZE was selected because future participants were assumed to be less familiar with it. The choice of this accent as the unfamiliar accent was also a practical one as COVID-19 case numbers were close to zero in New Zealand at the time, which allowed researchers to access the lab and recording equipment. NZE had a speech community with approximately 4.48 million members in 2018 (Stats New Zealand, 2021). Although New Zealand is larger in size than the United Kingdom (UK), (European) NZE is described by Bauer and Warren (2008, 40) as "remarkably homogeneous", which is in sharp contrast to the many locally distinguishable dialects in the UK. With regards to phonetic and phonological variation, which is the chief concern for the current research, this is not to say that all New Zealanders have the same accent. The Southland accent, for example, is identifiable through its rhoticity following NURSE and letter (Hay & Maclagan, 2008, 98-99). Other distinctive features of this accent, such as differentiating between [w] and [ʍ], are restricted to older speakers from the region. In fact, Hay and Maclagan (2008, 99) argue that lexical variation is more important in distinguishing the Southland variety and that, therefore, the use of the term *dialect* rather than *accent* is more suitable.

Regional variation aside, the vowel system of NZE is, in comparison with TE, characterised by the merging of NEAR and SQUARE, the central production of START as [a], the raising of TRAP to [ɛ], the raising of DRESS to [e], the centralisation of KIT to [ə] and, finally, the fronting and rounding of NURSE to [øː] (Bauer, Warren, Bardsley, Kennedy, & Major, 2007, 98).[6] Glottaling is more restricted in NZE in that it typically affects only /t/ word-finally. Glottal reinforcement can further be found for /p, k, tʃ/.

The recordings of the NZE speakers were conducted at the University of Canterbury in Christchurch, New Zealand in November and December 2020.[7] In line with the nomenclature for TE speakers, the NZE speakers will be referred to as NZE_1 and NZE_2 throughout this thesis. Both speakers were in their mid to late forties at the time of the recordings. They were female and of European heritage. They were born and raised in the Christchurch region although NZE_2 had spent four years in the United States. NZE_2's stay there, however, took place more than twenty years prior to the recording. Both NZE_1 and NZE_2 had identifiable NZE accents with, for example, raised productions of the DRESS and TRAP vowels.

As with the TE speakers, the sentences and stories were presented to the NZE speakers via PCIbex in five blocks with intermittent pauses. PCIbex was also used to inform them about the study and to secure their consent. Prior to the recording, the speakers were given examples of good and bad recordings. The main aim of this was to avoid misreading and productions that contained a lot of hesitations. The speakers were encouraged to produce the stimuli as many times as necessary. Recordings were made in Audacity with a head-mounted Beyerdynamic Opus 55.18 MK II microphone. The sampling rate was 44.1kHz. The file format was wav-16bit mono. The recordings for NZE_1 and NZE_2 were conducted in a room at the New Zealand Institute of Language, Brain and Behaviour of comparable size and acoustics to the room used in ECLS.

Unfortunately, there was some noise due to electrical interference on the recordings of NZE_2. Although a good amount of this noise could be removed with a noise reduction algorithm, the quality of certain stimuli remained poorer compared to the recordings of TE_1, TE_2 and NZE_1. Given the limited possibility of obtaining high-quality recordings due to the COVID-19 pandemic at the time, the pragmatic approach was taken to work with the available recordings of sufficient quality. As will be explained in more detail in the following subsections,

[6] This fronted and rounded variant of NURSE is also a variant in TE. Watt and Milroy (1999), for example, found that [øː] was frequently used by younger female speakers. Overall, however, NURSE has a more central quality [ɜː] in TE.

[7] Donald Derrick conducted the recordings with the NZE speakers. Special thanks go to him and his team for their patience and guidance throughout the process.

this resulted in a split of stimuli by accent such that, for example, one half of the stories for the recall task was presented in TE only while the other half was only presented in NZE (see rightmost column in Table 3.3).

3.2.2.3 Processing of Recordings

3.2.2.3.1 Sentences for Lexical Decision Task

The recordings for all tasks were processed in Audacity (Audacity Team, 2023) and Praat (Boersma & Weenink, 2023). First, the TE recordings were converted to mono. Next, the recordings for the lexical decision task were annotated by hand on a separate tier in Praat. Annotation was conducted at the sentence level and mainly with the help of the waveform. Spectrograms were consulted occasionally to determine the exact beginning and end of the sentences. The segments on the 'sentence' tier were labelled according to the (non)word at the end of the sentence. After segmentation, all labels were checked and, if necessary, corrected with a Praat script by DiCanio (2011). A Praat script by Lennes (2002) was subsequently used to save each sentence into a separate wav file.

The quality of all extracted files was checked next, which was essential for two reasons. First, there should be no structural differences between the recordings of each speaker to avoid potential confounds further on. Second, due to re-recording (see subsection 3.2.2.2.1), there were multiple recordings available for some of the sentences produced by the TE speakers and the best one had to be selected. To that end, all sentences were coded based on quality. The productions should not contain any hesitations, false starts, background noise or misreading. If any of the latter occurred in the recordings, the quality was marked down accordingly. Recordings were only included if they passed the quality criteria. 56 sentences were chosen for each speaker. The amplitude of these sentences was normalised in Audacity.

3.2.2.3.2 Short Stories for Recall Task

To prepare the recall stimuli for the experiment, all TE recordings were again converted from stereo to mono first. As explained in subsection 3.2.2.1.2, each story for the recall task consisted of three sentences. Three separate tiers were created in Praat to annotate the sentences by hand. As for the lexical decision stimuli, the sentences were annotated mainly based on the waveform although the spectrograms were examined more closely when necessary. The segments on each tier were labelled and mistakes in the labelling of the segments were

again corrected with DiCanio's (2011) Praat script, before each sentence was saved into its own wav file with Lennes's (2002) Praat script.

The next step was a quality check of the recordings. Since the three different versions of each story differed only in a single word, two of the three sentences were identical within each set. These sentences will henceforth be referred to as the 'frame' of a story. The speakers read each version of the story in full and, for TE_1 and TE_2, there were multiple recordings of some versions due to re-recordings and the number of recording sessions. Therefore, there were multiple versions of the same sentence, especially for the frame sentences and the TE speakers.

To make the versions within each set of stories as uniform as possible, the frame consisted of the same recordings for a set of stories. For example, the first and third sentence of the stories in (3a) through (3c) (see subsection 3.2.2.1.2) were the same recordings in each version of the story for a specific speaker. The second sentence, which contained the actual change, was spliced into this frame. The aim of this approach was to avoid the introduction of confounds by including frames of varying quality across the different versions of each story.

The quality of each recorded sentence was evaluated. Similar to the lexical decision stimuli, the quality was marked down if the productions were hesitant or contained misreadings, false starts or background noise. Sentences were used to make the recall stimuli if their quality was acceptable or could be edited accordingly in Audacity. The first important step here was to ensure that the productions of the target sentences, that is the sentences with the changed words, were usable. For the task to work, it was essential that these sentences were identical except for the intended word change. Any other differences between the sentences, even very minor ones such as using a contraction for auxiliary verbs, would have compromised the validity of the task. For TE_1 and TE_2, there were several versions of identical quality for a specific target sentence. In those cases, one production was chosen at random in R.[8]

Afterwards, the two sentences for the frame were determined. Here, several versions were available for all speakers since every speaker produced the frame of a specific story triplet (at least) three times. Where possible based on recording quality, the sentences of the frame were chosen to contain productions from different versions of the story. In most cases, this was more feasible than obtaining a usable frame from a single version of a story. Furthermore, if

[8] For a very limited set of stories, there are very minor differences in the wording of the stories across the speakers. For example, TE_1 struggled with the initial consonant cluster in *sceptical*, which is why *irritated* was used for this speaker in this context. Importantly, these changes never affected the target words. These changes are documented in Appendix B.2.

a frame from a single version was used (e.g. both sentences from the related version of the story), then there would be one version, in this case the related one, which would contain no splicing altogether. This might lead to structural differences between the versions, which was avoided by the current approach. If more than one suitable production was available for a specific frame sentence, one production was chosen at random in R. There were only very few instances of frames compiled from the same version of the story.

Once the target sentence for each version of the story and the two sentences of the frame had been determined, the corresponding wav files were concatenated with a Praat script by Daidone (2017). A pause of 400 ms was added between each sentence to make the stories sound more naturalistic. For the experiment to work as intended in PCIbex, a duplicate of the original version of each story was created and called 'unchanged'. Finally, 20 stories (12 experimental tokens and 8 fillers) were chosen for each accent (see subsection 3.2.2.1.2 and Table 3.3 for more detail). Audacity was again used to normalise the amplitude of the stories.

3.2.2.3.3 Sentences for Accent Matching and Transcription Task

The stimuli for these tasks were taken from Stringer and Iverson (2020) (see subsection 3.2.2.1.3). For the accent matching task, four audio files of sufficient quality (one per speaker) were chosen. Another sentence was recorded by the researcher and served as an attention check in the study (see subsection 3.2.3.4). For the transcription task, 8 sentences were chosen, which were available in sufficient quality for all speakers. This resulted in a total of 32 audio files, which were normalised in Audacity. None of the sentences used in these tasks were part of the lexical decision task.

3.2.3 Procedure

3.2.3.1 Headphone Check

Since the study was conducted online, it was more difficult to create a controlled environment. To ensure that the participants listened to the audio stimuli at a sufficient volume, paid attention to them and wore headphones, they first completed a headphone screening. The screening was developed by Woods, Siegel, Traer, and McDermott (2017) and consisted of six trials, during which the participants had to identify the quietest in a series of three tones. Importantly, one of these tones has "a phase difference of 180° between stereo channels [...] [and is] heavily attenuated when played through loudspeakers" (Woods et al., 2017, 2065)

but not over headphones. As a result, amplitude discrimination is only reliable when wearing headphones, as demonstrated by Woods et al. (2017) in both online and in-lab studies. The participants in the current study had to identify the quietest tone correctly at least five out of six times to pass the headphone check.

3.2.3.2 Lexical Decision Task

The participants heard 28 sentences produced by the four speakers for a total of 112 sentences. 56 of Stringer and Iverson's (2020) sentence frames were used and sentence frames were not repeated within accents. Thus, if a participant heard the word version of a sentence frame in TE (e.g. *My kids are playing with my* **niece**), they would hear the nonword version in NZE (e.g. *My kids are playing with my* **jadge**) and vice versa.[9] As a result, one specific sentence was only ever presented in either TE or NZE. This approach was chosen so that participants would not hear the same sentence frame twice in TE or NZE.

During each trial, the participants first saw fixation crosses for 500 ms, then heard the sentence and, finally, indicated the lexical status of the sentence-final target via key press. The participants were instructed to keep their index fingers on the 'M' (real word) and 'Z' (nonword) keys throughout the task. Participants were told that they would hear different speakers throughout the task to preempt surprise effects due to speaker changes as much as possible (Floccia et al., 2009, 392-400; see subsection 3.1.1). The participants were asked to respond as accurately and quickly as possible. For each speaker, the participants heard no more than three word or nonword sentences in a row. Which sentences were produced by the one or the other speaker of each accent was selected pseudorandomly by PCIbex. In addition to the 112 experimental items, the participants had the chance to practise the task with ten example sentences (5 word and 5 nonword sentences, no identical sentence frames). The wording of these sentences was identical for all participants but the speaker varied across groups. The participants heard the practice sentences from the first speaker that they heard during the task.

Participants were assigned iteratively to four different groups. The groups differed in terms of which speaker occupied which position throughout the experiment (see Table 3.4). The order of speakers followed a Latin square design such that each speaker occupied each slot once across the groups. As a result, potential learning and fatigue effects were mitigated because

[9] See Appendix B.1 for a complete list of the sentences used.

they affected the responses to each speaker to equal extents. The speakers were grouped together by accents such that the participants heard all sentences in TE or in NZE first.

Group	Practice	Speaker 1	Speaker 2	Speaker 3	Speaker 4
A	TE_1	TE_1	TE_2	NZE_1	NZE_2
B	TE_2	TE_2	TE_1	NZE_2	NZE_1
C	NZE_1	NZE_1	NZE_2	TE_1	TE_2
D	NZE_2	NZE_2	NZE_1	TE_2	TE_1

Table 3.4: Order of speakers across conditions in the lexical decision, recall and transcription tasks

In summary, the participants heard 112 sentences, 56 in each accent. The 112 sentences had 56 sentence frames, which were never repeated within accents. Within the 56 sentences for each accent, 28 were produced by one speaker and 28 were produced by the other speaker. For each speaker, there were 14 word and 14 nonword trials and there were never more than three trials of either type in a row. The allocation of the sentences to each speaker within an accent was pseudorandomised in PCIbex. The presentation of accents was blocked but the specific order of speakers was varied via the Latin square design shown in Table 3.4. Prior to the 112 experimental items, the participants heard 10 practice sentences with an equal number of word and nonword trials. The speaker that these practice items were produced by was identical to the first speaker they encountered during the experimental trials.

3.2.3.3 Recall Task

For one trial of the recall task, two stories were chosen from the story triplets described in subsection 3.2.2.1.2. The participants completed 10 trials for each of the four speakers for a total of 40 trials. The presentation of speakers followed the scheme in Table 3.4. As mentioned in subsection 3.2.2.2.2, there were some problems with the recording quality for NZE_2. Hence, 20 story triplets of good quality were chosen for each accent. 12 story triplets were experimental items, for which the change always occurred in the second sentence and the semantic distance between the original and the (potentially) changed word was controlled. The remaining eight triplets served as fillers, for which the change occurred in either the first sentence or the third sentence of the story.[10]

[10] See Appendix B.2 for a complete list of the stories used.

For each trial, the participants first saw fixation crosses for 500 ms and then heard the 'original' story. The fixation crosses then flashed for another 500 ms, followed by the second story. The second story was either identical to the first one ('unchanged' condition) or differed from it in a single word. In the 'related' condition, this changed word was semantically related to the original one (e.g. *doctor* → *nurse*). In the 'unrelated' condition the semantic distance between the changed and the original word was larger (e.g. *doctor* → *teacher*). The participants indicated via key press if there was a change between the stories. They indicated identical stories with the 'M' key and differing stories with the 'Z' key. If they indicated that the stories were identical, they immediately went on to the next trial. If they indicated that the stories were different, the second story they heard was displayed on their screen, along with two input boxes (see Figure 3.3). They were asked to identify the word that was changed in the second story and the corresponding word from the original story. The participants were instructed to do the initial key press as quickly and accurately as possible.

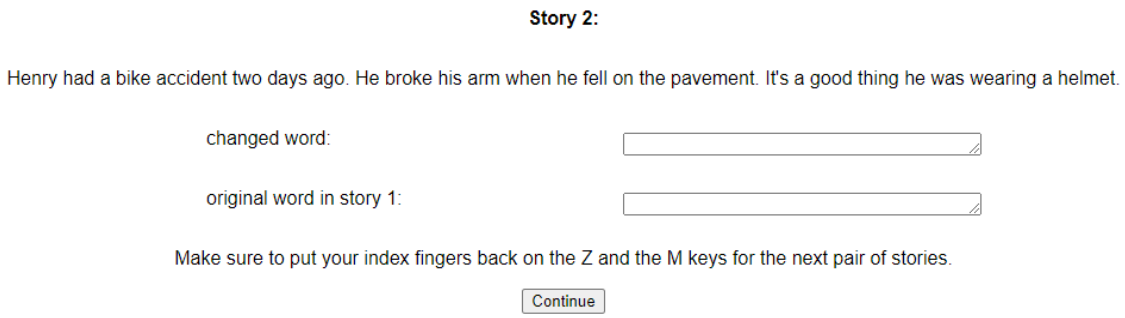


Figure 3.3: Correction screen for recall task

PCIbex pseudorandomly selected which of the available experimental and filler items were produced by the first and the second speaker of an accent. A breakdown of experimental and filler items for a single speaker is provided in Table 3.5. Most importantly, the likelihood of a change not occurring (2 unchanged experimental tokens and 3 unchanged fillers) was equal to the likelihood of a change in the stories (4 changed experimental tokens and 1 changed filler). Although there was an imbalance as to where the changes occurred (higher likelihood of a change in the second sentence), the difficulty of the task and the inclusion of a filler item with a change in the first or third sentence should have disrupted a strong learning effect. The consistent location of the change was further made more difficult by the sentences without a change.

Experimental Items (potential change in sentence 2)			**Filler Items** (potential change in sentence 1 or 3)	
unchanged	*related*	*unrelated*	*unchanged*	*changed*
2	2	2	3	1
total = 6			*total* = 4	
grand total = 10				

Table 3.5: Breakdown of experimental and filler items for one speaker in the recall task

The participants practised the task with four practice story pairs. The first and the last pair of stories included a change while both stories were identical for the second and third pair. In the first pair, the change occurred in the second sentence. In the last pair, there was a change in the third sentence, which established from early on that the change would not always occur in the second sentence. Following the scheme set out in Table 3.4, the practice stories were produced by the same speaker that was used as the first speaker for the actual recall task. In line with the lexical decision task, the presentation of TE and NZE speakers was blocked.

3.2.3.4 Accent Matching Task

The participants heard four sentences and identified the origin of the speaker from a list of nine labels.[11] For each sentence, the participants were instructed to choose where the speaker was from and, if unsure, to choose the most likely option. They were further told that they would not need each label. The participants could listen to each sentence as many times as necessary and all sentences were available on the same screen so the participants could compare them. The labels provided were: *Glasgow, Liverpool, Birmingham, London, Newcastle, United States, Australia, New Zealand* and *Germany*. The last label was provided because there was another audio file that served as an attention check. Specifically, the participants heard the following: 'This trial is to make sure that you are paying attention. Please select *Germany*.'

3.2.3.5 Transcription Task

The transcription task was less relevant for the experiment on familiarity. However, it was included in the pilot study because its format would become relevant for Experiment 2, when the effect of intelligibility on lexical processing and recall was assessed. Here, the participants

[11] See Appendix B.4 for a complete list of the sentences used.

transcribed two sentences produced by each of the four speakers for a total of eight trials.[12] The participants were instructed to transcribe each sentence as accurately as possible and to write down their best guess if they were unsure. For each trial, the participants first saw fixation crosses for 500 ms, followed by a single presentation of the audio stimulus. They then transcribed the sentence and moved on to the next one. The order of speakers varied between participants (see Table 3.4) and PCIbex pseudorandomly picked which sentence would be presented in which speaker's voice. TE and NZE speakers were blocked.

3.2.3.6 Demographic and Language Background Questionnaire

After the experimental tasks, the participants were asked for additional demographic data. To measure self-reported familiarity, they indicated on two separate sliding scales how familiar they were with the accent spoken in Newcastle[13] and New Zealand, respectively. The scales had 7 steps (from 0 to 6) and were labelled *very unfamiliar* on the left extreme and *very familiar* on the right extreme. Next, the participants reported how many members of their social network, defined as "family and friends", were from Newcastle and New Zealand. Again, 7-step Likert scales were used, this time with the labels *nobody/very few* and *very many* to the left and right, respectively. The participants were then asked if they had been "born and raised in Tyneside", what the first part of their UK postcode was and how long they had been living in this area. They further entered their age, gender (on a voluntary basis) and whether they were an L1 speaker of English or not. Another attention check was included in this part of the study. Participants were told to type the word *custard* as a response to one of the questions. Finally, the participants were debriefed and asked if, based on the information in the debrief, they provided their consent to have their data included.

3.2.4 Data Analysis

3.2.4.1 Lexical Decision Task

The results were exported from PCIbex and then read into R, where all data wrangling and analysis took place. Only the responses for words were considered for the lexical decision

[12] See Appendix B.3.1 for a complete list of the sentences used.

[13] The previous subsections referred to the accent of TE_1 and TE_2 as Tyneside rather than Newcastle English. Indeed, TE is not restricted to the city of Newcastle but encompasses the Tyneside conurbation. Here, the participants were asked about Newcastle (English) specifically to avoid confusion because of the terminology used. For example, some participants might have been unsure about the extension of the Tyneside conurbation, which could have influenced their responses. Since both speakers were from Newcastle, using this label in the questionnaire remained factually correct.

task, which is why all following considerations apply to word trials only. Nonword trials were not considered because the aim of this task was to test how fast and accurately entries in the mental lexicon could be accessed. The ability to dismiss targets as nonwords was not of primary interest. Accuracy was coded based on the participants' key press. LDLs are not directly measured by PCIbex. Instead, the tool saves time stamps when events occur during the study, such as the start and end of a trial or a behavioural response from the participant. Latencies were, therefore, measured with the following formula:

LDL = time stamp of key press - duration of stimulus - 500 ms - time stamp of trial start

The 500ms in the formula account for the presentation of fixation crosses preceding the stimulus. Stimulus duration was measured directly in R with the 'warbleR' package (Araya-Salas & Smith-Vidaurre, 2017, 2023). This information was then used to calculate the latencies for each trial. Outliers were identified via cut-off values. Following standards in the field, the lower cut-off was set at 250 ms while the upper cut-off was set subject-specifically at 2.5 standard deviations above one subject's mean (see Clopper et al., 2016, 92).

3.2.4.2 Recall Task

The results for the recall task were analysed in two steps. First, the accuracy of the participants' initial decision as to whether the two stories were identical or different was assessed ('key press accuracy').[14] Second, the participants' 'correction accuracy' was considered. As detailed in subsection 3.2.3.3, the participants entered the different word in the second story as well as the original word from the first story into two separate input boxes (see Figure 3.3). These responses were compared against the solutions in R and coded as accurate if both words from the input boxes matched the solutions. Responses were counted as accurate if participants identified the right words but put them into the wrong input boxes, that is if they put the original word from the first story into the input box that asked for the changed word from the second story. Since R simply matched the character strings of the participants' responses and the solutions, typos would automatically be counted as errors. Additionally,

[14] The latencies for this key press were also measured with the following formula:

Latency = time stamp of the button press - duration of the two story stimuli - 1000 ms - time stamp of trial start

The 1000ms in the formula account for the presentation of fixation crosses preceding each of the two stories. However, with the changes occurring in the second sentence of the story and not always as the final word of the third sentence, these latencies were not very informative and will not be considered in the analysis.

some participants added a question mark to their response, most likely to indicate uncertainty, which would also result in the coding of the response as an error. To prevent this, the responses were inspected manually and any typos (e.g. *rubys* for *rubies*) were corrected and additional punctuation marks were removed.

3.2.4.3 Accent Matching Task

For the accent matching task, the label that the participants chose for each of the audio files was compared against the origin of the speaker (either Newcastle or New Zealand) for a total of up to four correct choices. Performance in this task was primarily used as a measure for accent familiarity (see subsection 3.2.1.2).

3.2.4.4 Transcription Task

The aim for this task was to avoid time-consuming hand-coding of the participants' transcriptions. Matching the character strings for the entire sentence was problematic since small deviations, such as an additional space or a single typo, would have resulted in the coding of responses as errors. This task did not intend to test the participants' orthography but how well they understood the semantics of the sentences. Therefore, the pointer and key words in the sentences were used to assess transcription accuracy. The words set the semantic frame in Stringer and Iverson's (2020) stimulus set (see subsection 3.2.2.1.1). Hence, they were good indicators of how well the participants understood the sentence and, thus, how intelligible the different speakers were (see Floccia et al., 2009, 380).[15] Participants' responses were coded as correct if they included all pointer and key words from a sentence. Naturally, this did not solve the problem of spelling errors since the participants might have misspelled these pointer words. However, this approach allowed for automatic coding while avoiding the classification of responses as errors because of incorrectly spelled prepositions, articles or pronouns.

3.2.4.5 Statistical Analysis

The results for accuracy measures and LDLs were analysed with (generalised) linear mixed effects models in R with the 'lme4' package (Bates et al., 2023). This subsection details which fixed and random predictors were considered for each of the tasks in the experiment and gives reasons for the choice of each predictor. In general, significant predictors were identified via

[15] Stringer and Iverson (2019) used a similar measure in their sentence-recognition task (see subsection 2.5.1.2).

log-likelihood comparisons of nested models. To facilitate this, the `mixed()` function from the 'afex' package (Singmann et al., 2023) was used. For significant predictors, χ^2 and p values will be reported from these likelihood ratio tests. As will be shown below, there was a theoretical motivation to include all predictors in the models. Therefore, even if a predictor did not reach significance, it was kept in the model. Where necessary, the 'emmeans' package (Lenth et al., 2023) was used to conduct pairwise comparisons. Bonferroni corrections were applied for multiple pairwise comparisons to mitigate the increased chance of type I errors (Winter, 2019, 175-177).

3.2.4.5.1 Lexical Decision Task

Four fixed effects and one interaction were included for the analyses of both accuracy and LDLs in the lexical decision task:

- **Accent** was categorical and had two levels: 'TE/familiar accent' and 'NZE/unfamiliar accent'. This predictor was necessary to test the influence of accent familiarity on lexical processing. 'TE' was used as the reference level for this predictor.
- **Block** referred to the trial number in question. In total, there were 28 trials for each of the four speakers. Block was a categorical predictor and coded if the trial in question was part of the first 14 trials in a specific speaker's voice ('first half') or the second 14 trials ('second half'). It was included to detect adaptation effects, that is to say changes in accuracy and/or LDLs while listening to one speaker. The reference level for this predictor was 'first half'.
- The interaction between **accent** and **block** was included to track potentially differential improvements for either accent throughout the task. As previous research has shown, participants might initially perform worse for the unfamiliar accent but could improve over time (see subsection 3.1.1).
- **Word frequency** was a continuous predictor and was included to see if participants' performance was better for more frequent targets, as has been shown for LDLs by, for example, Grainger (1990) and Whaley (1978). While this predictor was only of secondary interest, it was included in an effort to replicate the widely attested effect of word frequency on lexical processing. The word frequency measures were part of the stimulus set by Stringer and Iverson (2020), who consulted the SUBTLEX database (Brysbaert & New, 2009). Word frequency (in occurrences per million words) was log10-transformed

and standardised before being added to the model. The log10-transformation is often used for word frequency (see Van Heuven, Mandera, Keuleers, & Brysbaert, 2014) and the standardisation was performed to allow for easier comparisons across different continuous predictors (Winter, 2019, 89).

- **Speech rate**, finally, was a continuous predictor and operationalised as the number of syllables per second. This factor was mainly included as a control factor to account for differences between the four speakers' voices used in the experiment. Naturally, the voices of the four speakers differed in several aspects beyond speech rate (see subsection 6.7.2). Hence, an alternative to account for the difference between the voices would be the use of a random predictor 'speaker'. However, this lead to singularity issues in the models. Therefore, a fixed predictor was used instead. Speech rate was also standardised.

Two random effects were included:

- Since repeated measures were taken for this task, random intercepts for **participants** were included in the model. If random slopes for accent by participant did not result in singularity issues, they were also included in the model.
- Following the same rationale, random intercepts for the trial targets or **items** were added to the model. Since these items were not repeated across accents, there was no need for random slopes.

There is a case to made for the inclusion of **speaker** as a random effect. However, as has been explained above, this was not possible due to singularity issues. Instead, speech rate was added to the model.

3.2.4.5.2 Recall Task

Four fixed effects and one interaction were included for the analyses of key press accuracy and correction accuracy in this task:

- **Accent** was categorical and had two levels: 'TE/familiar accent' and 'NZE/unfamiliar accent'. This predictor made it possible to test for the effect of familiarity on recall. The reference level was 'TE'.

- **Story type** referred to the (potential) changes between the two stories. It was a categorical predictor and had three levels: 'unchanged', 'related' and 'unrelated'. Story type was included to test the effect of semantic proximity between the changes on recall. Potential differences could be indicative of less-detailed processing or good-enough representations. The reference level used here was 'unchanged'.

- The interaction between **accent** and **story type** was added to track potential structural differences in the effect of different story types on the processing of familiar versus unfamiliar accents.

- **Block** referred to the trial number in question. In total, there were 10 trials for each of the four speakers. Block was a categorical predictor and coded if the trial in question was part of the first 5 trials in a specific speaker's voice ('first half') or the second 5 trials ('second half'). It served to identify adaptation effects, that is to say changes in key press and/or correction accuracy while listening to one speaker. 'First half' was set as the reference level for this predictor.

- The rationale behind including z-transformed **speech rate** has been explained in subsection 3.2.4.5.1.

The random effects included were the same as the ones for the lexical decision task (see subsection 3.2.4.5.1).

3.2.4.5.3 Transcription Task

For the accuracy in the transcription task, only **accent** was added as a predictor to the models to see if participants were better at transcribing TE (reference level) versus NZE. To avoid problems of singular fit for the small transcription dataset, the random effects structure only included random intercepts for **participants** as it was assumed that more variation would occur between participants than between items.

3.2.4.5.4 Presentation of Results

The results for each measure from the pilot study will be presented in three steps. First, a descriptive overview of the data is given. Second, the model that was used to analyse the data and its output are presented, along with an overview of which predictors did (not) reach significance. Finally, the results for each significant predictor are provided in more detail

to serve as a basis for the discussion of the results in Chapter 6. Priority is given here to main experimental predictors (e.g. accent) over predictors of secondary interest (e.g. word frequency). This three-step presentation of results will be applied throughout this thesis.

3.2.5 Results

3.2.5.1 Lexical Decision Task

3.2.5.1.1 Accuracy

After the exclusion of 77 outlier responses (see subsection 3.2.4.1), the dataset consisted of 969 responses to words in the lexical decision task. Accuracy was overall very high for the lexical decision task, with 936 correct responses (96.6%) and 33 incorrect responses (3.4%). As can be seen in Figure 3.4, accuracy rates were slightly higher for TE than NZE.

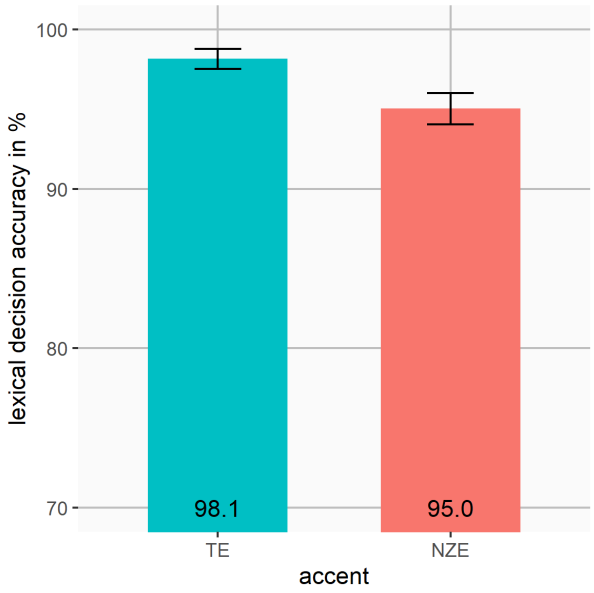


Figure 3.4: Lexical decision accuracy as mediated by accent (bars: mean, error bars: ±one SE, colour: accent)

Following the procedure outlined in subsection 3.2.4.5.1, the below model was used to analyse the accuracy data from the pilot study. The model output is provided in Table 3.6.

$$\text{glmer(accuracy} \sim \text{accent x block + speech\_rate\_z + freq\_log10\_z}$$
$$\text{(1|participant) + (1|item))}$$

Predictor	Estimate	Standard Error
Intercept	6.42	1.18
Accent (NZE)	-1.57	1.24
Block (second half)	-1.82	1.10
Accent (NZE) : Block (second half)	1.17	1.22
Speech Rate (z-scored)	0.10	0.31
Word Frequency (log10, z-scored)	0.79	0.32

Table 3.6: Model output: lexical decision accuracy

Model comparisons showed that there were two significant predictors: block, $\chi^2(1) = 5.71$, $p = .017$; and word frequency, $\chi^2(1) = 5.74$, $p = .017$. The other predictors and the interaction did not reach significance: accent, $\chi^2(1) = 1.54$, $p = .215$; speech rate, $\chi^2(1) = 0.10$, $p = .750$; and the interaction between accent and block, $\chi^2(1) = 1.07$, $p = .300$.

The main effect of block is visualised in Figure 3.5, which plots mean accuracy rates for the first versus second half of trials. As previously observed, accuracy was overall very high, which leaves little room for variation between blocks. While the effect of block was significant, $\beta = -1.24$, $SE = 0.61$,[16] the accuracy difference between word trials in the first 14 trials versus the second 14 trials produced by a specific speaker was only 2.1 percentage points. This effect seems to be renewed for each speaker that the participants heard. Put differently, even for the second speaker in TE or NZE, participants' accuracy was initially slightly higher than during the second 14 trials produced by this speaker.

[16] The model output in Table 3.6 shows simple effects of accent and block. This is how R usually prints the model coefficients. Here, there is a main effect of block and the corresponding coefficient of this main effect is provided. This coefficient was calculated manually based on the model output.

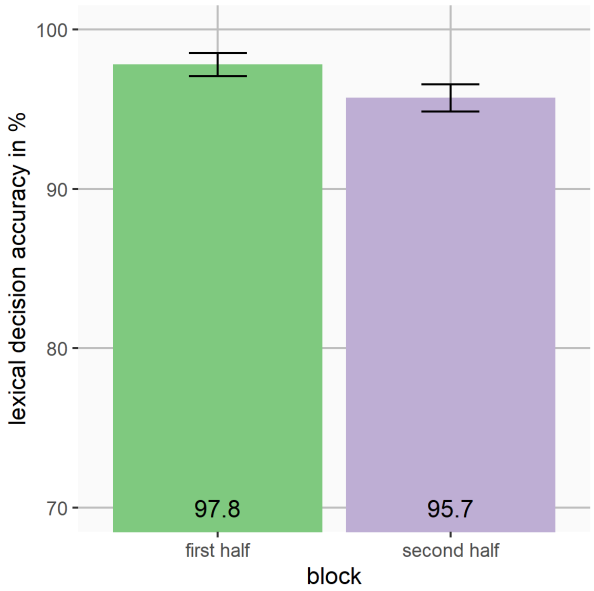


Figure 3.5: Lexical decision accuracy as mediated by block (bars: mean, error bars: ±one SE, colour: block)

In addition to the main effect of block, there was a significant effect of word frequency. As can be seen in Figure 3.6 accuracy increased with frequency, $\beta = 0.80$, $SE = 0.32$. The effect of word frequency seems to be driven by relatively low accuracies for words with frequency values below z = -2 (target: *calf*) and for frequency values between z = -1.0 and z = -1.25 (targets: *chef* and *bees*). Except for these words, accuracy was practically at ceiling in this task.

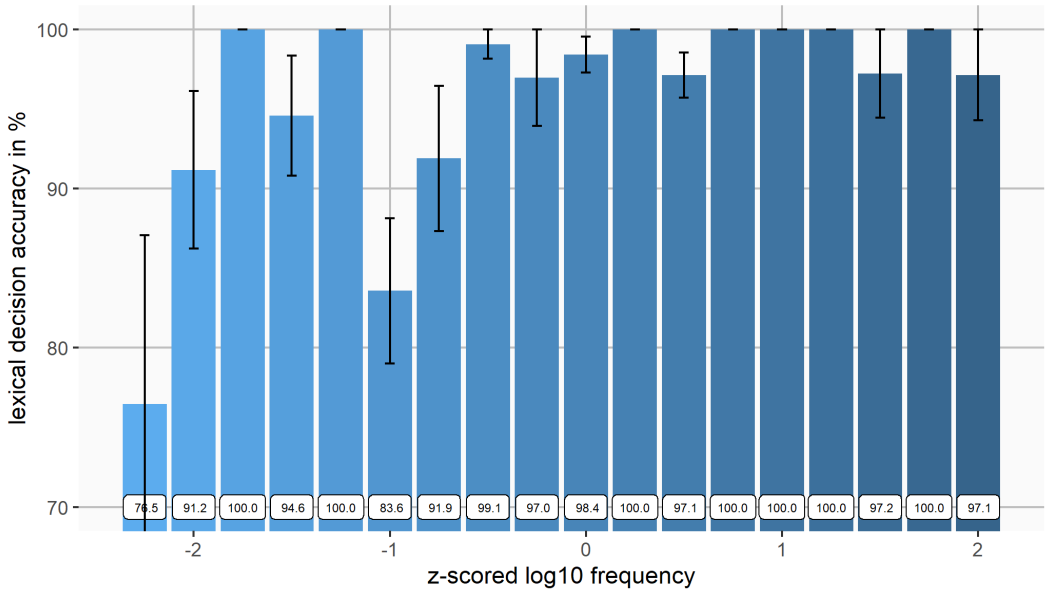
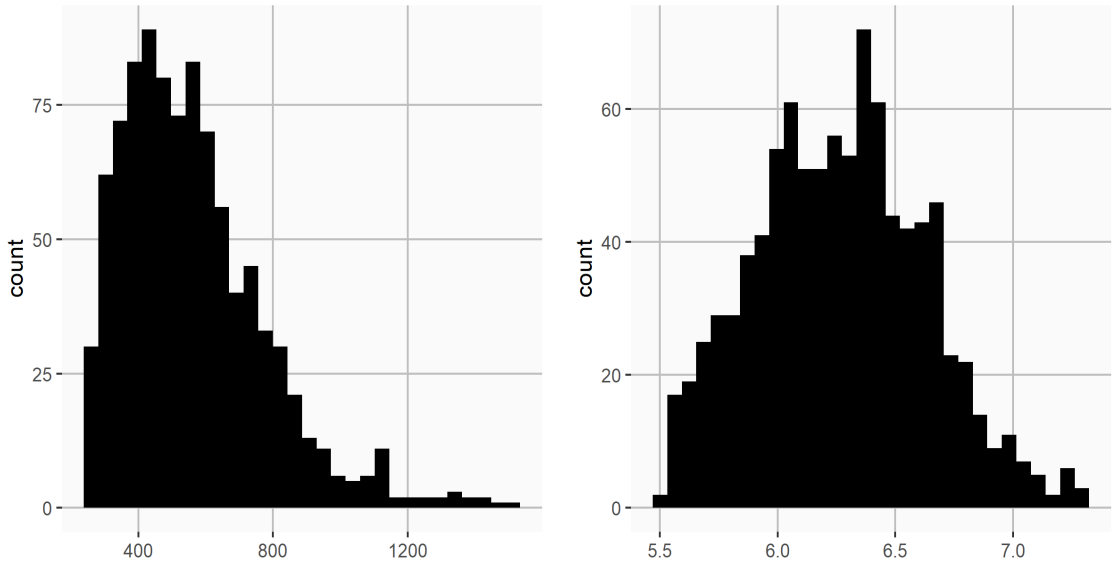


Figure 3.6: Lexical decision accuracy as mediated by word frequency (bars: mean, error bars: ±one SE)

3.2.5.1.2 Lexical Decision Latencies

LDLs were considered for the 936 correct responses. The histogram in Figure 3.7a shows that the latencies exhibited a positive skew, which would result in a similarly skewed distribution of the residuals of the model. To fulfil the normality assumption of linear regression, the latencies were log-transformed, the result of which is displayed in Figure 3.7b. Boxplots of the log-transformed latencies for each accent are provided in Figure 3.8. The medians were similar for TE and NZE although the spread of the data was slightly lower for TE. Importantly, due to the logarithmic transformation of the data, even small differences in the transformed latencies correspond to quite large differences in raw latencies. For example, logLDL = 6 corresponds to 403 ms while logLDL = 6.5 corresponds to 665 ms. Thus, a difference of 0.5 in log-transformed units can equal a difference of more than 250 ms in raw latencies.



(a) raw latencies *(b) log-transformed latencies*

Figure 3.7: Histograms for LDLs

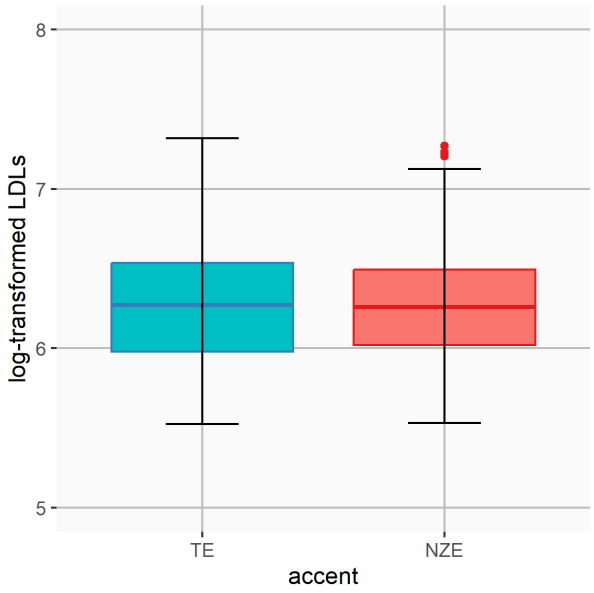


Figure 3.8: Log-transformed LDLs as mediated by accent (colour: accent)

The LDL data were fed into the following model, the output of which is shown in Table 3.7:

$$\text{lmer(LDLs} \sim \text{accent x block + speech\_rate\_z + freq\_log10\_z +}$$
$$\text{(1 + accent|participant) + (1|item))}$$

Predictor	Estimate	Standard Error
Intercept	6.29	0.06
Accent (NZE)	-0.02	0.04
Block (second half)	-0.03	0.03
Accent (NZE) : Block (second half)	-0.03	0.04
Speech Rate (z-scored)	0.02	0.01
Word Frequency (log10, z-scored)	-0.03	0.02

Table 3.7: Model output: LDLs

The only predictor to emerge as significant from the model comparisons was block, $\chi^2(1) = 5.01$, $p = .025$. The other model components were not significant: accent, $\chi^2(1) = 1.09$, $p = .296$; speech rate, $\chi^2(1) = 3.13$, $p = .077$; word frequency, $\chi^2(1) = 3.07$, $p = .080$; and the interaction between accent and block, $\chi^2(1) = 0.58$, $p = .448$.

Figure 3.9 shows the effect of block, β = -0.04, SE = 0.02 (see footnote 16). For ease of interpretation, raw rather than log-transformed LDLs are plotted. On average, participants reacted 23 ms faster to sentences that are among the second 14 sentences they hear in a specific speaker's voice. As for the accuracy data, it is important to bear in mind that this effect is renewed three times during the experiment. Hence, reactions did not become faster over the course of the entire experiment. Instead, they were initially slower and then faster for each speaker.

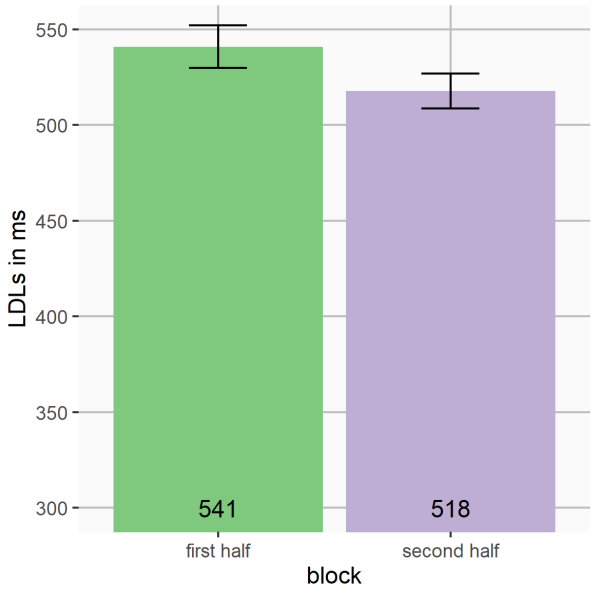


Figure 3.9: LDLs as mediated by block
(bars: mean, error bars: ±one SE, colour: block)

3.2.5.2 Recall Task

3.2.5.2.1 Key Press Accuracy

For the accuracy of the key press, the dataset consisted of 454 responses. Accuracy was overall very high for this measure, with 418 correct responses (92.1%) and 36 incorrect responses (7.9%). As can be seen in Figure 3.10, accuracy rates were generally highest for identical stories (unchanged story type). Accuracy was lowest for stories with semantically related changes presented in NZE. Overall, the participants were more accurate when listening to the familiar accent TE.

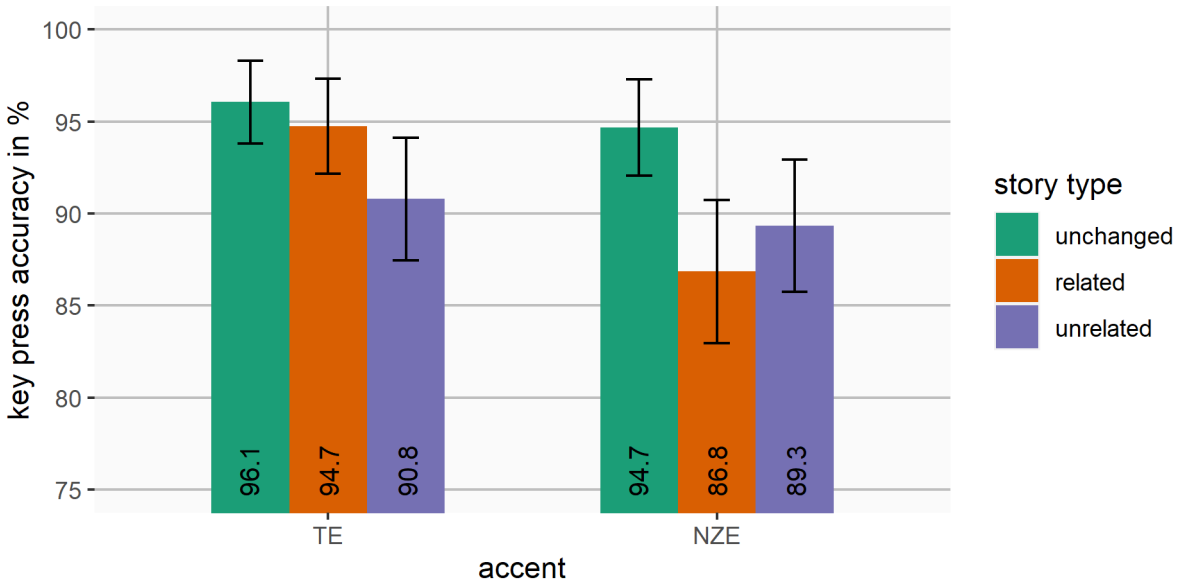


Figure 3.10: Key press accuracy as mediated by accent and story type (bars: mean, error bars: ±one SE, colour: story type)

The below model was used to analyse key press accuracy data. Table 3.8 presents the output of the model.

glmer(accuracy_key_press ~ accent x story_type + speech_rate_z + block + (1|participant) + (1|item))

Predictor	Estimate	Standard Error
Intercept	3.31	0.72
Accent (NZE)	0.13	0.93
Story Type (related)	-0.39	0.80
Story Type (unrelated)	-0.95	0.72
Accent (NZE) : Story Type (related)	-0.65	1.01
Accent (NZE) : Story Type (unrelated)	0.24	0.97
Speech Rate (z-scored)	-0.36	0.31
Block (second half)	-0.22	0.37

Table 3.8: Model output: key press accuracy

None of the predictors emerged as significant from the model comparisons: accent, $\chi^2(1) < .01$, $p = .985$; story type, $\chi^2(2) = 3.47$, $p = .177$; speech rate, $\chi^2(1) = 1.40$,

$p = .236$; block, $\chi^2(1) = 0.37$, $p = .543$; nor the interaction between accent and story type, $\chi^2(2) = 1.16$, $p = .559$.

3.2.5.2.2 Correction Accuracy

274 responses were considered for the accuracy of the correction. As pointed out above, this number represents the trials with a correct indication of a difference between the stories via key press. Accuracy was high overall: 248 corrections of the change were right (90.5%) and 26 corrections were wrong (9.5%). Correction accuracy rates per accent and story type are shown in Figure 3.11. Correction rates were generally lower for stories with semantically unrelated changes and for NZE trials.

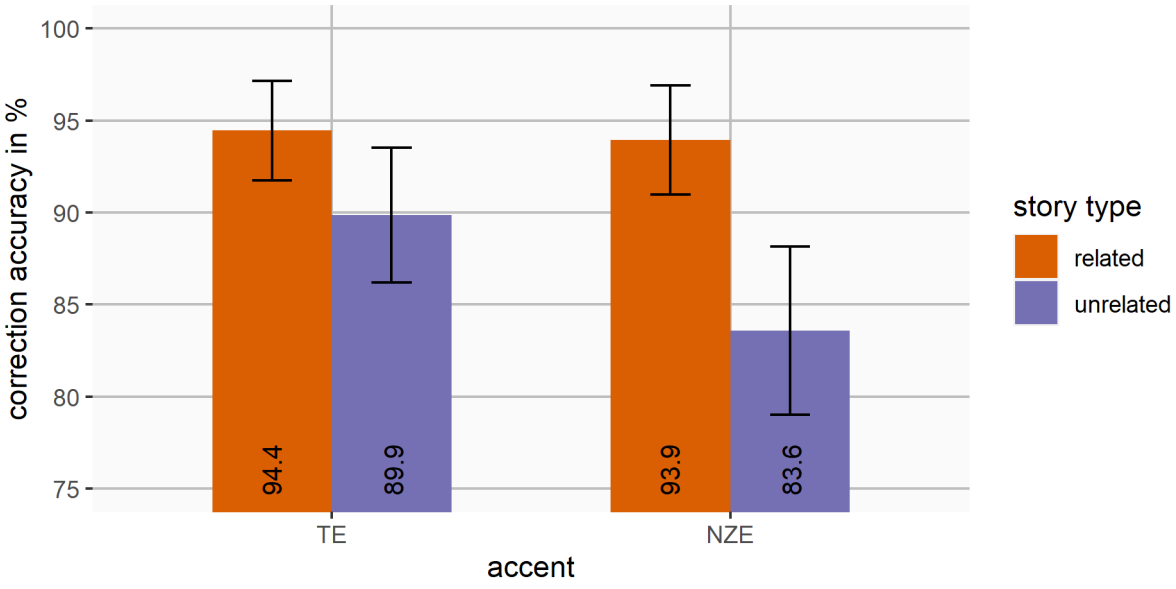


Figure 3.11: Correction accuracy as mediated by accent and story type (bars: mean, error bars: ±one SE, colour: story type)

The predictors for the correction accuracy were identical to the ones for the key press accuracy (see subsection 3.2.5.2.1). However, story type only included two levels since unchanged stories could logically not be considered for the correction accuracy measure. Hence, the reference level for this predictor was set to 'related'. The following model was applied to the data. Its output is provided in Table 3.9.

$$\text{glmer(accuracy\_correction} \sim \text{accent x story\_type} + \text{speech\_rate\_z} + \text{block} + (1|\text{participant}) + (1|\text{item}) \text{)}$$

Predictor	Estimate	Standard Error
Intercept	2.72	0.74
Accent (NZE)	0.48	1.01
Story Type (unrelated)	-0.62	0.68
Accent (NZE) : Story Type (unrelated)	-0.57	0.94
Speech Rate (z-scored)	-0.36	0.39
Block (second half)	0.31	0.46

Table 3.9: Model output: correction accuracy

Model comparisons showed that correction accuracy was mediated by a main effect of story type, $\chi^2(1) = 4.06$, $p = .044$. The other model predictors did not reach significance: accent, $\chi^2(1) = 0.06$, $p = .813$; speech rate, $\chi^2(1) = 0.82$, $p = .366$; block, $\chi^2(1) = 0.45$, $p = .502$; and the interaction between accent and story type, $\chi^2(1) = 0.38$, $p = .540$.

As can be seen in Figure 3.12, which shows the main effect of story type, accuracy was lower for the unrelated story type by 7.4 percentage points, β = -0.91, $SE = 0.79$ (see footnote 16). Thus, participants were better at recalling the change between the two stories if the changed words were semantically related rather than unrelated.

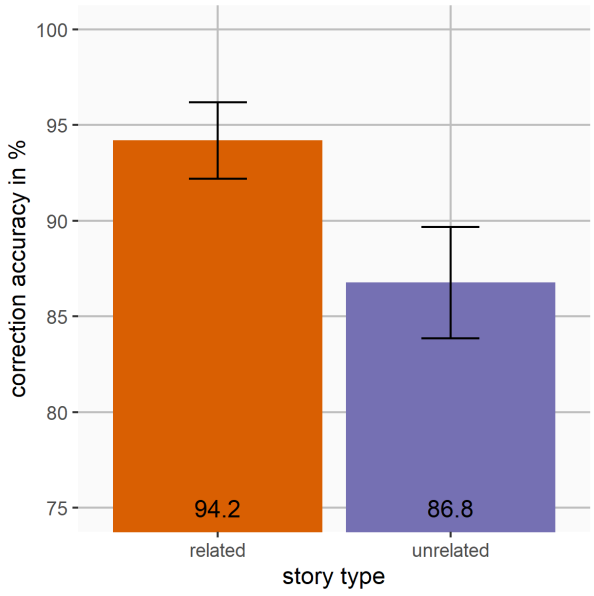


Figure 3.12: Correction accuracy as mediated by story type (bars: mean, error bars: ±one SE, colour: story type)

3.2.5.3 Transcription Task

As shown in Figure 3.13, the participants did extremely well for the transcription task (see subsection 3.2.4.4 for scoring procedure), for which 152 transcribed sentences were analysed. For both accents, the accuracy rate was above 95%. A closer inspection of incorrectly transcribed sentences showed no consistent pattern in that the five sentences that were transcribed incorrectly were all different (see Table 3.10).

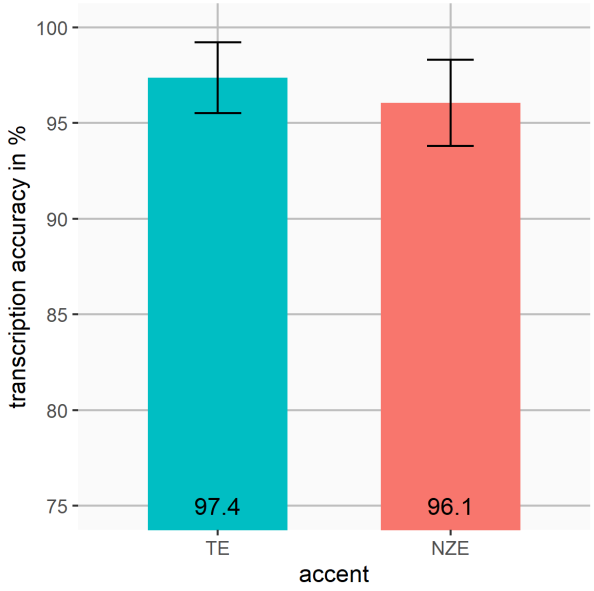


Figure 3.13: Transcription accuracy as mediated by accent (bars: mean, error bars: ±one SE, colour: accent)

Speaker	Transcription Sentence
TE_2	The **man** (main) in the corner is the captain.
TE_2	**In my city** (—) we have very good weather.
NZE_1	**The visitors** (they) thanked the kind driver.
NZE_2	**You must press** (—) the button on the control.
NZE_2	**He keeps** (puts) his stuff in the garage.

Table 3.10: Errors in the transcription task (bold: correct transcriptions, brackets: participant input)

The transcription accuracy data were fed into the following model. Its output is provided in Table 3.11.

$$\text{glmer}(\ \text{accuracy\_transcription} \sim \text{accent} + (1|\text{participant})\ )$$

Predictor	Estimate	Standard Error
Intercept	7.35	3.41
Accent (NZE)	-0.53	1.05

Table 3.11: Model output: transcription accuracy

Model comparisons showed that accent did not mediate transcription accuracy significantly, $\chi^2(1) = 0.26$, $p = .607$. Figure 3.13 demonstrates that this makes sense because accuracy was overall at ceiling and because there was little difference between performance for TE versus NZE speakers.

3.2.5.4 Summary

The lexical decision task showed a speed accuracy tradeoff between the two blocks for each speaker in that lexical decisions became faster but less accurate during the second half. This result suggests relatively quick adaptation to the speakers during the experiment. Given that a block consists of 14 sentences and sentences are, on average across all speakers, shorter than 2 seconds ($\bar{x} = 1858$ms), the participants' LDLs decreased after being exposed to a speaker for less than a minute. While this did concur with a decrease in accuracy, it must be kept in mind that accuracy was overall very high. Importantly, the interaction between accent and block did not reach significance. This is an interesting finding in that it would be sensible to assume that unfamiliar accents lead to a stronger initial disruption and then, as a consequence, leave more room for adaptation as compared to familiar accents. This pattern was not evident in the data. However, this null result does not mean that this interaction does not exist. Potentially, the task was too easy and an interaction between accent and block could have been observed under increased task difficulty. Further, there was an effect of word frequency in that lexical decisions were more accurate for more frequent targets.

For the recall task, none of the predictors significantly influenced key press accuracy. In terms of correction accuracy, there was a significant effect of story type. Participants were better at identifying the differences between semantically related as compared to unrelated stories, which suggest that semantic proximity helps in identifying differences.

Effects of accent and, thus, familiarity might not have emerged because the selection criteria for the participants in the pilot study were looser. In fact, native Tynesiders were explicitly excluded from the analysis here so that their data could be used for the main experiment (see

subsection 3.2.1.1). The aim of the pilot study was to show that the tasks were functioning as intended. As will be shown below, this aim was achieved. The tasks could be used for the main experiment with some small adjustments.

3.2.6 Reflections for Main Experiment

The pilot study showed that the main tasks of the study were functioning as intended, that PCIbex reliably saved the results and that the data could be analysed successfully in R. It further showed that it was possible to motivate participants to take part in this study, including them wearing headphones. Methodologically, only some minor changes were necessary before the main experiment could be launched. Two of these changes addressed shortcomings of the accent matching task. As described in footnote 5 in subsection 3.2.1.2, the participants might have thought that they could not use the labels more than once. In addition, the attention check trial for this task might have been confusing because the instructions to select *Germany* were provided in an American accent. Therefore, the audio file was re-recorded by the researcher with an audible German accent. Finally, a comprehensibility scale was added to the questionnaire at the end of the experiment. Specifically, the participants were asked to rate the comprehensibility of each of the four speakers on a seven-point Likert scale ranging from *very difficult* to *very easy*. An audio example was provided for each speaker.

3.3 Participants

3.3.1 Demographic Information

The participants for the main experiment, which was conducted online, were recruited across Newcastle University, through social media and Prolific. The aim was to find native Tynesiders to ensure high familiarity with TE and lower familiarity with NZE. The participants were compensated for their time with a gift voucher or payment. For the participants not recruited via Prolific, a pre-screening questionnaire was used to find out if they were native Tynesiders (*Have you been born and raised in Tyneside and have you lived here most of your life?*). Potential participants were further asked if they were at least 18 years old and had the necessary equipment for the study. The Newcastle University Form Builder (Newcastle University IT Service, 2023) was used to implement the pre-screening. In general, this proved to be an effective way of accessing participants that fit the sampling criteria. However, more than one third of the people who signed up for the study did not complete it. Participant slots

on Prolific filled up quite quickly but the screening for native Tynesiders could only be done indirectly via post codes on the platform.

In total, the study was completed by 52 participants, including five participants from the pilot study.[17] Four of the 52 participants were excluded for failing the attention check and 5 participants indicated at the end of the study that they were not native Tynesiders. None of the participants reported hearing problems. In sum, all following analyses are applied to the data from 43 participants. As can be seen in Table 3.12, the participants were split relatively evenly across the four groups of the Latin-square design and each group had at least ten participants.[18]

Group	A	B	C	D
Number of Participants	12	11	10	10

Table 3.12: Distribution of participants across groups from Latin-square design

With regards to gender, 20 participants identified as female, 22 as male and one as genderfluid. On average, they were 30.9 years old (sd = 13.9 years).

3.3.2 Familiarity Measures

The participants' familiarity with TE and NZE was measured via their ability to identify the accents during the accent matching task and their self-reports from the questionnaire (see subsection 3.2.3.6). Figure 3.14 shows the results from the accent matching task. Since 9 labels were provided per speaker, the likelihood of choosing the correct response at random was $\frac{1}{9} \approx 0.11$. As can be seen, accent matching accuracy was above that level for both accents but much higher for TE. Incorrectly chosen labels for TE utterances included Liverpool (40.0% of incorrect responses), London (40.0%) and Birmingham (20.0%) . For NZE utterances, incorrect choices included Australia (84.3%), London (11.8%), Birmingham (2.0%) and Newcastle (2.0%). A simple logistic regression model with only accent as a predictor showed significantly lower accent matching accuracy for the NZE utterances, β = -3.16, SE = 0.51, $p < .001$.

[17] Some participants from the pilot study were native Tynesiders. Their data were excluded from the pilot study dataset so that they could be used for the main experiment (see subsection 3.2.1.1).

[18] Following the procedure of the pilot study (see subsection 3.2.3), PCIbex assigned the participants iteratively to one of four groups, which differed with regards to the order of the four speakers. However, in some cases, the participants had to reload the website due to technical issues, which resulted in their assignment to the next group. Furthermore, the exclusion of some participants due to not being native Tynesiders or failing the attention check caused the slightly uneven numbers of participants across groups.

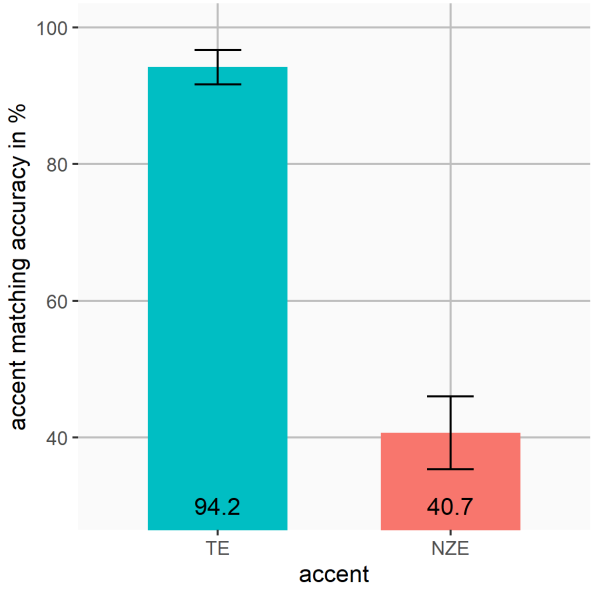


Figure 3.14: Accent matching accuracy as mediated by accent (bars: mean, error bars: ±one SE, colour: accent)

Table 3.13 shows the mean values and standard deviations for the self-reported familiarity and social network measures. For both measures, the means were at the opposite ends of the scales for TE versus NZE. Simple linear models that predicted familiarity and social network based on accent/location showed significantly lower values for NZE (familiarity: β = -3.86, SE = 0.29, $p < .001$; social network: β = -5.33, SE = 0.18, $p < .001$). Thus, the data strongly supported the participants' familiarity with TE and unfamiliarity with NZE.

	Newcastle (English)	**New Zealand (English)**
Familiarity with Accent	5.8 (0.9)	1.9 (1.7)
Social Network	5.6 (1.0)	0.2 (0.6)

Table 3.13: Self-reported familiarity and social network information: mean and standard deviation from seven-point Likert scales (0 to 6)

The difference between TE and NZE was much smaller in terms of comprehensibility ratings. As these ratings were not elicited in the pilot study, they were only available for 41 out of the 43 participants. Mean values per accent are shown in Table 3.14. While comprehensibility ratings were generally higher for the TE speakers than the NZE speakers and this difference did become significant in a simple linear model, β = -1.12, SE = 0.18, $p < .001$, the ratings were a lot closer to each other than both the familiarity and the social network ratings. This discrepancy between familiarity and comprehensibility ratings will be addressed in more detail in section 3.8.

Accent	TE	NZE
Comprehensibility Rating	5.8 (0.8)	4.6 (1.4)

Table 3.14: Self-reported comprehensibility ratings for the four speakers: mean and standard deviation from seven-point Likert scales (0 to 6)

3.4 Stimuli

The stimuli in the main experiment were identical to the ones used in the pilot study. Subsection 3.2.2 provides detailed information on them. Briefly, the stimuli were recorded with four middle-aged female speakers, two with a TE accent and two with a NZE accent. For the lexical decision task, 28 short sentences with two pointer words and a sentence-final monosyllabic target word from Stringer and Iverson (2020) were used. To create the nonword sentences, the target word was replaced with a nonword from Rastle et al. (2002). Thus, there were 56 stimuli for each speaker, which resulted in a total of 224 recorded sentences. The stimuli for the recall task were story triplets. Each story consisted of three sentences and, during one trial of the task, the participants heard two of the three versions of a story (see subsection 3.2.2.1.2). The versions differed in a single word only, which, for experimental triplets, was either semantically related (e.g. *doctor* → *nurse*) or semantically unrelated (e.g. *doctor* → *teacher*) to the word in the original version. 12 experimental and 8 filler triplets were selected for a total of 20 triplets per accent. Another four triplets were identical across all speakers and were used during the practice trials. The transcription task included 8 sentences per speaker. The sentences were also taken from Stringer and Iverson (2020) and their final word was bisyllabic. In total, there were 32 stimuli for this task. For the accent matching task, finally, another sentence from Stringer and Iverson (2020) with a bisyllabic word at its end was recorded per speaker for a total of four sentences. All sentences and stories were thoroughly checked in terms of recording quality before they were used in the experiment.

3.5 Procedure

The procedure for the main experiment mostly followed the one from the pilot study. However, in order to recruit native TE speakers, participants that were not sourced via Prolifc filled in a short pre-screening questionnaire (see subsection 3.3.1). The main tasks are summarised again here. Further details on each task can be found in subsection 3.2.3.

(1) **Headphone check**: To ensure that participants were wearing headphones, they had to identify the quietest in a series of pure tones five out of six times correctly.

(2) **Lexical decision task**: After 10 practice sentences, 112 sentences, either ending in a real word or a nonword, were presented auditorily (28 for each speaker). Trials were blocked by speaker and the order of the speakers was counterbalanced via a Latin-square design (see Table 3.4). For each speaker, the participants completed no more than three word or nonword trials in a row. The participants were instructed to indicate the lexical status of the final (non)word via key press as quickly and accurately as possible.

(3) **Recall task**: After 4 practice trials, 40 pairs of stories were presented auditorily (10 for each speaker). The stories were either identical or differed in a single word, which was either semantically related or semantically unrelated to the original story (see Table 3.5). The presentation of speakers was blocked and the order of the speakers was counterbalanced via a Latin-square design (see Table 3.4). The participants indicated via timed key press if the stories were identical or not. If they indicated that the stories were different, they saw the second story on a new screen and were instructed to identify both the changed word and the original word from the first story. If they indicated that the stories were identical, they moved on to the next trial.

(4) **Accent matching task**: The participants heard each of the four speakers and selected their geographical origin from a list of 9 labels (*Glasgow, Liverpool, Birmingham, London, Newcastle, United States, Australia, New Zealand* and *Germany*). Each label could be used more than once and the participants could listen to the speakers as many times as necessary. A catch trial was included here to monitor that participants were paying attention.

(5) **Transcription task**: The participants transcribed 8 sentences (two for each speaker). Speakers were blocked and the order of the speakers was counterbalanced via a Latin-square design (see Table 3.4).

(6) **Demographic and language background questionnaire**: The questionnaire elicited demographic information. It further included questions on familiarity (*How familiar are you with the accent spoken in Newcastle/New Zealand?*), social network (*How many members of your social network (family and friends) are from Newcastle?*) and comprehensibility. For all measures, seven-point Likert scales were used. The

comprehensibility measure had not been part of the pilot study. For this measure, the participants were asked to listen to each speaker included in the study again and rate their comprehensibility. The audio file included for each speaker was one of the practice stories from the recall task.[19] Since the participants heard the practice stories from one of the four speakers (see Table 3.4), they had encountered one of the comprehensibility audio recordings from the questionnaire earlier in the experiment. They could listen to each audio file as many times as necessary. The rating scales ranged from *very difficult* to *very easy* in response to the prompt: *Listen to each speaker again and state how easy or difficult you found it to understand them.*

3.6 Data Analysis

The data analysis procedure for the main experiment followed the steps taken for the data from the pilot study (see subsection 3.2.4). For the lexical decision task, the accuracy as well as the latencies of the lexical decisions were considered. The measures for the recall task included key press accuracy and correction accuracy. Key press accuracy refers to the participants' initial decision as to whether the two stories were identical or not. Correction accuracy refers to the accurate recall of the change between the two versions of the story. Hence, it was only considered for related and unrelated stories and if the participants' initial key press was correct. Accuracy in the transcription task was measured based on the pointer and key words in the sentences. Where statistical analyses were conducted, the following subsections include the (generalised) linear mixed effects models used, their output and detailed information on significant model predictors, as identified by log-likelihood model comparisons with stepwise reductions of the full model.

3.7 Results

3.7.1 Lexical Decision Task

3.7.1.1 Accuracy

After the exclusion of 192 outlier responses (based on latencies with a lower cut-off value of 250 ms and an upper cut-off value of 2.5 participant-specific standard deviations; see subsection 3.2.4.1), the dataset consisted of 2216 reactions to words in the lexical decision task. Accuracy

[19] Specifically, the participants heard the following story from each speaker: *I took my parents to the old castle in town yesterday. The iron door of the castle was renovated recently. It looked great and the history behind it was interesting.*

was overall very high for the lexical decision task, with 2159 correct responses (97.4%) and 57 incorrect responses (2.6%). As can be seen in Figure 3.15, the accuracy rate was generally higher for the TE speakers (99.4%) as compared to the NZE speakers (95.5%).

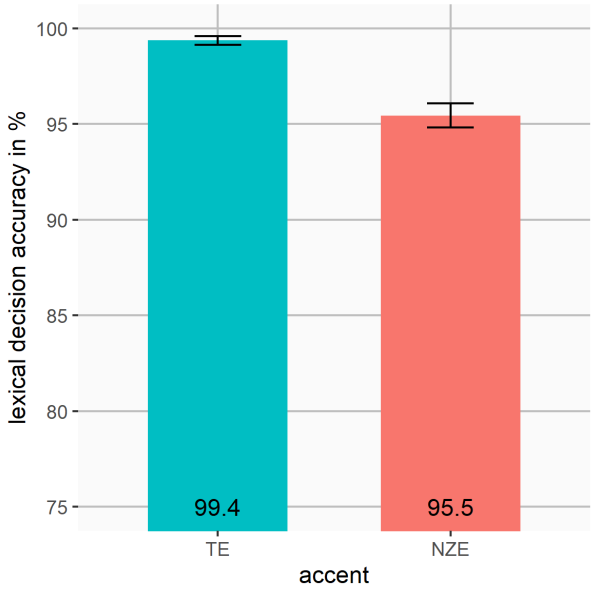


Figure 3.15: Lexical decision accuracy as mediated by accent (bars: mean, error bars: ±one SE, colour: accent)

In order to investigate which predictors mediated accuracy in the lexical decision task, the data were fed into the following model, the output of which is provided in Table 3.15:

$$\text{glmer(accuracy} \sim \text{accent x block + speech\_rate\_z + freq\_log10\_z}$$
$$\text{(1 + accent|participant) + (1|item))}$$

Predictor	Estimate	Standard Error
Intercept	5.67	1.00
Accent (NZE)	-0.99	1.07
Block (second half)	1.26	0.86
Accent (NZE) : Block (second half)	-1.78	0.94
Speech Rate (z-scored)	0.01	0.23
Word Frequency (log10, z-scored)	0.79	0.29

Table 3.15: Model output: lexical decision accuracy

Model comparisons demonstrated that there were significant main effects of accent, $\chi^2(1) = 6.44$, $p = .011$; and word frequency, $\chi^2(1) = 7.10$, $p = .008$. The interaction

between accent and block also reached significance, $\chi^2(1) = 3.95$, $p = .047$. The remaining predictors did not emerge as significant: block, $\chi^2(1) = 0.66$, $p = .418$; and speech rate, $\chi^2(1) < .01$, $p = .971$.

Figure 3.15 above shows the effect of accent on lexical decision accuracy, β = -1.88, SE = 0.99 (see footnote 16). Accuracy was overall very high in this task. The difference between TE and NZE trials was 3.9 percentage points.

Figure 3.16 serves to interpret the significant interaction between accent and block. The main effect of accent is clearly visible here. Follow-up pairwise comparisons showed significant differences in accuracy between TE and NZE for the second half (higher accuracy for TE), $z = 2.45$, $p = .014$; but not for the first half, $z = 0.93$, $p = .354$. Put differently, the accent difference emerges only during the second 14 trials per speaker. If the speaker has a NZE accent, the participants' lexical decision were significantly less accurate than for the TE speakers.

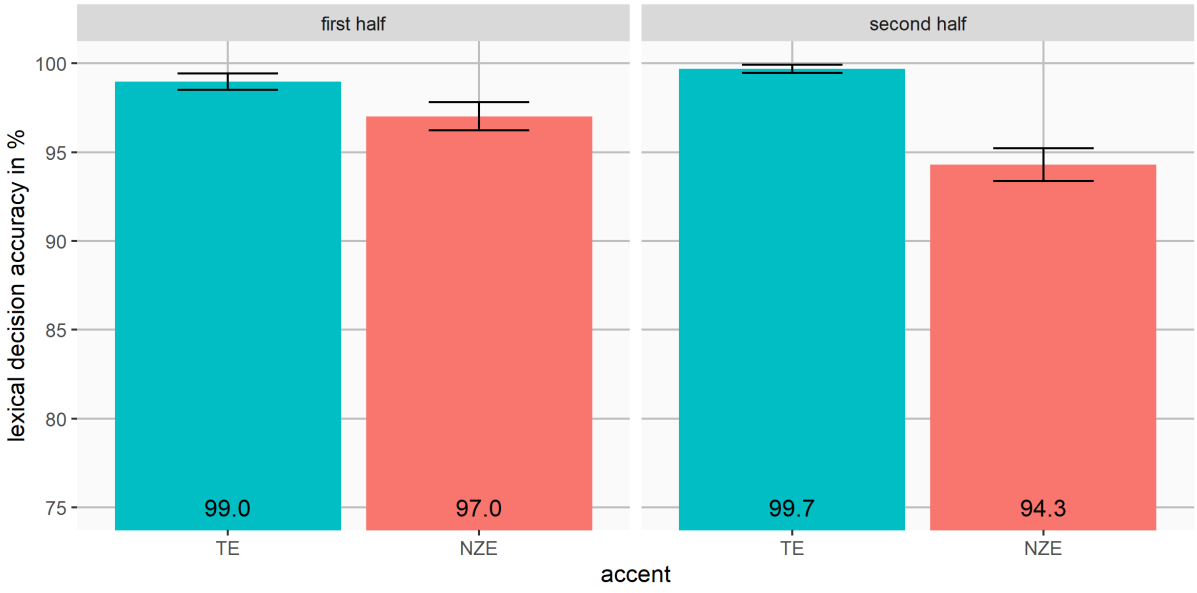


Figure 3.16: Lexical decision accuracy as mediated by accent and block (bars: mean, error bars: ±one SE, colour: accent)

With regards to word frequency, lexical decision accuracy was higher for more frequent targets, $\beta = 0.79$, $SE = 0.29$. This effect is visualised in Figure 3.17, which further shows that participants made more mistakes for targets with word frequencies around z = -1.0 (targets: *bees, chef, rice, soap* and *zoo*), z = -1.5 (targets: *ink* and *jar*) and z = -2.0 (targets: *grapes, vase* and *calf*).

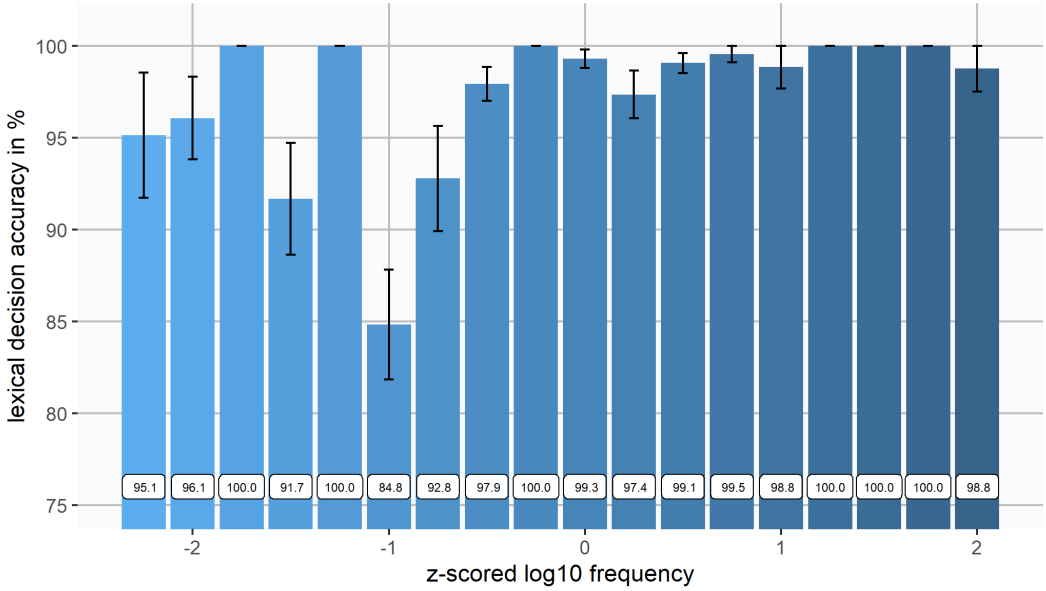
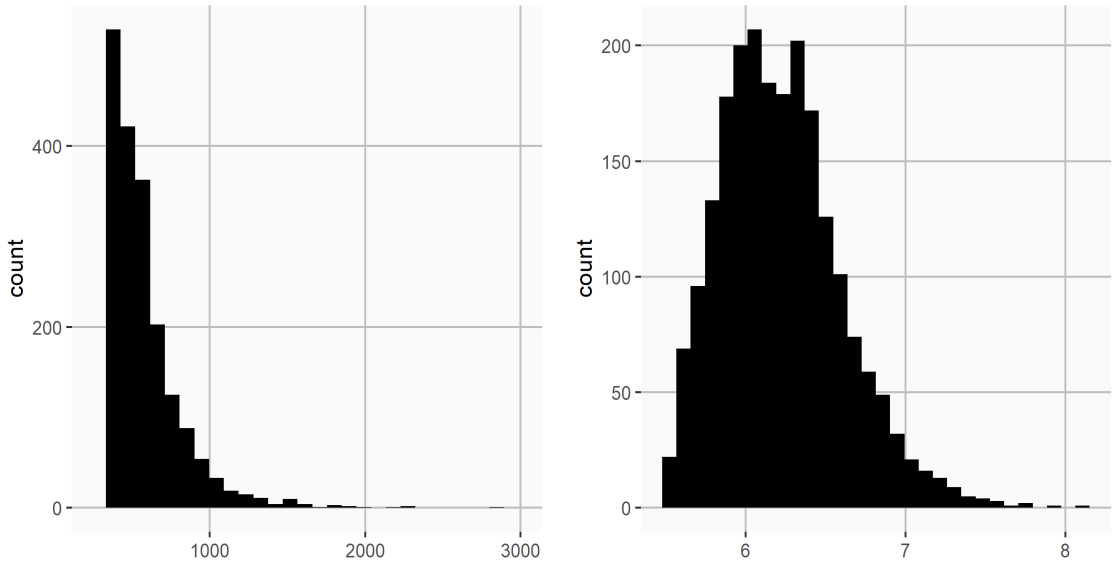


Figure 3.17: Lexical decision accuracy as mediated by word frequency (bars: mean, error bars: ±one SE)

3.7.1.2 Lexical Decision Latencies

LDLs were considered for the 2159 correct responses. The histogram in Figure 3.18a shows that the latencies exhibited a positive skew. To mitigate this skew, the latencies in ms were log-transformed, which, as can be seen in Figure 3.18b, resulted in a distribution closer to the normal distribution. Boxplots of the log-transformed latencies for both accents are provided in Figure 3.19. While the medians were similar for TE and NZE, the interquartile range and general spread of the data were slightly higher for NZE than TE.



(a) raw latencies *(b) log-transformed latencies*

Figure 3.18: Histograms for LDLs

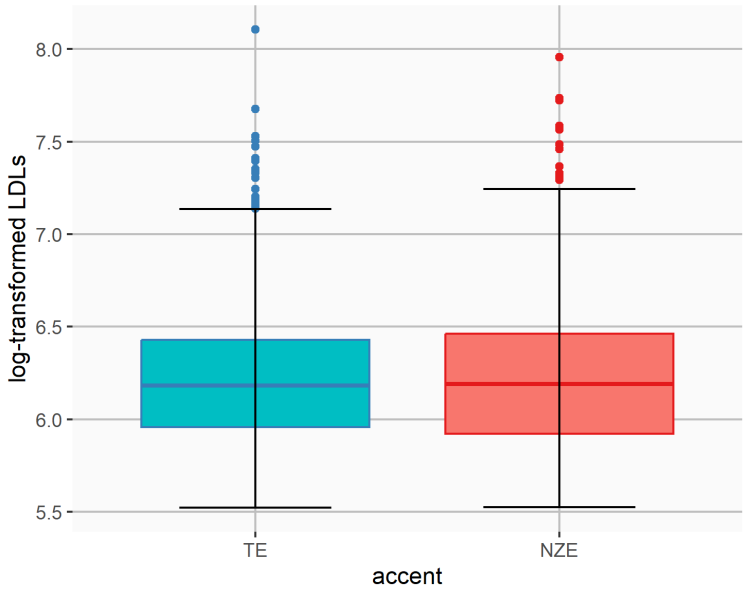


Figure 3.19: Log-transformed LDLs as mediated by accent (colour: accent)

The LDLs were analysed by means of the model below. The output of the model is shown in Table 3.16.

$$\text{lmer}(\ \text{LDL\_log} \sim \text{accent x block} + \text{speech\_rate\_z} + \text{freq\_log10\_z} + (1 + \text{accent}|\text{participant}) + (1|\text{item})\ )$$

Predictor	Estimate	Standard Error
Intercept	6.23	0.04
Accent (NZE)	0.01	0.04
Block (second half)	-0.02	0.02
Accent (NZE) : Block (second half)	-0.04	0.03
Speech Rate (z-scored)	0.01	0.01
Word Frequency (log10, z-scored)	< 0.01	0.02

Table 3.16: Model output: LDLs

Only block emerged as a significant predictor from the model comparisons, $\chi^2(1) = 7.33$, $p = .007$. The remaining predictors and the interaction did not reach significance: accent, $\chi^2(1) = 0.14$, $p = .705$; speech rate, $\chi^2(1) = 0.60$, $p = .439$; word frequency, $\chi^2(1) < 0.01$, $p = .949$; and the interaction between accent and block, $\chi^2(1) = 2.11$, $p = .147$.

With regards to block, as shown in Figure 3.20, the participants' lexical decisions were 18ms faster during the second half of trials, β = -0.03, SE = 0.04 (see footnote 16). Since the effect of block is a main effect, it applies to all four speakers in the study. Thus, latencies did not generally decrease over the course of the experiment but were longer when a new speaker was presented first and were then lower for word trials that are part of the second 14 trials for this speaker.

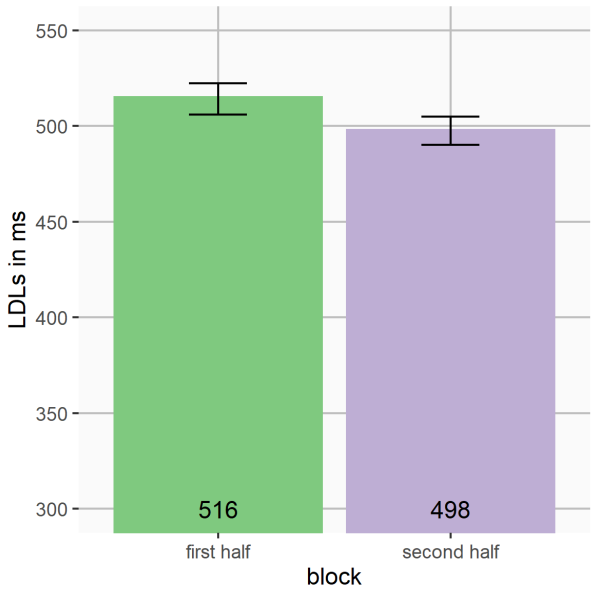


Figure 3.20: LDLs as mediated by block
(bars: mean, error bars: ±one SE, colour: block)

3.7.2 Recall Task

3.7.2.1 Key Press Accuracy

For the accuracy of the initial key press, 1032 responses were recorded. Overall, key press accuracy was high. There were 922 correct responses (89.3%) and 110 incorrect responses (10.7%). An overview of the accuracy rates per accent and story type is shown in Figure 3.21. As can be seen, stories with semantically related changes elicited the lowest accuracy rates for the TE speakers (83.1%). For NZE, accuracy rates were lowest for unrelated stories (87.2%). Altogether, the participants performed better for NZE than TE.

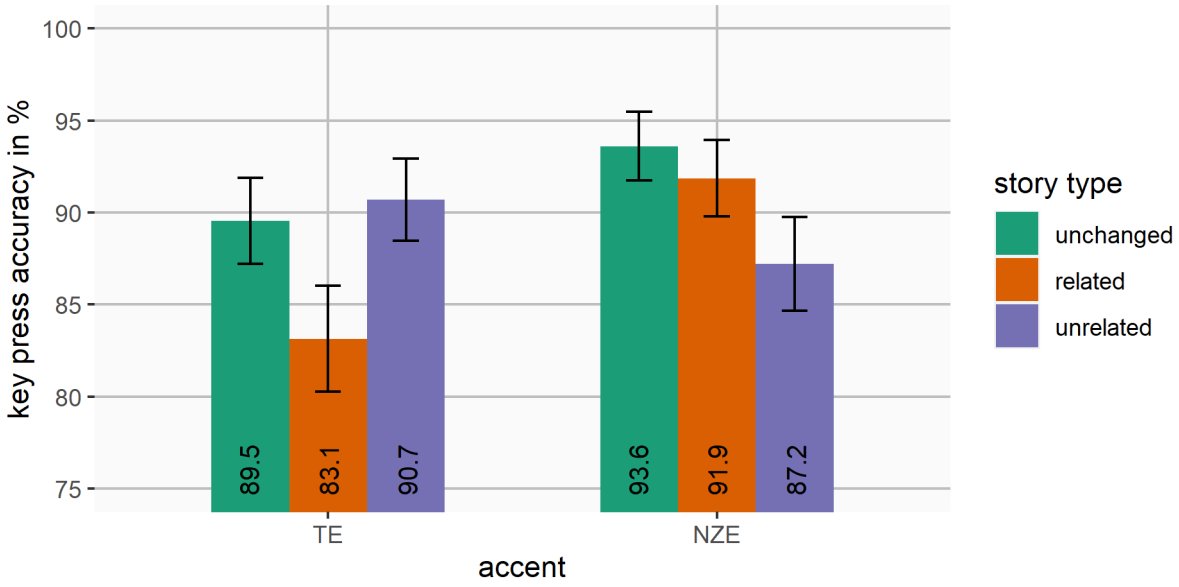


Figure 3.21: Key press accuracy as mediated by accent and story type (bars: mean, error bars: ±one SE, colour: story type)

The accuracy data were fed into the following model. The model output is given in Table 3.17.

$$\text{glmer(accuracy\_key\_press} \sim \text{accent x story\_type} + \text{speech\_rate\_z} + \text{block} + (1|\text{participant}) + (1|\text{item}) \text{)}$$

Predictor	Estimate	Standard Error
Intercept	2.40	0.40
Accent (NZE)	1.15	0.54
Story Type (related)	-0.70	0.35
Story Type (unrelated)	0.20	0.39
Accent (NZE) : Story Type (related)	0.36	0.57
Accent (NZE) : Story Type (unrelated)	-1.11	0.57
Speech Rate (z-scored)	-0.36	0.19
Block (second half)	-0.00	0.22

Table 3.17: Model output: key press accuracy

Model comparisons demonstrated a significant main effect of accent, $\chi^2(1) = 4.44$, $p = .035$; as well as an interaction between accent and story type, $\chi^2(2) = 8.36$, $p = .015$. The remaining model components did not emerge as significant: story type, $\chi^2(2) = 3.53$, $p = .172$; speech rate, $\chi^2(1) = 3.21$, $p = .073$; and block, $\chi^2(1) < 0.01$, $p > .999$.

In terms of accent, the key press accuracy was 3.1 percentage points higher for the unfamiliar NZE accent, $\beta = 0.90$, $SE = 0.43$ (see footnote 16) than the familiar TE accent. This effect is visualised in Figure 3.22.

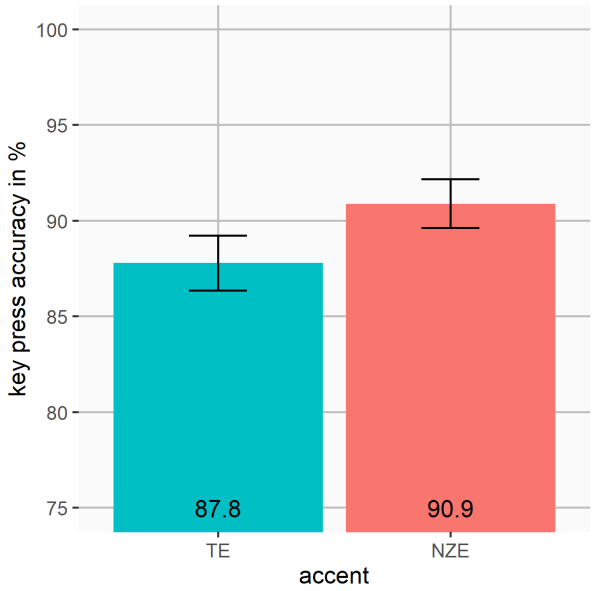


Figure 3.22: Key press accuracy as mediated by accent (bars: mean, error bars: ±one SE, colour: accent)

Figure 3.21 above visualises the interaction between accent and story type. When the accuracy rates are broken down by story type, it becomes clear that the pattern for the main effect of accent only emerged for unchanged and related stories. For unrelated stories, the average accuracy was lower for NZE than TE. This impressionistic observation is, however, not fully supported by pairwise comparisons between the accents. These comparisons show significant differences in key press accuracy for unchanged stories (higher key press accuracy for NZE), z = -2.14, p = .033; and related stories (higher key press accuracy for NZE), z = -2.97, p = .003; but not for unrelated stories, z = -0.08; p = .935. Importantly, the accuracy difference was higher for related stories (8.8 percentage points) than stories without a change (4.1 percentage points). Thus, the effect of accent is more pronounced for the related story type.

3.7.2.2 Correction Accuracy

For this measure, only those trials were considered for which the participants correctly indicated via key press that there was a difference between the two stories. Consequently, the dataset consisted of 607 responses. The overall correction accuracy was slightly lower than the accuracy of the initial key press, with 533 correct responses (87.8%) and 74 incorrect responses (12.2%). Figure 3.23 breaks down accuracy rates per story type for the two accents included. Only related and unrelated stories are included here because the definition of this measure does not include responses to pairs of stories without any changes. For NZE, correction accuracy was almost identical for the related and unrelated stories. For TE, on the other hand, accuracy rates were much higher for stories with semantically related changes (91.6%) as opposed to those with semantically unrelated changes (82.1%). If the accuracy rates are averaged per accent, there was only a small difference between TE (86.6%) and NZE (89.0%).

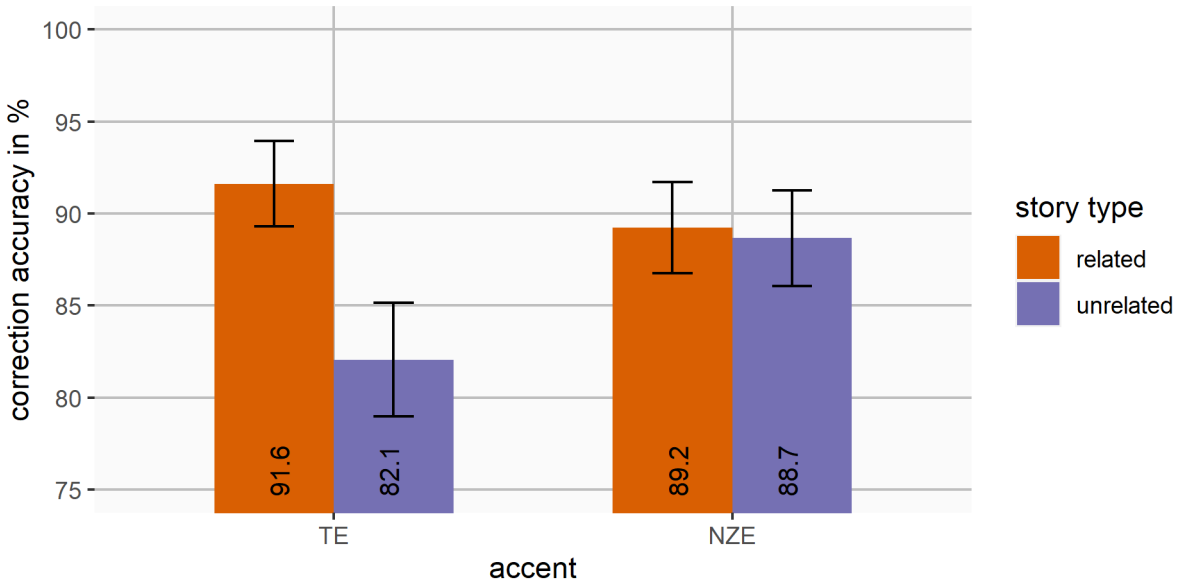


Figure 3.23: Correction accuracy as mediated by accent and story type (bars: mean, error bars: ±one SE, colour: story type)

For correction accuracy, the story type predictor only included two levels: related stories and unrelated stories (see subsection 3.2.5.2.2). All predictors are provided in the model below, with Table 3.18 showing its output.

glmer(accuracy_correction ~ accent x story_type + speech_rate_z + block + (1|participant) + (1|item))

Predictor	Estimate	Standard Error
Intercept	2.62	0.44
Accent (NZE)	0.14	0.54
Story Type (unrelated)	-0.97	0.40
Accent (NZE) : Story Type (unrelated)	0.79	0.56
Speech Rate (z-scored)	-0.28	0.21
Block (second half)	-0.10	0.27

Table 3.18: Model output: correction accuracy

The model comparisons showed that the only predictor to reach significance was story type, $\chi^2(1) = 4.32$, $p = .038$. The other predictors and interaction tested remained unsignificant: accent, $\chi^2(1) = 1.49$, $p = .222$; block, $\chi^2(1) = 0.14$, $p = .713$; speech rate, $\chi^2(1) = 1.66$, $p = .198$; and the interaction between accent and story type, $\chi^2(1) = 2.04$, $p = .153$.

As can be seen in Figure 3.24, which plots average accuracy rates per story type, performance was better for related stories by 5.1 percentage points, β = -0.57, SE = 0.46 (see footnote 16). This seems to be primarily driven by the performance for stories narrated by TE speakers (see Figure 3.23).

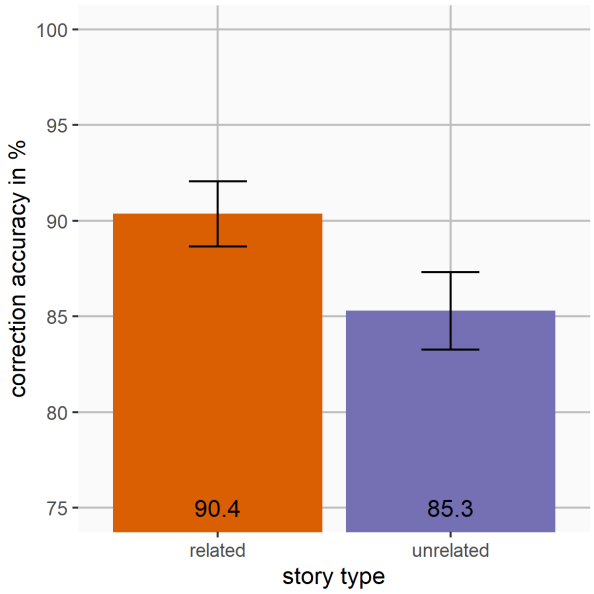


Figure 3.24: Correction accuracy as mediated by story type (bars: mean, error bars: ±one SE, colour: story type)

3.7.3 Transcription Task

The dataset for the transcription task consisted of 344 trials. Overall, transcription accuracy was almost at ceiling, with 330 correct (95.9%) and only 14 incorrect responses (4.1%). Figure 3.25 breaks down transcription accuracy per accent. On average, performance was slightly better for NZE (96.5%) than TE (95.3%).

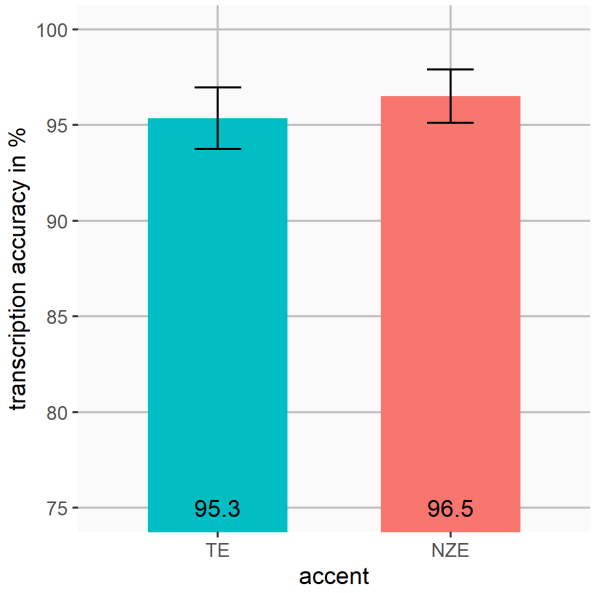


Figure 3.25: Transcription accuracy as mediated by accent (bars: mean, error bars: ±one SE, colour: accent)

The transcription accuracy data were analysed by means of the below model.[20] Its output is shown in Table 3.19.

$$\text{glmer(accuracy\_transcription} \sim \text{accent} + (1|\text{participant}) + (1|\text{item}) \text{)}$$

Predictor	Estimate	Standard Error
Intercept	4.40	0.98
Accent (NZE)	0.29	0.61

Table 3.19: Model output: transcription accuracy

As demonstrated by model comparisons, the effect of accent on transcription performance did not reach significance, $\chi^2(1) = 0.33$, $p = .566$. This is likely due to little room for variation in transcription accuracy (see Figure 3.25).

3.8 Summary of Findings

In the lexical decision task, there was a main effect of accent on accuracy in that lexical decisions were overall more accurate for TE. A speed accuracy tradeoff for lexical decisions was

[20] Since the dataset for the main experiment was larger than the one for the pilot study, random intercepts for items could be added here without issues of singular fit.

only found for the NZE speakers. For these speakers, the participants' lexical decisions were faster but less accurate during the second half. The main effect of block on the LDLs demonstrates that latencies were also shorter during the second half for the TE stimuli. Importantly, accuracy was not negatively affected for this accent during the second block. These results suggest a processing cost for the unfamiliar accent since lexical decisions were less accurate for the NZE stimuli (see Floccia et al., 2009, 2006; Impe et al., 2009). This processing cost is further evident from the finding that faster decisions were detrimental in terms of accuracy only for the unfamiliar accent. Thus, exposure to the unfamiliar accent did not seem to improve accuracy throughout the experiment, which might also have been due to little room for variation with performance close to ceiling. Across all speakers, lexical decisions were more accurate for more frequent targets (see Grainger, 1990; Whaley, 1978).

The accuracy of the key press in the recall task was generally better for the unfamiliar NZE than the familiar TE. Assuming that the lexical processing cost for the unfamiliar accent inhibits recall, this result is surprising. In addition to this main effect of accent, the interaction between accent and story type emerged as significant in that the difference in key press accuracy between the two accents was higher for related stories. Correction accuracy was mediated by story type. Interestingly, the participants were better at identifying the difference between two semantically related versus two semantically unrelated stories. Taken together, these results suggest that accent had an effect on recall performance, with generally better recall of the unfamiliar accent, potentially because of more item-specific processing of the unfamiliar accent (see Clopper et al., 2016; Frances et al., 2018; Grohe & Weber, 2018). Additionally, semantic proximity was generally beneficial for recall performance. Semantically related changes were detected more frequently, potentially because the changed word was primed by the original word, which aided recall (see McNamara, 2005; Perea & Rosa, 2002).

With regards to the research questions from subsection 3.1.3, these are the key findings from Experiment 1:

(1) How does accent familiarity affect lexical processing?

Participants' accuracy in the lexical decision task was generally higher for the familiar accent. For the unfamiliar accent, fast lexical decisions were associated with a decreased accuracy.

(2) Is there evidence for adaptation in terms of lexical processing?

LDLs decreased from the first to the second half of trials in a speaker's voice. For the unfamiliar accent, this was accompanied by lower accuracy rates.

(3) What is the effect of word frequency on lexical processing?

Lexical decision accuracy was higher for more frequent targets.

(4) How do accent familiarity and semantic proximity of the change affect recall?

For the key press accuracy, performance was better for the unfamiliar accent. While the interaction between accent and story type reached significance for key press accuracy, pairwise comparisons only showed significant differences between accents for unchanged and related but nor for unrelated stories.

Correction accuracy was generally higher for related story pairs.

(5) Overall, is there evidence for less-detailed processing of the unfamiliar versus familiar accent?

The results do not suggest less-detailed processing of the unfamiliar accent. In fact, recall is better for this accent, especially for stories with semantically related changes. This suggests a potential priming effect from the original to the changed word in related stories.

The comprehensibility ratings presented in subsection 3.3.2 indicate that the participants did not find NZE much more difficult to understand than TE. Perhaps comprehensibility trumps familiarity and lexical processing as well as recall might be more dependent on how comprehensible an accent is to the listeners. It could also be the case that stronger differences between the accents and/or story types emerge if the task is more difficult. Much of the research from subsection 2.5.2.2 suggested that processing differences between familiar and unfamiliar accents surface more visibly if noise is added to the stimuli. Therefore, Experiment 2, which will be presented in the next chapter, investigates lexical processing and recall performance under adverse listening conditions. The results from the transcription task in Experiment 1 are used as a starting point for the pilot study for Experiment 2.

Chapter 4

Experiment 2: Intelligibility

4.1 Introduction

The results from Experiment 1 suggested a familiarity benefit for lexical processing. In terms of recall, performance was better for the unfamiliar accent and for stories with semantically related rather than unrelated changes. Based on these findings, the experiment in this chapter investigates the combined effects of stimulus intelligibility and accent familiarity on lexical processing and recall as previous research found that effects of familiarity might only emerge under adverse listening conditions (see subsection 4.1.1). Intelligibility is here manipulated through the presentation of stimuli in noise versus in quiet. Familiarity with an accent refers to long-term familiarity as it would be expected from speakers who were born and raised in a region where a specific accent is spoken. The self-reported measures from Experiment 1 showed that the participants, while giving lower ratings for the unfamiliar accent in terms of familiarity and social network, found the four speakers almost equally comprehensible (see subsection 3.3.2). To investigate the effects of adverse listening conditions on the processing of familiar versus unfamiliar accents, Experiment 2 presents the Tyneside English (TE) and New Zealand English (NZE) stimuli in the experimental tasks in noise. Apart from the use of noise-masked stimuli, the procedure of this experiment was largely congruent with the one of Experiment 1, which made it possible to draw comparisons between Experiments 1 and 2 and, thus, research the interaction between familiarity and intelligibility.

In the following subsections, relevant previous research will be reviewed. Next, the research questions for Experiment 2 will be presented, along with an overview of the remaining sections in this chapter.

4.1.1 Lexical Processing under Adverse Listening Conditions

The research considered here has already been discussed in detail in subsection 2.5.2.2. The insights from these studies are summarised here to contextualise the research questions and the methods for this experiment (see subsections 4.1.3 and 4.6).

A key study in terms of the effect of noise on lexical and semantic processing was conducted by Adank et al. (2009), whose participants did a sensibility judgement task in quiet and at three different noise levels. The participants were speakers of Standard Southern British English (SSBE) or Glaswegian English (GE) and heard sentences in both accents. In quiet, there was no difference between the processing of SSBE and GE. Under adverse listening conditions, however, a processing imbalance emerged in that GE participants did equally well for both accents but SSBE participants became slower and less accurate for GE. Adank et al. (2009) replicated this effect in their second experiment and showed an even larger processing delay when L2 English is processed by SSBE participants in noise. Thus, increased task difficulty by means of noise-masked stimuli might be necessary to elicit effects of familiarity.

Although Stringer and Iverson (2019) used a different framework to explain their findings, they made similar observations. In one of their experiments, their participants completed a sentence-recognition task with sentences recorded in, amongst others, SSBE and GE. Similar to Adank and Janse (2010), the stimuli were presented in quiet and three noise conditions. The SSBE participants in this experiment performed equally well for SSBE and GE sentences in quiet and under light noise levels. When the noise masking was stronger, their performance for GE as opposed to SSBE sentences dropped significantly. These findings again suggest an interaction between familiarity and intelligibility in that effects of familiarity on lexical processing might only emerge under adverse listening conditions.

As part of their longitudinal study, Evans and Iverson (2007) had their participants complete a sentence recognition task in noise. The participants, who were all students from the North of England that attended different universities in the UK, heard noise-masked sentences recorded with a Northern English and an SSBE speaker and had to repeat them back to the researchers. In a stepwise procedure, the participants' speech recognition threshold was determined for both accents. Their findings were twofold. First, sentence recognition was generally better for SSBE. Second, participants performed better for this task if their accent had changed towards SSBE during their university education. These results suggest that life-long exposure to SSBE as a variety with high prestige aids lexical processing in noise to the extent that it became more

efficient for the standard than for the participants' northern home accent. Importantly, the participants were not exposed to an unfamiliar accent in this study, which might have elicited worse performance than both the Northern accent and SSBE.

Lastly, Adank et al. (2009) presented an artificial accent in noise, which guaranteed that their participants were unfamiliar with it. In their stimuli, they changed the vowels of Dutch such that, for example, /ɑ/ was realised as [aː], /ɔ/ was realised as [oː], and vice versa. Young and old L1 Dutch participants' speech perception thresholds were measured for this accent and for Standard Dutch. The participants generally performed better for the standard accent. In addition, there was an adaptation effect over the course of the experiment in that performance increased for the artificial accent. In line with the studies above, Adank et al.'s (2009) findings indicate that the lexical processing of familiar (and standard) accents is more robust under adverse listening conditions.

4.1.2 Recall under Adverse Listening Conditions

Past research on how noise affects recall (see subsection 2.5.2.3) was less concerned with the participants' (un)familiarity with the accents used and revolved instead around hearing conditions, working memory capacity and task demands. The findings generally showed that recall was poorer for materials previously presented in noise, both for word lists and more coherent narratives. The reason behind this might be that word recognition in noise requires more cognitive resources, which are then available to a lesser extent for subsequent encoding and recall. In Ward et al. (2016), for example, participants' free recall of the main semantic content of short stories was worse when they heard vocoded versions of the stories rather than stories in quiet. The comparatively small size of this effect might have been due to the high intelligibility of vocoded speech compared to, for example, noise-masked speech. Additionally, higher verbal working memory scores were associated with better recall. Similar findings for the effect of more challenging listening conditions on narrative recall were made by Piquado et al. (2012), Schneider et al. (2000), Tye-Murray et al. (2008) and Wasiuk et al. (2021).

With regards to word lists, Kjellberg et al. (2008) found that participants' recall accuracy of 50-item lists was significantly worse when noise was added to the recordings. Participants' performance under adverse listening conditions was more robust when they scored higher on the working memory task. Similarly, Ljung and Kjellberg's (2008) results showed that longer reverberation time in the speech signal, which makes word identification more challenging, was associated with worse recall of word lists. Specifically, the participants recalled fewer items

from the lists if they heard them with longer reverberation time. However, in this experiment, no mediating effect of working memory was found, potentially due to how the latter was measured.

Marsh et al. (2015) measured to what extent their participants engaged in semantic processing for the recall of word lists in quiet versus in noise. They assumed that word identification under adverse listening conditions would require more cognitive resources and, thus, inhibit semantic processing. This pattern was found for all their measures in that, for example, not only were fewer words recalled in noise but the type of recalled words and the order of recall also showed less semantic processing. Working memory was not measured by Marsh et al. (2015).

Taken together, these results suggest that noise negatively affects recall but also induces more item-specific or verbatim rather than semantically based gist processing. Item-specific processing might aid recall in the current design because only one word is changed between stories. However, it might also be the case that noise increases the task difficulty to such an extent that a potential benefit due to item-specific processing does not surface.

4.1.3 Research Questions

Based on the research reviewed in the previous subsections and the findings from Experiment 1, this chapter is concerned with the following research questions:

(1) How do adverse listening conditions and accent familiarity affect lexical processing?

(2) Is there evidence for adaptation in terms of lexical processing?

(3) What is the effect of word frequency on lexical processing?

(4) How do adverse listening conditions, accent familiarity and semantic proximity of the change affect recall?

(5) Overall, is there evidence for less-detailed processing in quiet versus in noise and/or for the unfamiliar versus familiar accent?

The following two sections provide information on the pilot studies, which were run to test if the noise masking decreased intelligibility and, thus, task difficulty, but not to the extent of eliciting a floor effect. Next, the participants, stimuli, procedure and data analysis steps

for the main experiment are explained in detail. The presentation of results is followed by a summary of the key results from this experiment.

4.2 Pilot Study 1

The key difference between the stimuli used in Experiment 1 versus Experiment 2 was that the latter were masked with noise. Specifically, the stimuli were overlaid with multi-speaker babble noise (see subsection 4.2.2). The pilot studies were conducted to ensure that the noise masking decreased intelligibility to the desired extent. Both pilot studies were conducted online via LabVanced (Finger, Goeke, Diekamp, Standvoß, & König, 2017). For Pilot Study 1, participants from the North East were recruited and asked to complete a transcription task, which had been trialled in Experiment 1. It included the transcription of short sentences presented in noise and in quiet.

4.2.1 Participants

Pilot Study 1 was completed by 18 North East participants. All participants were recruited via Prolific and the pre-screening criteria on Prolific were set to only include L1 speakers of English who were born and currently resided in North East England. In addition, participants could not take part in the study if they had completed Experiment 1 or its pilot study on Prolific. The sampling criteria here were less strict than those for the main experiment because its aim was to verify the decrease in intelligibility through noise masking rather than produce insights into the interaction between familiarity and intelligibility. The participants were paid for their participation.

All participants passed the attention check, which consisted of a catch trial at the end of the experiment (see subsection 4.2.3). At the time of the study, 15 of them identified as female, with the remaining 3 identifying as male. The mean age of the participants was 32.8 years (sd = 13.0 years). Similar to Experiment 1, familiarity[1] and social network measures[2] were collected from the participants on seven-point Likert scales. The mean values and standard deviations of both measures are provided in Table 4.1. Simple linear regression models, which only included accent/location as a predictor, were applied to both measures. Values were significantly lower for New Zealand (English) in terms of both familiarity, β = -4.22,

[1] Familiarity measures were elicited with the following question: *How familiar are you with the accent spoken in Newcastle/New Zealand?*

[2] Social network measures were elicited with the following question: *How many members of your social network (family and friends) are from Newcastle/New Zealand?*

SE = 0.06, $p < .001$; and social network, β = -5.22, SE = 0.37, $p < .001$. These self-reports demonstrate that, although the sampling criteria were looser, there was still an evidently higher familiarity with TE than NZE. In fact, only 3 participants chose values on the upper end of the familiarity scale (selection of 4 and above) for NZE.

	Newcastle (English)	**New Zealand (English)**
Familiarity with Accent	5.7 (1.0)	1.4 (1.5)
Social Network	5.4 (1.5)	0.2 (0.4)

Table 4.1: Self-reported familiarity and social network information: mean and standard deviation from seven-point Likert scales (0 to 6; North East participants)

4.2.2 Stimuli

Recordings for the lexical decision task from Experiment 1 were used to create the stimuli for the transcription task in the pilot study. As described in more detail in subsection 3.2.2, the sentences were selected from Stringer and Iverson's (2020) set of non-native speech recognition (NNSR) sentences. Each item contained two pointer words and a sentence-final monosyllabic target, which together construed the semantic content of the sentence (e.g. *You should put*$_{pointer\ 1}$ *your tickets*$_{pointer\ 2}$ *in the bin*$_{target}$). 40 sentences were chosen, for which high-quality recordings were available of all speakers (TE_1, TE_2, NZE_1 and NZE_2). The recordings were selected based on the evaluations that had been conducted for Experiment 1 (see subsection 3.2.2.3).

During the pilot study, the participants transcribed sentences in quiet and in noise. The 160 chosen audio files (40 per speaker) had to be processed for both conditions. The amplitude of the stimuli to be presented in quiet was normalised to 65 dB in Praat so as to resemble the loudness of a normal conversation (Centers for Disease Control and Prevention, 2022). For the noise condition, an adapted version of a Praat script developed by McCloy (2013) was used to mask the stimuli with multi-speaker babble noise. The script read the wav files into Praat, normalised their amplitude to 65 dB at the sentence level, masked the signal with noise at a signal-to-noise ratio (SNR) of 0 dB and, finally, ensured that the overall intensity of the output was identical to that of the input. The script further added fade-in and fade-out portions of 500 ms each to the stimuli. These portions contained the noise signal exclusively, with increasing amplitude at the beginning of the stimulus and a reversed amplitude contour towards the end of the stimulus. They were included so that participants had a moment to get used to the noise before they would hear the noise-masked sentence.

Multi-speaker babble noise was chosen as the noise signal because it ensured higher ecological validity than, for example, white or pink noise. The "Party Crowd" recording by Simion (2023)[3] was ideal for the purposes of the current experiment because it effectively induced the cocktail party problem (Cherry, 1953; see also Wang & Xu, 2021) without resulting in lexical competition as the recording does not contain discernible words. Apart from the fade-in and fade-out portions, the amplitude of the noise signal was relatively constant.

This procedure resulted in a total of 320 stimuli (160 per condition). Interestingly, although the same noise masking procedure was applied to all four speakers (TE_1, TE_2, NZE_1 and NZE_2), the comprehensibility of the NZE stimuli seemed to decrease disproportionately more, as determined impressionistically by phonetically trained listeners. NZE_1, in particular, was quite difficult to understand over the added babble noise. This could be due to voice quality, which varies between speakers and, thus, is very difficult to control. Voice quality has been shown to interact with adverse listening conditions (e.g. Lyberg Åhlander, Haake, Brännström, Schötz, & Sahlén, 2015; see subsection 6.7.2).

Individual differences in comprehensibility over noise might confound the results of the current experiment, especially because the NZE stimuli were more affected than the TE stimuli. If the results were to show an effect of accent on lexical access in noise, it would be difficult to decide if this was due to an interaction between (un)familiarity with an accent and noise or merely the result of using specific speakers whose voices happen to be disproportionately harder to hear when presented in noise. This question is also discussed below and was addressed by eventually including two groups of participants: one from Tyneside and one from New Zealand (see subsection 4.3.6).

4.2.3 Procedure

After the participants were informed about the purpose of the pilot study and informed consent had been secured, they completed the following three tasks:

(1) **Headphone check**: The participants heard six sequences of three tones and had to identify the quietest one in each case. This task was possible if the participants wore headphones but nearly impossible if they used the loudspeakers of their computer (Woods et al., 2017). To pass the headphone check, five out of six responses had to be correct.

[3] The noise file was converted to mono before being mixed with the sentence recordings.

(2) **Transcription task**: The transcription task included eight practice and 32 experimental trials. The participants were informed that they would hear different speakers in different noise conditions. For each trial, the participants first saw fixation crosses for 500 ms and then heard the stimulus, which was presented either in quiet or with the added multi-speaker babble noise. After the presentation of the stimulus, the participants transcribed the sentence they heard in a textbox. If unsure, they were instructed to enter their best guess. The two conditions were blocked, with 8 participants hearing sentences in quiet first while the other 10 heard sentences in noise first.[4] The practice round included one sentence per speaker per condition for a total of eight trials. The main part of the task included four sentences per speaker per condition for a total of 32 trials. Sentences were never repeated within one session and, within one block, the participants were exposed to a maximum of three trials with the same speaker in a row. Which sentence was presented by which speaker was determined pseudorandomly in R. This information was then fed into LabVanced so that the software could order the practice and experimental trials accordingly.

(3) **Demographic and language background questionnaire**: The questionnaire consisted of familiarity and social network measures (Likert scales from 0 to 6; see subsection 3.2.3.6 for more information and subsection 4.2.1 for results) as well as questions on basic demographic information. It further included a catch trial to ensure that participants were paying attention throughout the experiment (*Please type the word 'custard' into the box on the right*).

The pilot study was run on LabVanced rather than the PennController for IBEX (PCIbex; used in Experiment 1), which made a smooth implementation of the study on Prolific possible. LabVanced was further developed to guarantee high precision in the temporal domain. This was important for the reaction time measures included in the tasks of the main experiment. While LabVanced includes a randomisation function, the randomisation of all experiments in the current thesis was conducted in R to make the process more easily customisable. The resulting trial orders were then uploaded to LabVanced.

[4] The allocation of participants to the 'quiet first' and 'noise first' groups was done iteratively in LabVanced through a system-internal counter. The reason why numbers in each group were not equal is that some participants started the study but did not complete it. For these participants, no data is available for the transcription task but the internal counter increased nonetheless.

4.2.4 Data Analysis

The aim of the transcription task was to measure if the participants understood the main semantic content of each sentence. The pointer and target words from Stringer and Iverson's (2020) NNSR sentences were chosen to assess transcription accuracy because they include the main semantic content of the sentence. This allowed for an efficient analysis of the data since the presence or absence of the pointer and target words in the participants' transcriptions could be assessed (semi-)automatically in R. A similar procedure was used by Evans and Iverson (2007), Adank and Janse (2010) and Stringer and Iverson (2019) for their studies on speech perception in noise. An initial problem with the analysis in R was that the software simply matches character strings. Thus, typos (e.g. *neice* instead of *niece*) would be flagged as incorrect. To avoid this, the participants' input for the pointer and target words was corrected manually before the analysis in R.

Two accuracy measures were devised for the participants' performance in the transcription task. Each sentence contained a total of three semantic words of interest (two pointer words and one target word). The **total accuracy** measure was binary. The participants' input for a given trial was only classified as correct if it included all three words. This measure was used to analyse the transcription data from Experiment 1 (see subsection 3.2.4.4).[5] The **proportional accuracy** measure was ordinal and was calculated by adding up the number of correct words included in the participants' input (0, 1, 2 or 3) and then dividing it by three. The total accuracy was more conservative whereas the proportional accuracy allowed for a more nuanced picture of the participants' performance. Proportional accuracy was eventually chosen as the operationalisation measure of transcription accuracy because it provided more detail and because this measure was also used by Stringer and Iverson (2019), which allowed for an easier comparison. The comparison between this pilot study and Stringer and Iverson (2019) was further facilitated by the fact that both studies used weakly constrained or 'neutral' sentences from Stringer and Iverson (2020).

For an SNR of 0 dB, Stringer and Iverson (2019, 2218-2219) found that their SSBE participants could, on average, repeat back approximately 90% of pointer words and targets for SSBE stimuli and approximately 60% for the unfamiliar GE. These findings were used as

[5] While transcription data in quiet were available from Experiment 1, they were not included in the analysis here for two reasons. First, only very few data points were collected per participant. Second, and more importantly, the transcription sentences were only presented in quiet. The advantage of the new pilot studies for intelligibility is that participants transcribed both in quiet and in noise, which resulted in a dataset that was less susceptible to undesired between-participant variation.

a reference point for the current research. Thus, the aim for the current pilot study was to corroborate that the proportional accuracy for the stimuli presented in noise was around 30 percentage points lower than that of the stimuli in quiet. The 30 percentage points measure was set as an overall target for all stimuli in quiet versus noise combined. The main tasks of Experiment 2 (see section 4.6) were a lexical decision and a recall task. Performance in these tasks would crucially depend on the participants' understanding of target words. Thus, a decrease in proportional accuracy of 30 percentage points was deemed ideal to avoid ceiling or floor effects during Experiment 2.

The data of the pilot study were visualised and analysed statistically in R. Since proportional accuracy was expressed through percentage values, it might be considered a continuous variable, which would have allowed for the use of linear mixed effects models. However, the accuracy measure was ordinal rather than continuous as, effectively, it could only be one of four values for each trial: 0.00 (or 0%) if none of the key words were transcribed correctly, 0.33 (or 33%) if one key word was transcribed correctly, 0.67 (or 67%) if two key words were transcribed correctly or 1.00 (or 100%) if all three key words were transcribed correctly. For that reason, ordinal logistic regressions were used instead (Ackerman, 2019; Barlaz, 2023). The use of these models was implemented in R with the 'ordinal' package (Christensen, 2022). Ordinal logistic regression assumes that the individual categories of the ordinal data are equidistant, which is usually difficult to confirm, for example when Likert scales are used (Barlaz, 2023). Equidistance could be assumed for these data since one step on the proportional accuracy scale corresponded to one correctly transcribed key word. Differences between key words across sentences and differences between participants were mediated by including random factors in the model.

The above considerations resulted in an ordinal logistic mixed effects model with two fixed effects and one interaction term:

- **Accent** was categorical and had two levels: 'TE/familiar accent' and 'NZE/unfamiliar accent'. The reference level was set to 'TE'. This predictor was included to test for effects of familiarity on transcription performance.
- **Condition** coded how the stimuli were presented. It was categorical and had two levels: the reference level 'quiet' and 'noise'. It was included to test if noise masking had the desired effect.

- The interaction between **accent** and **condition** was included to further investigate if the noise masking resulted in a stronger performance decrease for one accent as compared to the other.

Two random effects were added to the models:

- Each participant completed 32 experimental trials during the transcription task. To account for these repeated measures, the models included random intercepts for **participants**. Furthermore, random slopes for accent by participant were included if they did not result in singularity issues for the model.
- Random intercepts for sentences or **items** were included because the same sentence occurred multiple times in the dataset. If possible, random slopes for accent by item were also added.

Significant predictors were identified by log-likelihood model comparisons in R. The model comparisons were conducted manually via a stepwise reduction of the model. The first predictor to be taken out of the model was the interaction, followed by condition and accent.

4.2.5 Results

As soon as five complete submissions for the study were available, the data were checked for potential floor or ceiling effects in the noise condition. Since none of these were detected, data were collected until the final 18 submissions were achieved. The full dataset of the pilot study included 576 transcribed sentences. The overall effect of the added noise, without considering individual speakers, is visualised in Figure 4.1, which shows an accuracy decrease of 34 percentage points for the noise-masked stimuli. Transcription accuracy was close to ceiling in the quiet condition. The difference between the two conditions was close to the desired 30 percentage points.

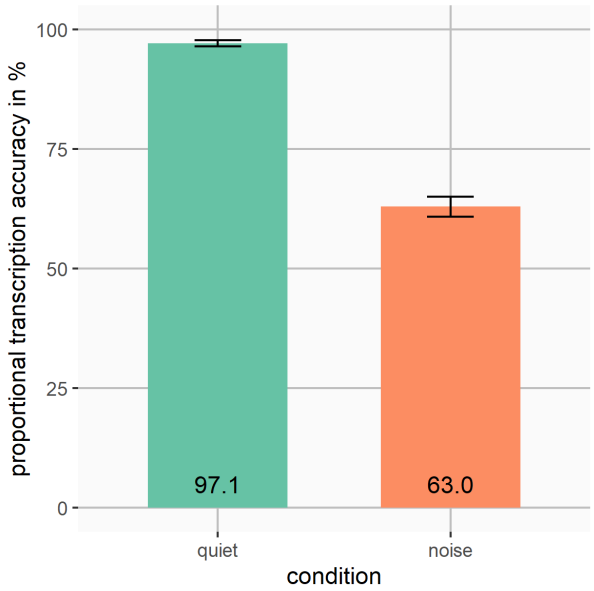


Figure 4.1: Proportional transcription accuracy as mediated by condition (bars: mean, error bars: ±one SE, colour: condition; North East participants)

The effect of noise was not equal across the four speakers. As can be seen in Figure 4.2, transcriptions of TE sentences were least affected by the added noise, followed by NZE_2's and NZE_1's sentences. The data were fed into the following ordinal logistic regression model:

clmm(accuracy_proportional ~ accent x condition + (1 + accent|participant) + (1|item))

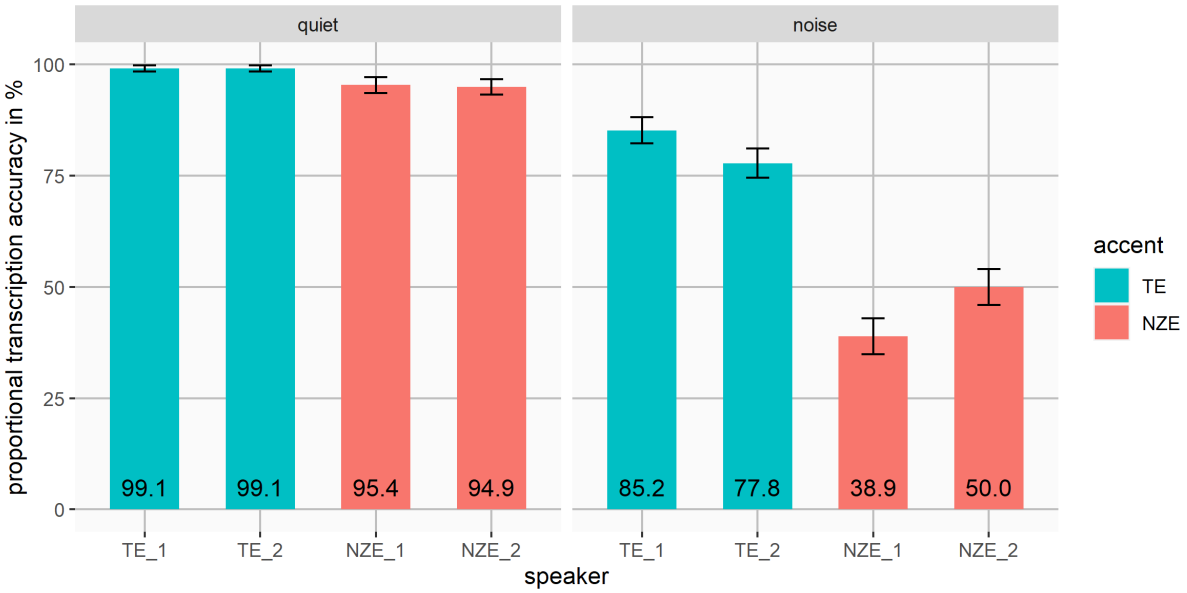


Figure 4.2: Proportional transcription accuracy as mediated by condition and speaker (bars: mean, error bars: ±one SE, colour: accent; North East participants)

Model comparisons demonstrated that there were significant main effects of both accent, $LR(1) = 24.34$, $p < .001$; and condition, $LR(1) = 289.46$, $p < .001$. Although the proportional accuracy values from Figure 4.2 point towards an interaction between accent and condition, this term did not reach significance, $LR(1) = 2.29$, $p = .130$. Transcription performance was worse in noise, β = -3.85, SE = 0.60; and for NZE trials, β = -1.68, SE = 0.71.

Figure 4.3 shows the predicted rates of occurrence for each level of proportional accuracy. The latter are plotted along the x-axis while the former are shown on the y-axis. For example, the model predicted a proportional accuracy of 1.00, that is to say all key words were transcribed correctly, to occur with a likelihood of $0.90 \leq x \leq 0.99$ for NZE speakers presented in quiet. For the noise-masked NZE trials, on the other hand, the model predicted a proportional accuracy of 1.00 to occur with a likelihood of $0.04 \leq x \leq 0.18$. These values correspond to the decreased accuracy associated with NZE_1 and NZE_2 for the noise-masked trials (see Figure 4.2).

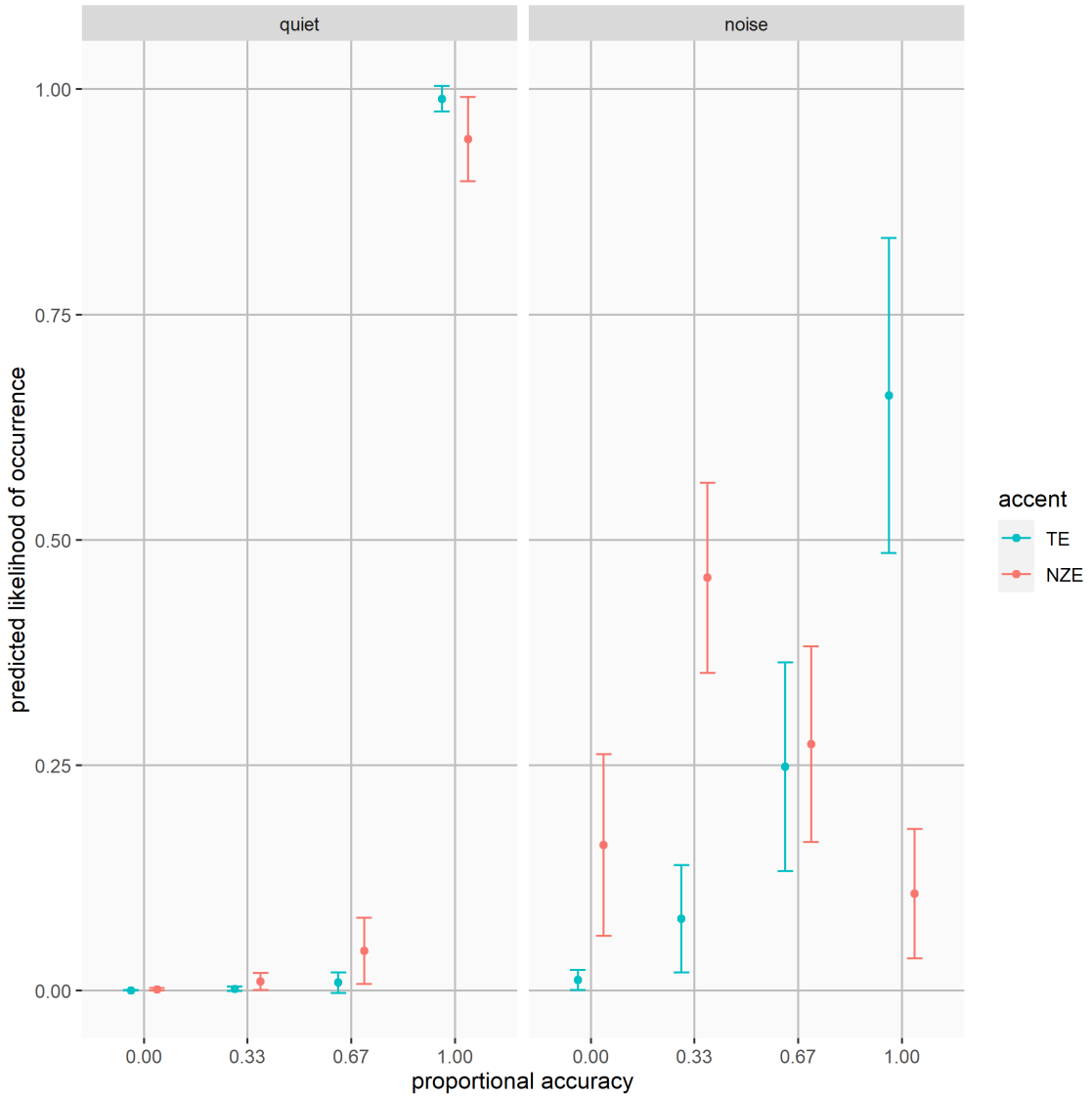


Figure 4.3: Predicted likelihoods of occurrence for each level of proportional accuracy as mediated by accent and condition
(points: means, error bars: confidence intervals, colour: accent; North East participants)

4.2.6 Summary

At first sight, the results of Pilot Study 1 displayed the desired effects of the noise masking. Overall, the transcription accuracy under adverse listening conditions was 34 percentage points lower, which was very close to the intended 30 percentage points. However, as mentioned in subsection 4.2.2, the two NZE speakers seemed very hard to hear in the noise-masked stimuli. This impression was supported by the main effect of accent on transcription accuracy and the very low accuracy for NZE_1 in noise (38.9%). To further investigate the seemingly disproportionate masking of the speakers and to avoid confounds in the main experiment,

another pilot study was conducted with participants from New Zealand. If the results above are due to familiarity, New Zealand participants should perform worse for the TE than the NZE speakers in noise. On the other hand, if individual voice characteristics play a more important role here than (un)familiarity with an accent, then New Zealand participants should also perform worse for NZE_1 and NZE_2 in noise.

4.3 Pilot Study 2

4.3.1 Participants

The dataset for the New Zealand participants included nine complete submissions. The participants were harder to recruit and were mostly sourced via social media, in particular through the Facebook page of the Linguistics and Language Student Society at the University of Auckland. The aim was to recruit participants who were born and raised in New Zealand and had lived there most of their lives. Two participants were excluded from the dataset: one for not being an L1 speaker of English and one for not being from New Zealand originally. Thus, the final dataset for this pilot study included the transcriptions from seven participants. The participants received gift vouchers for completing the study.

All seven participants passed the attention check. When the pilot study was conducted, six participants identified as male and one participant identified as female. The mean age of the participant pool was 22.1 years (sd = 3.2 years), which is unsurprising as most of them were likely students at the University of Auckland. Table 4.2 shows the self-reported familiarity and social network values for the two accents/locations. As can be seen, the participants were much more familiar with the accent spoken in New Zealand and had a larger social network there. The difference between Newcastle (English) and New Zealand (English) reached significance for both measures in simple linear regression models with accent/location as the only predictor: familiarity, $\beta = 4.71$, $SE = 0.05$, $p < .001$; social network, $\beta = 5.43$, $SE = 0.45$, $p < .001$.

	Newcastle (English)	**New Zealand (English)**
Familiarity with Accent	1.3 (1.1)	6.0 (0.0)
Social Network	0.4 (1.1)	5.9 (0.4)

Table 4.2: Self-reported familiarity and social network information: mean and standard deviation from seven-point Likert scales (0 to 6; New Zealand participants)

4.3.2 Stimuli

The stimuli in this pilot study were identical to the ones from Pilot Study 1 (see subsection 4.2.2). Sentences from Stringer and Iverson (2020) were used. There was a total of 320 stimuli: 40 sentences, recorded each with 4 speakers and either presented in quiet or overlaid with multi-speaker babble noise.

4.3.3 Procedure

Once informed consent was in place, the participants completed the headphone check, transcription task and language and demographic background questionnaire. Further details on each part of the pilot study can be found in subsection 4.2.3. During the transcription task, three participants heard the stimuli in quiet first. The remaining four participants transcribed noise-masked sentences first. There were 32 transcription trials, preceded by eight practice trials.

4.3.4 Data Analysis

The same data analysis steps as in Pilot Study 1 were taken for the data in Pilot Study 2 (see subsection 4.2.4). The ordinal logistic mixed effects model included accent, condition and the interaction between the two as fixed predictors as well as random intercepts for participants and items. When considering the results for Pilot Study 2, especially the output of the model, the small number of New Zealand participants must be taken into account, which likely resulted in an underpowered analysis.

4.3.5 Results

The dataset for Pilot Study 2 consisted of 224 transcribed sentences. Figure 4.4 demonstrates that, on average, the New Zealand participants transcribed the sentences more accurately than the North East participants in both conditions. The North East participants' varying performance for the four speakers in noise was also evident in the New Zealand participants (see Figure 4.5). In principle, the pattern was identical for the participants from both locations. Transcriptions were most accurate for TE_1, followed by TE_2, NZE_2 and, finally, NZE_1. However, the values for the New Zealand participants were more levelled in that, for example, the difference between TE_2 and NZE_2 was only 1.2 percentage points,

as compared to 27.8 percentage points for the North East participants. In general, the New Zealand participants performed much better at transcribing NZE_1 and NZE_2 in noise, which strongly suggests an effect of familiarity. Comparing the transcription accuracy for each speaker in quiet versus in noise further demonstrates that noise masking was largely effective for the New Zealand participants although the difference between the two conditions was quite low for TE_1.

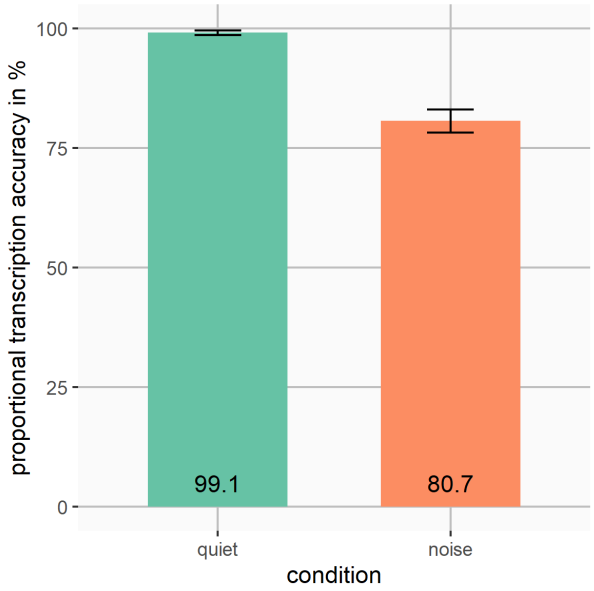


Figure 4.4: Proportional transcription accuracy as mediated by condition (bars: mean, error bars: ±one SE, colour: condition; New Zealand participants)

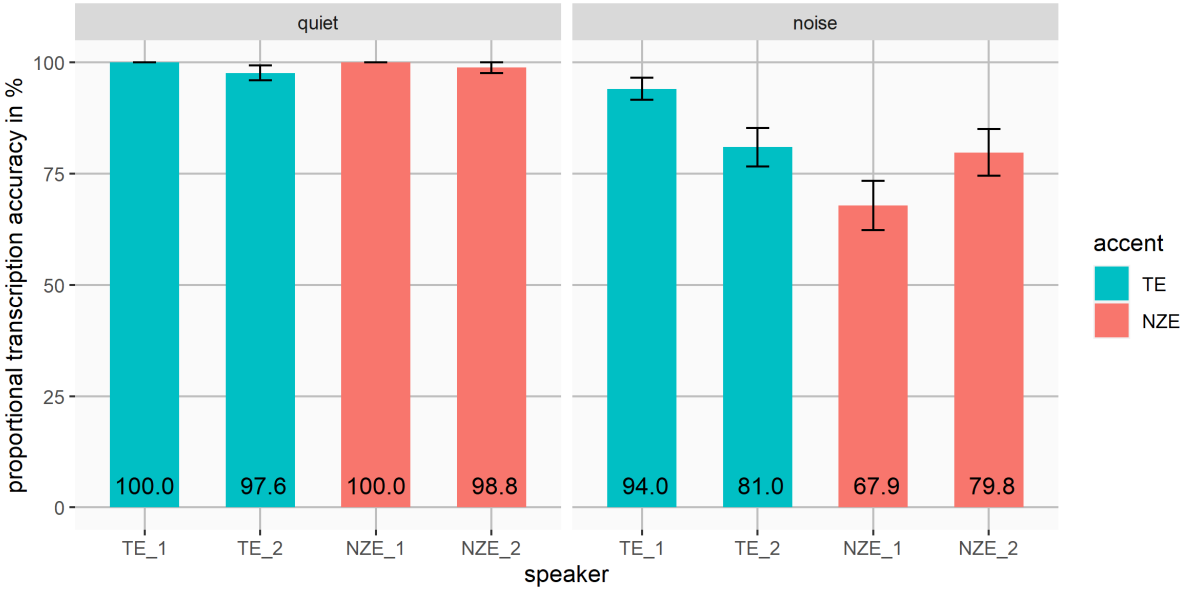


Figure 4.5: Proportional transcription accuracy as mediated by condition and speaker (bars: mean, error bars: ±one SE, colour: accent; New Zealand participants)

One interesting finding was that New Zealand participants transcribed the TE speakers more accurately than the NZE speakers in noise. This difference will be discussed below in conjunction with the results of the statistical analysis. The model that was fitted to the data from Pilot Study 2 was the following:

$$\text{clmm(accuracy\_proportional} \sim \text{accent x condition} + (1|\text{participant}) + (1|\text{item}) \text{)}$$

In line with what was found for Pilot Study 1, model comparisons showed significant effects of accent, $LR(1) = 8.74$, $p = .003$; and condition, $LR(1) = 63.6$, $p < .001$; but not of the interaction between the two, $LR(1) = 2.85$, $p = .091$. Transcription performance was worse in noise, β = -2.90, $SE = 0.88$; and for NZE trials, $\beta = 0.66$[6], $SE = 1.31$.

The predicted likelihoods for each level of transcription accuracy are displayed in Figure 4.6 and were in line with the above observations. Unsurprisingly, there was hardly any variation between the two accents in quiet. In noise, on the other hand, the predicted likelihood of a perfect score was much higher for TE than NZE. Note, however, that the likelihood for this score for the NZE trials was higher in Pilot Study 2 than in Pilot Study 1 (Figure 4.6 versus Figure 4.3). This complies with the finding from above that transcription accuracy was higher for the noise-masked NZE speakers in New Zealand participants compared to participants from the North East of England.

[6] The positive sign of the coefficient shows that the model actually predicted better performance for NZE trials. However, the standard error given for the accent predictor actually creates an interval that does include negative values (0.66 - 1.31 = -0.65), which corresponds better to the pattern in the actual results. It might also be the case that, due to the small size of the dataset, there was a Type S error, which refers to the incorrect estimation of the direction of an effect by a model (Gelman & Carlin, 2014).

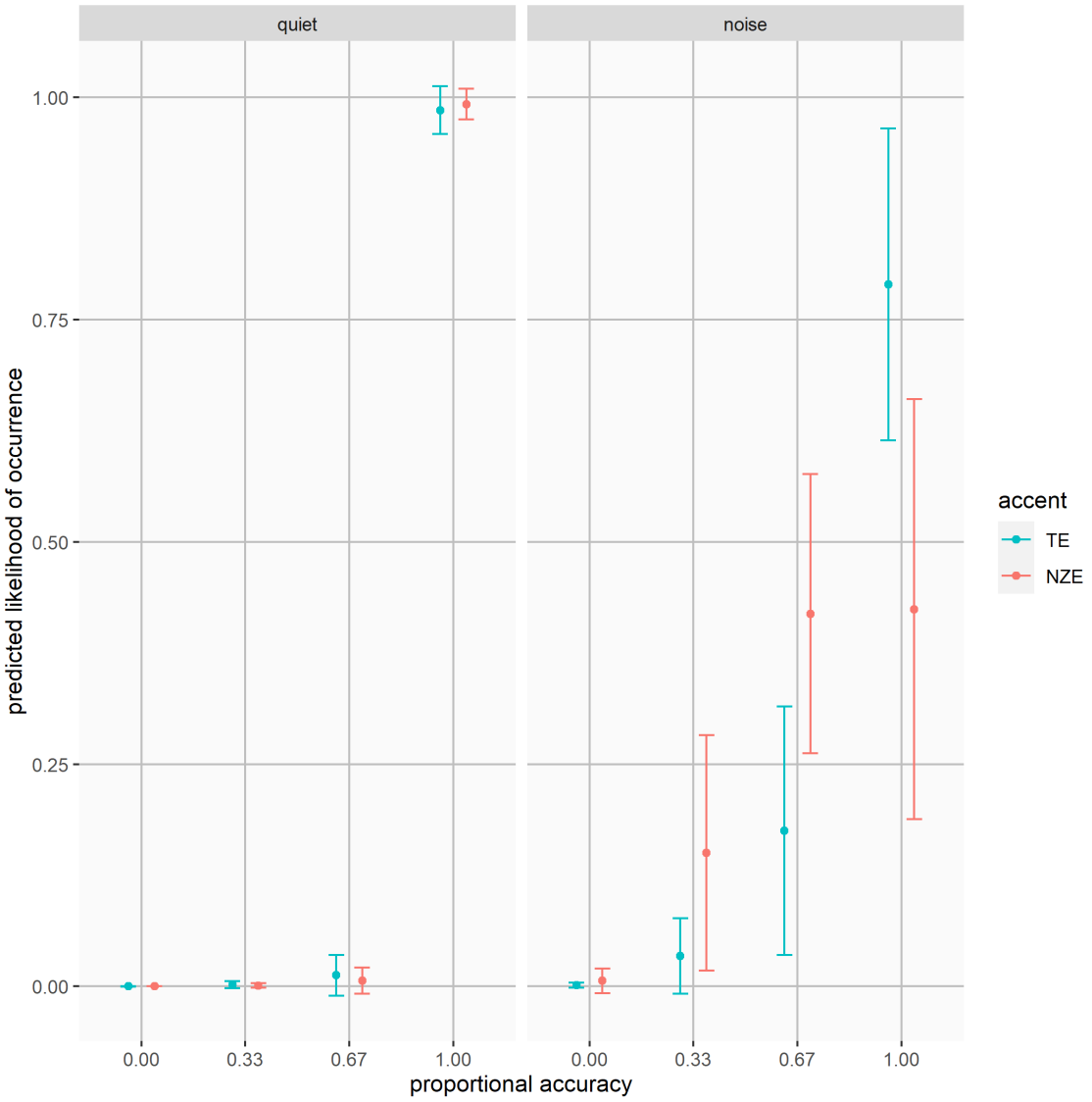


Figure 4.6: Predicted likelihoods of occurrence for each level of proportional accuracy as mediated by accent and condition
(points: means, error bars: confidence intervals, colour: accent; New Zealand participants)

What remains to be addressed is the generally lower transcription accuracy for the NZE as compared to the TE speakers although the participants were New Zealanders who reported that they were more familiar with NZE (see subsection 4.3.1). It is reasonable to expect higher accuracy for transcription in noise for an accent that the participants are more familiar with, which is contrary to what was found in Pilot Study 2. This suggests that, as originally mentioned in subsection 4.2.2, the noise masking was more effective, or to put it differently, more severe for the NZE than the TE speakers. While the interaction between accent and condition did not reach significance in the two pilot studies, it will be decisive to keep this point in mind when discussing the results of the main experiment so as not to confound the effect of accent with an effect of individual voice characteristics in noise. Most importantly

though, a comparison of the data for the North East and the New Zealand participants in the pilot studies showed that, when the NZE speakers were considered in isolation, there still seemed to be a marked effect of familiarity. New Zealanders transcribed NZE better in noise than the participants from the North East.

4.3.6 Summary and Reflections for Main Experiment

Altogether, two pilot studies were conducted that measured transcription performance of TE and NZE sentences presented both in quiet and in noise. Pilot Study 1 was conducted with participants from the North East of England while New Zealanders participated in Pilot Study 2. In both pilot studies, transcription accuracy decreased in noise. Additionally, there was a non-significant trend in the data that the cost of noise was disproportionately higher for NZE sentences. While this interaction did not reach significance, both accent and condition were significant predictors of transcription accuracy in the ordinal logistic regression models.[7]

There are two main conclusions to be drawn from the results of the two pilot studies. First, the noise masking was effective in that it evidently increased the difficulty of the transcription task. Performance was (almost) at ceiling in both pilot studies in quiet but more varied in noise. The increased difficulty of processing speech in noise might lead to the surfacing of stronger effects of familiarity in Experiment 2 as compared to Experiment 1. Second, it seems that the individual voice characteristics of the four speakers used in this study resulted in more effective or severe noise masking for the stimuli produced by NZE_1 and NZE_2. As stated above, this trend in the data was not supported by the ordinal regression as there was no significant interaction between accent and condition. However, it does emerge in Figures 4.2 and 4.5 and is congruent with an auditory inspection of the data. Voice characteristics are very hard to control in an experimental setting, especially because of additional restrictions at the time of recording due to the COVID-19 pandemic (see subsection 3.2.2). While it might be argued that different NZE speakers should be used for the main experiments, this was rather impractical given the circumstances and, more importantly, would not have allowed for a direct comparison of the results from Experiments 1 and 2. This comparison was critical to the research project to further explore the potential interaction between the roles of familiarity and intelligibility in lexical processing and recall.

As a result, the decision was taken to proceed with the available stimuli for Experiment 2. However, in order to mitigate individual voice characteristics in noise as a confound, Experiment

[7] Note, however, the potential Type S error for the effect of accent in Pilot Study 2 (see subsection 4.3.5).

2 was conducted with two participant samples: Tynesiders and New Zealanders. This allowed for comparisons similar to the ones made between the two pilot studies in addition to the contrastive analysis of Experiments 1 and 2 (for the Tyneside participants). The participants, stimuli, data analysis steps and results of Experiment 2 will be addressed in the following subsections.

4.4 Participants

4.4.1 Tyneside Participants

4.4.1.1 Demographic Information

The Tyneside participants[8] for Experiment 2, which was conducted online via LabVanced, were recruited from the SONA participant pool within the School of Education, Communication and Language Sciences (ECLS) at Newcastle University as well as other schools at Newcastle University and Northumbria University Newcastle. Further participants were recruited via social media. The participants received either SONA credits or a gift voucher for their time spent on the study. Similar to Experiment 1, participants born and raised in Tyneside were identified by means of a pre-screening questionnaire and a questionnaire at the end of the experiment.

Altogether, 58 participants were recruited. The data from only one of the SONA participants could be used for the current analysis because the other SONA participants were not native Tynesiders. From the participants recruited from other sources, five were excluded from the final dataset: One participant failed the attention check and four participants indicated that they had or might have speech, hearing or learning disorders. Consequently, the total number of Tyneside participants in Experiment 2 was 42. Table 4.3 shows the distribution of the participants across the four groups from the Latin-square design, which controlled the order of the four speakers in the experimental tasks (see Table 4.9). Given the total number of participants, the distribution was as equal as possible and each group included at least 10 participants.

Group	A	B	C	D
Number of Participants	10	11	10	11

Table 4.3: Distribution of Tyneside participants across groups from Latin-square design

[8] See section 3.3 for further information on the Tyneside participants from Experiment 1.

In terms of gender, 31 participants identified as female, 10 as male and one preferred not to disclose this information. On average, the participants were of similar age to the ones from Experiment 1, with a mean of 28.6 years (sd = 11.6 years).

4.4.1.2 Familiarity Measures

As for Experiment 1, two measures were used to determine the participants' familiarity with TE and NZE: their performance in the accent matching task and their responses on the Likert scales in the questionnaire (see section 4.6). Importantly, for this experiment, the stimuli in the accent matching task were noise-masked and the participants were asked to provide comprehensibility ratings for all speakers both in quiet and in noise. The results from the accent matching task in noise are displayed in Figure 4.7. Accent identification was still above chance ($\frac{1}{9} \approx 0.11$) for both accents although the accuracy rate for the NZE utterances suggests that participants were not performing greatly above chance. Incorrect choices for TE utterances included Birmingham (44.4% of incorrect responses), New Zealand (22.2%), Liverpool (22.2%) and London (11.1%). For NZE utterances, the following labels were incorrectly selected by the participants: Australia (51.7%), London (34.5%), Birmingham (6.9%), Glasgow (3.4%) and no response (1.7%). A simple logistic regression model with accent as the only predictor showed significantly lower accuracy for the NZE speakers, β = -2.92, SE = 0.42, $p < .001$.

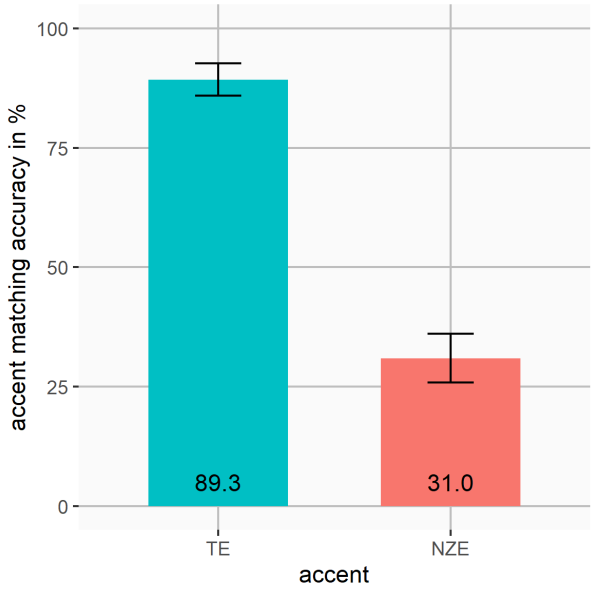


Figure 4.7: Accent matching accuracy in noise as mediated by accent (bars: mean, error bars: ±one SE, colour: accent; Tyneside participants)

The means and standard deviations of the participants' self-reported familiarity and social network measures are shown in Table 4.4. Simple linear regressions including accent/location as the only predictor showed that values were significantly lower for New Zealand (English) on the familiarity scale, β = -4.47, SE = 0.20, $p < .001$; as well as the social network scale, β = -5.47, SE = 0.13, $p < .001$. These results show that the participants were far more familiar with TE than NZE. In fact, only two out of the 42 participants indicated levels of familiarity with NZE of at least 4 on the scale.

	Newcastle (English)	**New Zealand (English)**
Familiarity with Accent	5.9 (0.4)	1.4 (1.3)
Social Network	5.5 (0.8)	0.1 (0.3)

Table 4.4: Self-reported familiarity and social network information: mean and standard deviation from seven-point Likert scales (0 to 6; Tyneside participants)

The comprehensibility ratings[9] for the four speakers in quiet and in noise are provided in Table 4.5. As can be seen, the differences between TE and NZE speakers were less extreme here than for the familiarity and social network measures. A linear regression on the ratings including accent, condition and the interaction between the two as predictors showed significant effects of accent and condition but not of the interaction, β = -0.25, SE = 0.31, p = .423. Comprehensibility ratings were generally lower for NZE speakers, β = -1.89, SE = 0.22, $p < .001$; and when the speakers were presented in noise, β = -1.29, SE = 0.22, $p < .001$.

Speaker	TE_1	TE_2	NZE_1	NZE_2
Comprehensibility Rating (Quiet)	6.0 (0.2)	5.9 (0.4)	4.0 (1.8)	3.9 (2.0)
Comprehensibility Rating (Noise)	4.3 (1.5)	5.0 (1.2)	2.3 (1.5)	2.6 (1.5)

Table 4.5: Self-reported comprehensibility information: mean and standard deviation from seven-point Likert scales (0 to 6; Tyneside participants)

[9] Three Tyneside participants stated in the feedback section of the study that their comprehensibility ratings might not be accurate. For example, some of them were unaware that they would be asked for ratings in quiet **and** in noise and, thus, only provided an overall rating in the first instance. The comprehensibility ratings of these participants were removed such that the data in Table 4.5 represent the responses from 39 participants.

4.4.2 New Zealand Participants

4.4.2.1 Demographic Information

Recruiting participants from New Zealand was more challenging than sourcing Tyneside participants. Different schools at universities in New Zealand were contacted to find potential participants. The link to the sign-up form was also posted in different Facebook groups related to New Zealand. The sign-up form included a pre-screening questionnaire to ensure that individuals were born and raised in New Zealand, had the necessary equipment to complete the study and were at least 18 years old. The participants received a gift voucher after finishing the study.

In total, there were 48 complete submissions by New Zealand participants to the study on LabVanced. 11 participants had to be excluded from the final dataset for the following reasons: Four failed the attention check, three were not born and raised in New Zealand, two were not L1 speakers of English and two more did not complete the recall task properly. As a result, the data from 37 New Zealand respondents were included and will be analysed in subsection 4.8.1. The split of the New Zealand participants across the four groups of the Latin-square design is displayed in Table 4.6. The distribution was fairly even and each group consisted of at least eight participants.

Group	A	B	C	D
Number of Participants	10	11	8	8

Table 4.6: Distribution of New Zealand participants across groups from Latin-square design

When Experiment 2 was conducted, 31 participants identified as female, five as male and there was one participant who preferred not to disclose information on gender. With regards to age, the New Zealand participants were, on average, slightly younger than the Tyneside ones. Here, the mean age was 24.5 years (sd = 7.8 years).

4.4.2.2 Familiarity Measures

Following the procedure from subsection 4.4.1.2, the results from the accent matching task in noise are considered first, before the self-reported scales will be reported on. Figure 4.8 shows the accuracy rates per accent in the accent matching task. Again, accent identification

was generally above the chance rate of 0.11. However, the accuracy for the TE utterances was so low that they were close to the participants merely guessing the origin of the speakers. Incorrect answers for the TE utterances included Glasgow (32.2% of incorrect responses), Birmingham (23.7%), Liverpool (23.7%), London (18.6%) and Australia (1.7%). For the NZE utterances, the following incorrect labels were selected: Australia (68.0%), London (12.0%), Newcastle (8.0%), United States (4.0%), Birmingham (4.0%) and Liverpool (4.0%). A simple logistic regression on accent matching accuracy as mediated by accent showed a significant effect of accent. The accuracy rates were generally higher for NZE utterances, $\beta = 2.04$, $SE = 0.38$, $p < .001$. Thus, participants from New Zealand were much better at identifying their own accent than TE, which corresponds to the reverse results found for participants from Tyneside (see subsection 4.4.1.2).

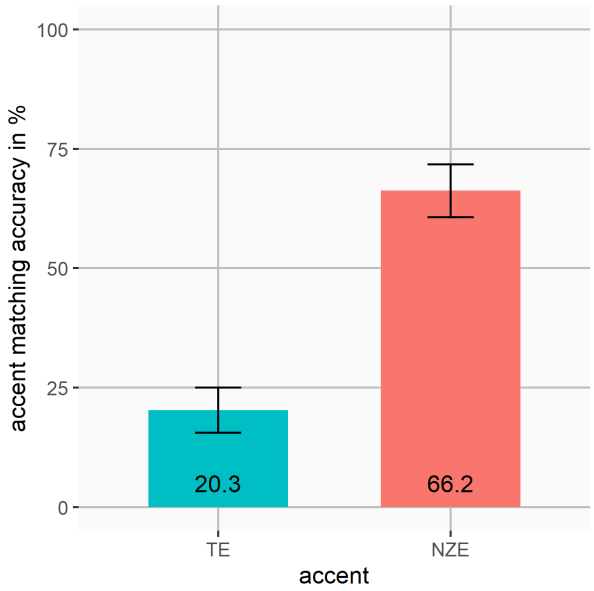


Figure 4.8: Accent matching accuracy in noise as mediated by accent (bars: mean, error bars: ±one SE, colour: accent; New Zealand participants)

As expected, the self-reports of familiarity and social network indicated very high familiarity with New Zealand (English) and very low familiarity with Newcastle (English) (see Table 4.7). The mean values were close to the extreme ends of the scales for both measures. However, six out of the 36 New Zealand participants (16.7%) stated familiarity levels with Newcastle English on the upper end of the scale (i.e. 4 and above). Simple linear regressions including accent/location as the only predictor were run for both scales. The models showed significantly higher values for New Zealand (English) in terms of both familiarity, $\beta = 4.70$, $SE = 0.28$, $p < .001$; and social network, $\beta = 5.30$, $SE = 0.21$, $p < .001$.

	Newcastle (English)	New Zealand (English)
Familiarity with Accent	1.1 (1.5)	5.8 (0.7)
Social Network	0.6 (1.2)	5.9 (0.4)

Table 4.7: Self-reported familiarity and social network information: mean and standard deviation from seven-point Likert scales (0 to 6; New Zealand participants)

The participants' mean comprehensibility ratings and the corresponding standard deviations are given in Table 4.8.[10] The New Zealand participants found their own accent more comprehensible than TE, both in quiet and in noise. Similar to the results for the Tyneside participants, the differences in comprehensibility ratings between TE and NZE speakers were less extreme than the differences on the familiarity and social network scales. The comprehensibility data were fed into a linear regression model that included accent, condition and the interaction between the two. Ratings were generally higher for the NZE speakers, $\beta = 1.70$, $SE = 0.25$, $p < .001$; but lower in noise, $\beta =$ -2.18, $SE = 0.25$, $p < .001$; with no significant interaction, $\beta = 0.49$, $SE = 0.36$, $p = .175$.

Speaker	TE_1	TE_2	NZE_1	NZE_2
Comprehensibility Rating (Quiet)	4.5 (1.5)	3.3 (1.7)	5.6 (0.9)	5.5 (1.1)
Comprehensibility Rating (Noise)	2.0 (1.7)	1.4 (1.4)	3.5 (1.6)	4.2 (1.7)

Table 4.8: Self-reported comprehensibility information: mean and standard deviation from seven-point Likert scales (0 to 6; New Zealand participants)

Overall, self-reported measures and the results from the accent matching task supported the assumption that New Zealand participants were more familiar with NZE than TE and vice versa.

4.5 Stimuli

The stimuli for Experiment 2 were identical to the ones used in Experiment 1, except that they were overlaid with multi-speaker babble noise. Further details on the stimuli and the noise masking procedure have been provided in subsections 3.2.2 and 4.2.2, respectively. In short, the recordings were made with four middle-aged female speakers, two with a TE and two with an NZE accent. For the lexical decision task, there was a total of 224 noise-masked sentences (56 per speaker), which ended either in a real English word or in a nonword. For

[10] The ratings from one participant were excluded because they misunderstood the rating task for this measure (see footnote 9). Thus, Table 4.8 represents the data of 36 New Zealand participants.

the recall task, noise-masked story triplets were used. Each story consisted of three sentences and participants heard two out of three story versions for one trial of the recall task. The only difference between the three versions of a story was a single word (e.g. related: *doctor* → *nurse*; unrelated: *doctor* → *teacher*). For each accent, 20 story triplets (12 experimental and 8 fillers) were selected. For the accent matching task, one short sentence was presented per speaker for a total of four sentences. All stimuli were masked at a SNR of 0dB in Praat. The stimuli did not contain fade-out portions so that latencies could be measured correctly.[11]

4.6 Procedure

For the most part, the procedure of Experiment 2 is the same as the one for Experiment 1 and its pilot study. It was run online and a summary of the components of the experiment is provided below, with more detailed information to be found in subsection 3.2.3.

(1) **Headphone check**: The participants identified the quietest tone in a series of pure tones. To pass the headphone test, the participants had to provide the accurate response in at least five (out of six) trials.[12]

(2) **Lexical decision task**: Following 10 practice sentences, the participants completed 28 trials per speaker for a total of 112 trials. They indicated via key press if the last word of the noise-masked sentence was a real word or a nonword. No more than three word or nonword trials were presented successively for each speaker. Trials were blocked by speaker. The Latin-square design in Table 4.9 was used to counterbalance the order of speakers across participants.

(3) **Recall task**: After 4 practice trials, 40 pairs of noise-masked stories were presented (10 for each speaker). The story pairs either contained no change or a semantically related change or a semantically unrelated change. Trials were blocked by speaker and the order of speakers was determined via a Latin-square design (see Table 4.9).

[11] The measurement of a participant's latency started as soon as the audio file in a trial ended. If the audio files had contained a fade-out portion of 500 ms, the participants might have made their lexical decision within this time frame and, thus, latency measurements would have been artificially low. The option of keeping the fade-out portion and generally subtracting 500 ms from each measurement was disregarded because this could have resulted in negative values and would not have been a valid measure of lexical access.

[12] The headphone check included a minor error. Rather than six, only five different sound files were used. Put differently, one of the sound files was accidentally used twice. However, due to the nature of the task, the duplicate sound file was very hard to spot. Therefore, it remained a reliable check as to whether or not the participants were wearing headphones.

The participants pressed different keys to indicate if the stories were identical or different. If they indicated a difference, they had to report the difference and the original word from the first story. Otherwise, they moved on to the following trial.

(4) **Accent matching task**: The participants heard the four speakers in noise and selected their geographical origin from 9 options (*Glasgow, Liverpool, Birmingham, London, Newcastle, United States, Australia, New Zealand* and *Germany*). The participants could listen to the speakers as many times as necessary and a catch trial was included here to ensure that participants paid attention during the experiment.

(5) **Demographic and language background questionnaire**: The questionnaire asked the participants for demographic information and elicited familiarity, social network and comprehensibility measures. A catch trial was included in the questionnaire.

Group	Practice	Speaker 1	Speaker 2	Speaker 3	Speaker 4
A	TE_1	TE_1	TE_2	NZE_1	NZE_2
B	TE_2	TE_2	TE_1	NZE_2	NZE_1
C	NZE_1	NZE_1	NZE_2	TE_1	TE_2
D	NZE_2	NZE_2	NZE_1	TE_2	TE_1

Table 4.9: Order of speakers across groups in the lexical decision and recall tasks

4.7 Data Analysis

4.7.1 Lexical Decision Task

For the lexical decision task, accuracy and latencies for word trials were considered. Thus, the nonword trials are not included in the results of the following subsections. Unlike PCIBEX, LabVanced recorded latencies directly so they did not have to be calculated in R. Outliers in the dataset were identified based on latencies. Trials with latencies below 250ms and above 2.5 standard deviations from the subject-specific mean were excluded (see Clopper et al., 2016, 92).

Statistical analyses on the data were conducted by means of (generalised) linear mixed effects models with the 'lme4' package in R. The model components are provided below and when the results are presented. In general, log-likelihood comparisons of nested models were used to identify significant predictors. These comparisons were conducted with the 'afex'

package. Even if a predictor did not reach significance, it was not dropped from the model because there was a theoretical justification for each predictor to be in the model. The 'emmeans' package was used to do pairwise comparisons. For multiple pairwise comparisons, Bonferroni corrections were applied.

4.7.1.1 Noise Only Data

This dataset only included data from Experiment 2. The models for the lexical decision task included the following predictors:

- **Accent** was categorical and had two levels: 'TE' and 'NZE'. This predictor was used to investigate the influence of accent familiarity on lexical processing. 'TE' was used as the reference level for this predictor.
- **Participant origin** was categorical and coded if the participants were 'Tyneside participants' (reference level) or 'New Zealand participants'. This predictor was necessary because two participant groups were recruited for Experiment 2.
- **Block** referred to the trial number in question. It was a categorical predictor and showed if a specific trial was part of the first 14 trials produced by a specific speaker ('first half') or the second 14 trials ('second half'). It was included to detect adaptation effects, that is to say changes in accuracy and/or LDLs while listening to one speaker. The reference level for this predictor was set to 'first half'.
- All possible two-way and three-way interactions between **accent**, **participant origin** and **block** were included to track potentially differential performance for either accent and participant group throughout the task.
- **Word frequency** was a continuous predictor (of secondary interest) and was added to see if participants performed better for more frequent targets (see e.g. Grainger, 1990 and Whaley, 1978). This predictor was included to replicate the widely attested effect of word frequency on lexical processing. The word frequency measures were taken from the SUBTLEX database (Brysbaert & New, 2009). Word frequency (in occurrences per million words) was log10-transformed and standardised. The log10-transformation is common for word frequency (see Van Heuven et al., 2014) and the standardisation allowed for easier comparisons across different continuous predictors (Winter, 2019, 89).

- **Speech rate** was a continuous predictor and measured in syllables per second. This factor was included as a control factor to account for differences between the four speakers used in the experiment. Naturally, the voices of the four speakers differed not only in terms of speech rate (see subsection 6.7.2). Hence, the difference between the speakers could also have been accounted for by a random predictor 'speaker'. However, this lead to singularity issues in the models. Thus, a fixed predictor was used instead. Speech rate was also standardised.

Two random effects were included:

- Since several measurements were taken per participant, random intercepts for **participants** were included in the model. Random slopes for accent by participant were also included if this did not result in singularity issues.
- Following the same rationale, random intercepts for the trial targets or **items** were added. As these items were not repeated across accents, random slopes were not included.

4.7.1.2 Experiment 1 versus Experiment 2

The dataset for this analysis included the data from Tyneside participants in Experiments 1 and 2. It allowed for a comparison of speech processing in quiet versus in noise among the same participant population. The models for this analysis included these predictors:

- **Accent**: categorical with the levels 'TE' (reference level) and 'NZE'.
- **Condition** was a categorical predictor with two levels: 'quiet' (reference level) and 'noise'. It coded if the trial in question was presented in noise or in quiet and, thus, allowed for a comparison of speech processing in different listening conditions.
- **Block**: categorical with the levels 'first half' (reference level) and 'second half'.
- All possible two-way and three-way interactions between **accent**, **condition** and **block** were included to track the potentially differential performance for either accent and condition throughout the task.
- **Word frequency**: continuous, log10-transformed and standardised.
- **Speech rate**: continuous, measured in syllables per second and standardised.

Two random effects were included:

- random intercepts for **participants** and, if possible, random slopes for accent by participant
- random intercepts for **items**

4.7.2 Recall Task

Two measures were considered for the recall task. 'Key press accuracy' refers to the participants' decision as to whether or not the two stories they heard were identical or different. 'Correction accuracy' was measured for trials that included different stories and a correct indication thereof via key press by the participants. The correction was recorded as accurate if the participants entered the word that changed in the second story as well as the word that was used in its place in the first story. The participants' responses were first checked manually, for example to correct typos, and then analysed automatically in R.

Statistical analyses followed the procedure from subsection 4.7.1. The model components for the two analyses (noise only and Experiment 1 versus Experiment 2) are provided below.

4.7.2.1 Noise Only Data

All trials in this dataset included noise-masked stimuli. The models consisted of:

- **Accent** was categorical and had two levels: 'TE' and 'NZE'. This predictor was used to test for the effect of familiarity on recall and its reference level was 'TE'.
- **Participant origin** was categorical and referred to 'Tyneside participants' (reference level) versus 'New Zealand participants'. This predictor contrasted the two participant groups recruited for Experiment 2.
- **Story type** coded the (potential) changes between the story pairs. It was a categorical predictor and included the levels 'unchanged', 'related' and 'unrelated'. Story type was added to test the effect of semantic proximity between the changes on recall. Potential differences could be indicative of less-detailed processing or good-enough representations. The reference level used here was 'unchanged'.

- All possible two-way and three-way interactions between **accent**, **participant origin** and **story type** were added to track potential structural differences in the effect of different story types on the processing of familiar versus unfamiliar accents in both participant groups.
- **Block** referred to the trial number in question. In total, there were 10 trials for each of the four speakers. Block was a categorical predictor and coded if a trial was part of the first 5 trials produced by a specific speaker ('first half'; reference level) or the second 5 trials ('second half'). It served to identify effects of adaptation, that is to say changes in key press and/or correction accuracy while listening to one speaker.
- **Speech rate** was a continuous predictor and measured in syllables per second. It was standardised and included to control differences between the four speakers used in the study.

Two random effects were included:

- As several measurements were taken per participant, random intercepts for **participants** were added in the model. If possible, random slopes for accent by participant were also added.
- Following the same rationale, random intercepts for the trial targets or **items** were added. Items were not repeated across accents. Thus, random slopes were not included.

4.7.2.2 Experiment 1 versus Experiment 2

Data from Tyneside participants in Experiments 1 and 2 were used for this analysis to compare speech processing in quiet versus in noise. The models included:

- **Accent**: categorical with the levels 'TE' (reference level) and 'NZE'.
- **Condition**: categorical with the levels 'quiet' (reference level) and 'noise'.
- **Story type**: categorical with the levels 'unchanged' (reference level), 'related' and 'unrelated'.
- All possible two-way and three-way interactions between **accent**, **story type** and **condition** were included to investigate potential structural differences in the effect of different story types on the processing of familiar versus unfamiliar accents in both conditions.

- **Block**: categorical with the levels 'first half' (reference level) and 'second half'.
- **Speech rate**: continuous, measured in syllables per second and standardised.

Two random effects were included:

- random intercepts for **participants** and, if possible, random slopes for accent by participant
- random intercepts for **items**

4.7.3 Accent Matching Task

For this task, the participants' choice of location labels was checked against the true origin of the four speakers (Newcastle and New Zealand). The results for this task have been reported for Tyneside participants in subsection 4.4.1.2 and for New Zealand participants in subsection 4.4.2.2.

4.8 Results

4.8.1 Noise-Only Data

This subsection presents the results from Experiment 2 exclusively. It includes the data from participants from Tyneside and New Zealand. The data is analysed to investigate how intelligibility and accent familiarity affect lexical processing and recall. There were 42 participants from Tyneside and 37 participants from New Zealand (see subsections 4.4.1 and 4.4.2).

4.8.1.1 Lexical Decision Task

4.8.1.1.1 Accuracy

Following the analysis procedure from subsection 4.7.1, 546 outlier responses were excluded. Thus, the final dataset included 3878 lexical decisions for trials with words as the target. Figure 4.9 shows the mean accuracy rates as mediated by accent and participant origin. For the participants from Tyneside, the data patterned as expected. Accuracy was lower for the unfamiliar NZE than the familiar TE. The participants from New Zealand did not mirror this

effect of accent familiarity under adverse listening conditions. Their accuracy for the familiar NZE was also lower than for the unfamiliar TE. However, a comparison across the two groups of participants shows that accuracy rates for TE speakers were lower for participants from New Zealand than for participants from Tyneside. Conversely, the NZE trials were associated with lower accuracy for the Tyneside than the New Zealand participants.

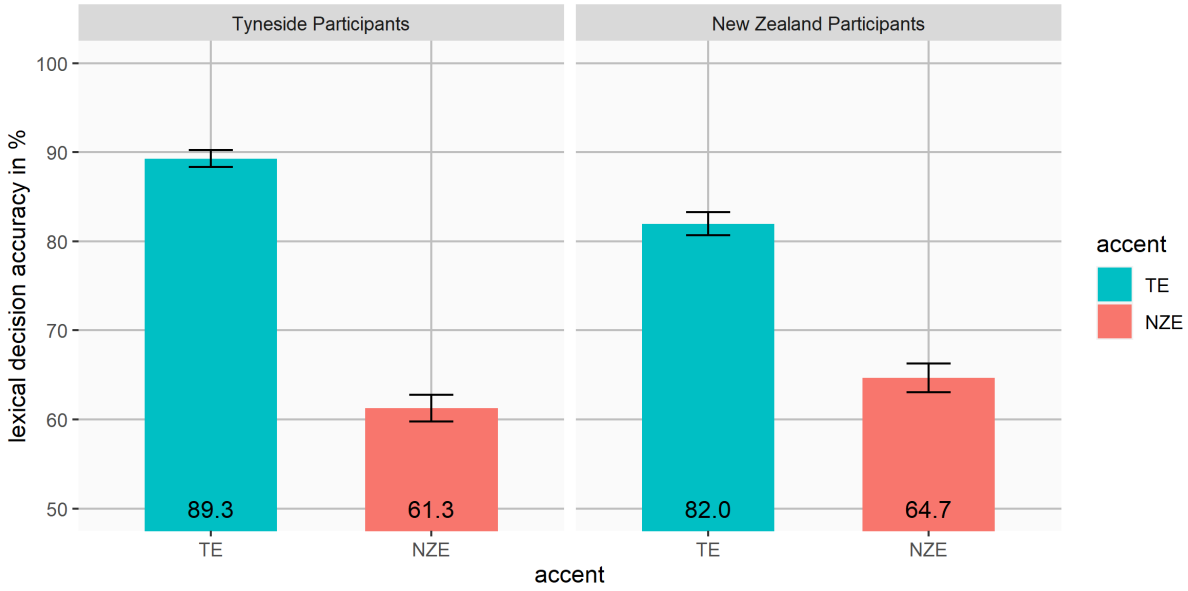


Figure 4.9: Lexical decision accuracy as mediated by participant origin and accent (bars: mean, error bars: ±one SE, colour: accent; noise data)

The accuracy data were fed into the below model, the output of which is given in Table 4.10.

glmer(accuracy_lex_dec ~ accent x particip_origin x block + speech_rate_z + freq_log10_z + (1 + accent|participant) + (1|item))

Predictor	Estimate	Standard Error
(Intercept)	2.78	0.28
Accent (NZE)	-2.29	0.35
Participant Origin (New Zealand)	-0.72	0.27
Block (second half)	-0.10	0.21
Accent (NZE) : Participant Origin (New Zealand)	0.87	0.29
Accent (NZE) : Block (second half)	0.20	0.25
Participant Origin (New Zealand) : Block (second half)	0.02	0.28
Accent (NZE) : Block (second half) : Participant Origin (New Zealand)	0.20	0.35
Speech Rate (z-scored)	0.11	0.08
Word Frequency (log10, z-scored)	0.01	0.15

Table 4.10: Model output: lexical decision accuracy (noise data)

Model comparisons showed significant effects of accent, $\chi^2(1) = 25.91$, $p < .001$; and of the interaction between accent and participant origin, $\chi^2(1) = 16.42$, $p < .001$. The other model components did not emerge as significant: participant origin, $\chi^2(1) = 2.07$, $p = .150$; block, $\chi^2(1) = 0.42$, $p = .516$; speech rate, $\chi^2(1) = 1.90$, $p = .169$; word frequency, $\chi^2(1) < .01$, $p = .948$; the interaction between accent and block, $\chi^2(1) = 2.75$, $p = .097$; the interaction between participant origin and block, $\chi^2(1) = 0.50$, $p = .480$; and the interaction between accent, participant origin and block, $\chi^2(1) = 0.32$, $p = .570$.

Figure 4.10 shows the main effect of accent. Across all participants, performance was generally better for trials presented in TE, β = -1.98, SE = 0.28.[13] Lexical decisions for TE were more accurate by 23.2 percentage points as compared to NZE.

[13] The model output in Table 4.10 shows simple effects of accent, participant origin and block. This is how R generally prints the model coefficients. Here, there is a main effect of accent and the corresponding coefficient of this main effect is provided. This coefficient was calculated manually based on the model output.

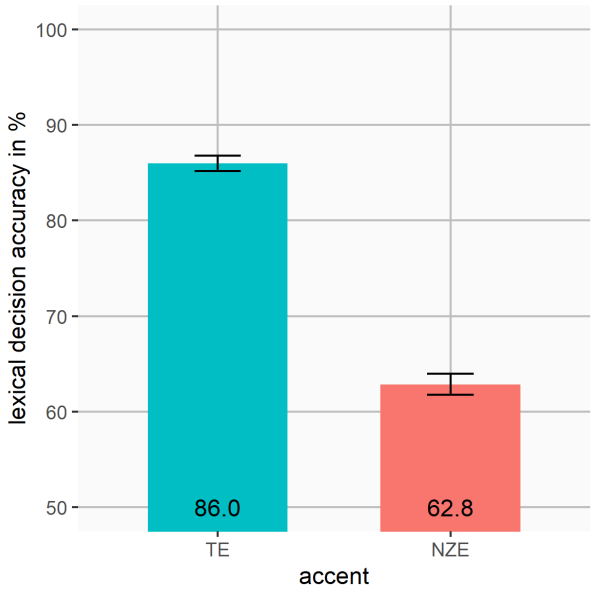
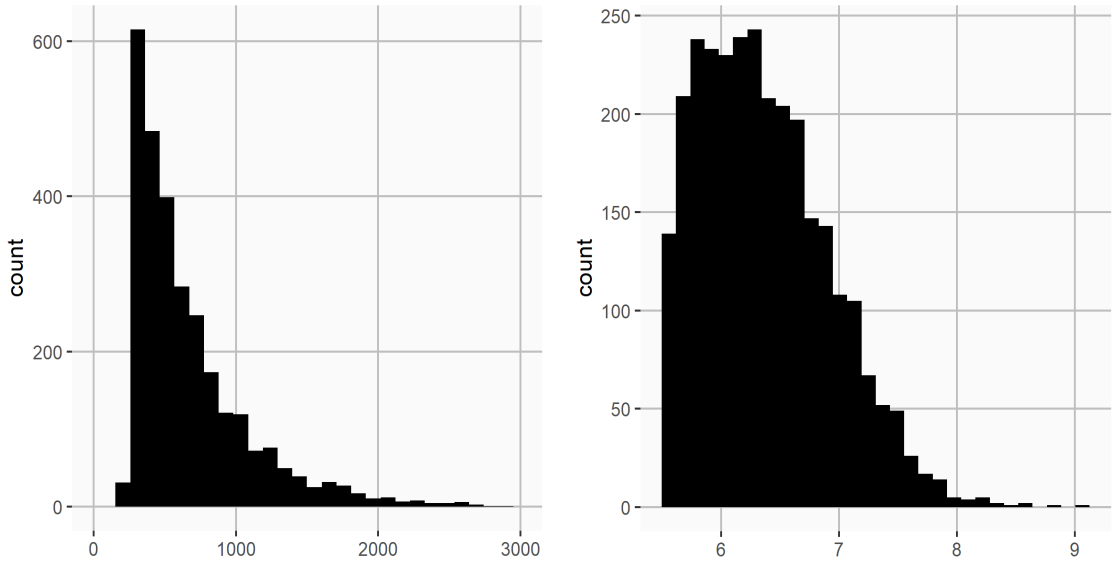


Figure 4.10: Lexical decision accuracy as mediated by accent (bars: mean, error bars: ±one SE, colour: accent; noise data)

The interaction between accent and participant origin is visualised in Figure 4.9 above. The main effect of accent addressed just now is shown by the higher accuracy rates for TE in both participant groups. Pairwise comparisons between TE and NZE showed significant accuracy differences in Tyneside participants (higher accuracy for TE speakers), $z = 6.65$, $p < .001$; as well as New Zealand participants (higher accuracy for TE speakers), $z = 3.74$, $p < .001$. The interaction stems from the size of the difference, which is larger for Tyneside participants. While New Zealand participants performed better for TE than NZE, the difference between the two accents is less marked than for Tyneside participants. This suggests an effect of familiarity under adverse listening conditions, which will be commented on in section 4.9.

4.8.1.1.2 Lexical Decision Latencies

The latencies dataset for the analyses presented in this subsection included all correct lexical decisions for words in Experiment 2 that were not excluded as outliers. This resulted in a dataset consisting of 2889 data points. The distributions of both the raw and the log-transformed latencies are shown in Figure 4.11. As can be seen, the logarithmic transformation yielded a distribution closer to the bell shape of a normal distribution. Towards the left end of the range, there is a hard cut-off at 250 ms/5.52 log-transformed units, which is due to the exclusion criteria for outliers.



(a) raw latencies *(b) log-transformed latencies*

Figure 4.11: Histograms for LDLs (noise data)

The boxplots in Figure 4.12 present the log-transformed LDLs, as mediated by participant origin and speaker. For the Tyneside participants, the NZE speakers were associated with longer latencies and a larger spread of the data. The latencies of participants from New Zealand were roughly equal for TE and NZE trials, both in terms of central tendency and dispersion.

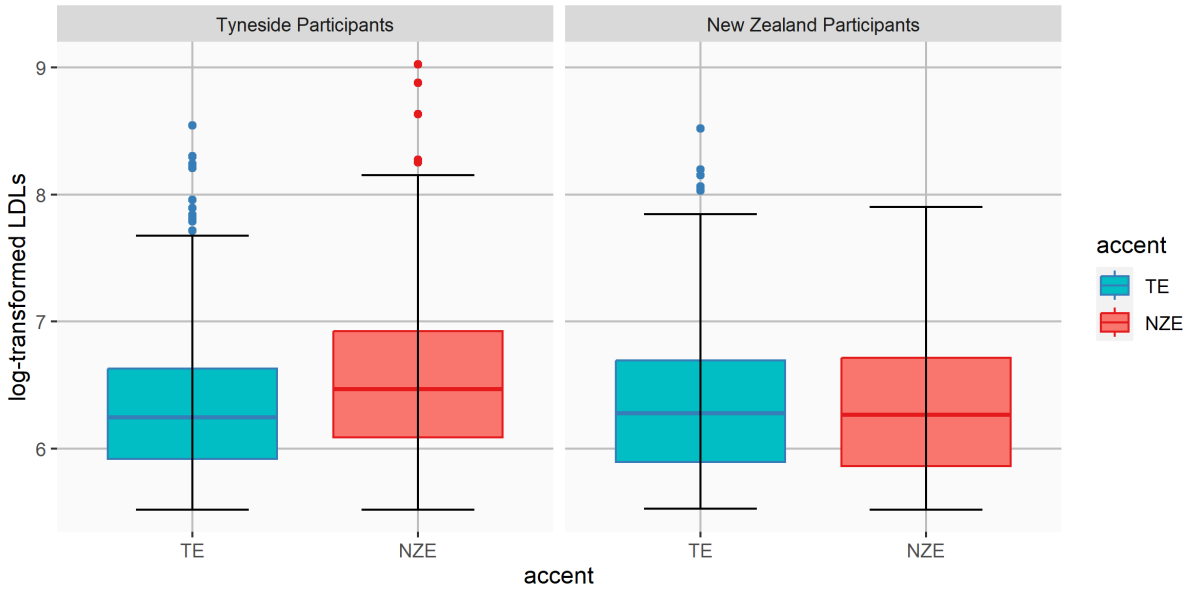


Figure 4.12: Log-transformed LDLs as mediated by participant origin and accent (colour: accent; noise data)

The following model was used to analyse the latencies. Table 4.11 contains the output of the model.

lmer(LDL_log ~ accent x participant_origin x block + speech_rate_z + freq_log10_z + (1 + accent|participant) + (1|item))

Predictor	Estimate	Standard Error
(Intercept)	6.38	0.06
Accent (NZE)	0.16	0.06
Participant Origin (New Zealand)	0.00	0.07
Block (second half)	-0.12	0.03
Accent (NZE) : Participant Origin (New Zealand)	-0.19	0.06
Accent (NZE) : Block (second half)	0.10	0.05
Participant Origin (New Zealand) : Block (second half)	-0.00	0.04
Accent (NZE) : Block (second half) : Participant Origin (New Zealand)	-0.00	0.07
Speech Rate (z-scored)	-0.00	0.01
Word Frequency (log10, z-scored)	-0.02	0.02

Table 4.11: Model output: LDLs (noise data)

The results of model comparisons were significant effects of accent, $\chi^2(1) = 5.03$, $p = .025$; and block, $\chi^2(1) = 15.96$, $p < .001$. Additionally, two interaction terms reached significance: the interaction between accent and participant origin, $\chi^2(1) = 15.37$, $p < .001$; and the interaction between accent and block, $\chi^2(1) = 8.03$, $p = .005$. The other predictors and interactions did not emerge as significant from the model comparisons: participant origin, $\chi^2(1) = 1.98$, $p = .160$; speech rate, $\chi^2(1) = 0.10$, $p = .757$; word frequency, $\chi^2(1) = 0.61$, $p = .433$; the interaction between participant origin and block, $\chi^2(1) < .01$, $p = .947$; and the interaction between accent, block and participant origin, $\chi^2(1) < .01$, $p = .994$.

Figure 4.13 shows the main effect of accent, with LDLs provided in milliseconds rather than log units for ease of interpretation. As can be seen, the participants' reactions were generally faster for the TE speakers, $\beta = 0.13$, $SE = 0.06$ (see footnote 13), with an average difference of 93ms between the two accents.

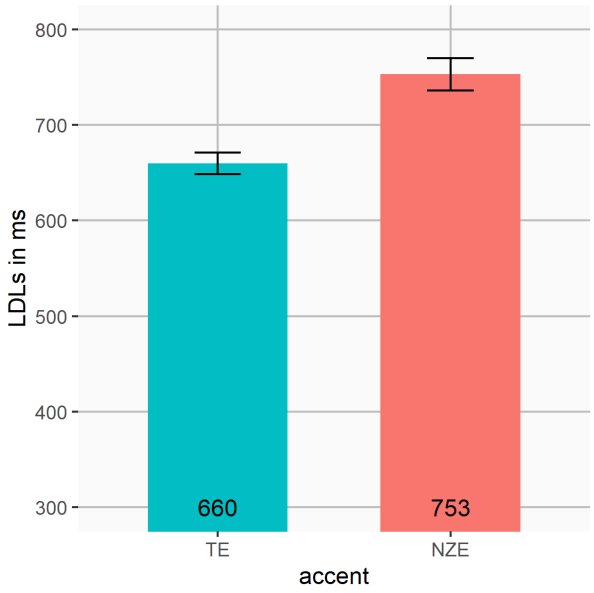


Figure 4.13: LDLs as mediated by accent
(bars: mean, error bars: ±one SE, colour: accent; noise data)

The main effect for block demonstrates that, on average, latencies decreased from the first half of trials presented in a specific speaker's voice to the second half of trials, β = -0.05, SE = 0.06 (see footnote 13). As will be discussed below, this effect was qualified by an interaction with the speaker's accent. The extent of the difference between the first and second half is visualised in Figure 4.14, which shows an average decrease of 30 ms.

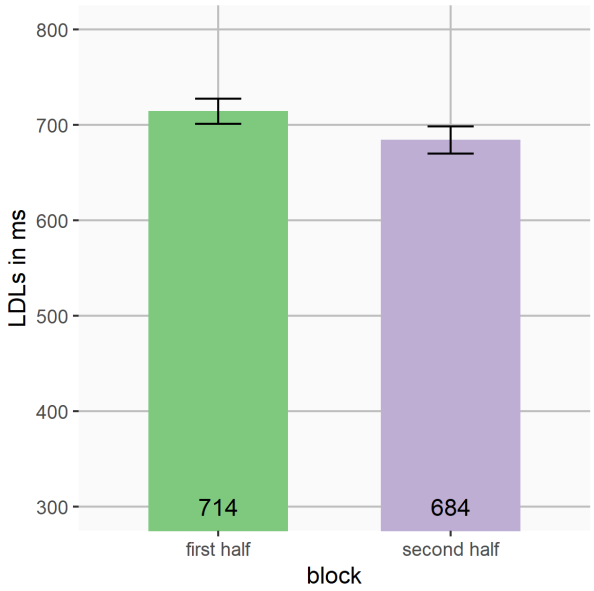


Figure 4.14: LDLs as mediated by block
(bars: mean, error bars: ±one SE, colour: block; noise data)

The interaction between accent and participant origin is already shown in Figure 4.12 above. To ease interpretation, Figure 4.15 displays latencies in milliseconds rather than log units. The graph demonstrates that the main effect of accent described above was due to the variation in the Tyneside participants' data. Pairwise comparisons show significant differences for TE versus NZE trials in Tyneside participants (longer LDLs for NZE speakers), t = 2.55, p = .013; but not in New Zealand participants, t = -0.04, p = .969. As can be seen, latencies for the New Zealand participants were, on average, identical across the two accents. The Tyneside participants, however, were 176 ms slower in making lexical decisions for the NZE as compared to the TE speakers.

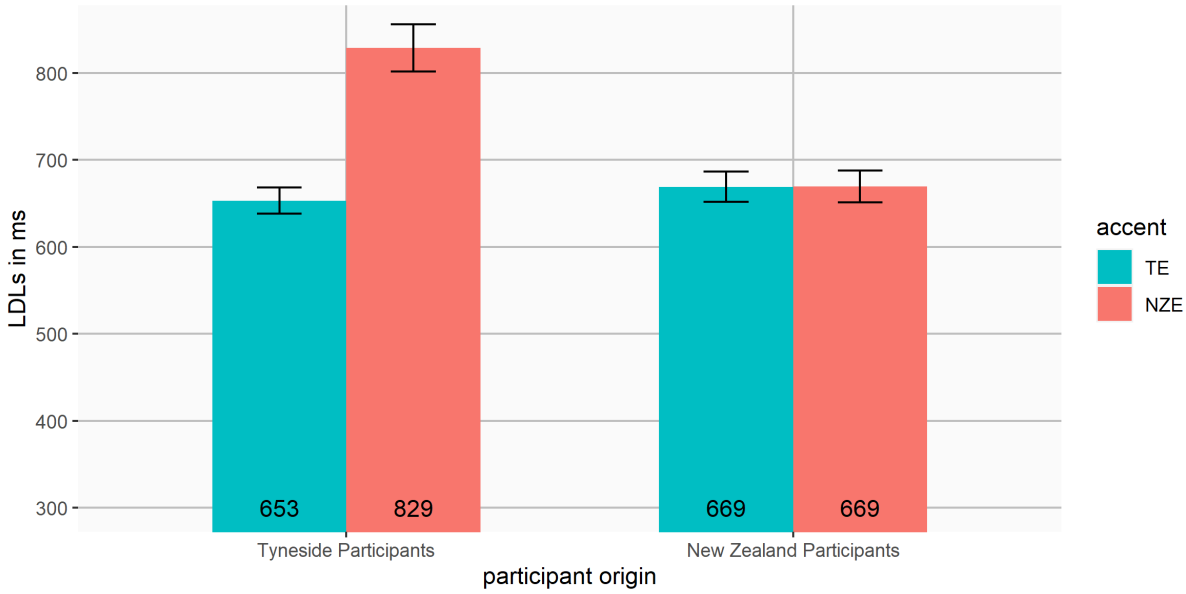


Figure 4.15: LDLs as mediated by accent and participant origin (bars: mean, error bars: ±one SE, colour: accent; noise data)

The significant interaction between accent and block is visualised in Figure 4.16. The previously reported main effects of accent and block are apparent in the bar plot. First, the latencies for the TE speakers were generally shorter than the latencies for the NZE speakers. Second, the average of latencies from the second half of trials was smaller than the average during the first half. Pairwise comparisons between the first and second half of trials showed significant differences for TE trials (longer LDLs during the first half), t = 5.26, $p < .001$; but not for NZE trials, t = 0.77, p = .442. Thus, latencies decreased from the first to the second half for TE speakers only. Importantly, this interaction is not qualified by where the participants were from. Put differently, latencies decreased for the TE trials in participants from both Tyneside and New Zealand.

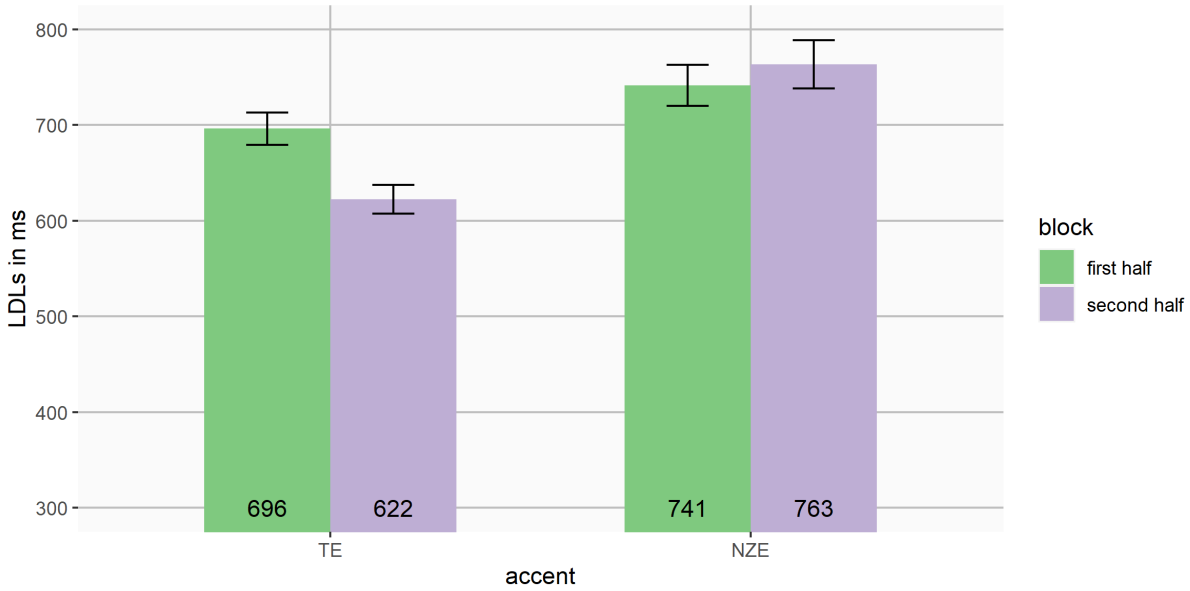


Figure 4.16: LDLs as mediated by accent and block (bars: mean, error bars: ±one SE, colour: block; noise data)

4.8.1.2 Recall Task

4.8.1.2.1 Key Press Accuracy

The dataset for the accuracy of the key press included 1896 observations. In total, there were 1546 correct (81.5%) and 350 incorrect responses (18.5%). The average accuracy rates, as mediated by accent, story type, and participant origin, are shown in Figure 4.17. The effect of where the participants came from was evident for the TE speakers in that Tyneside participants generally made fewer errors than the New Zealand participants for this accent. The reverse pattern was true for the NZE speakers in that participants from New Zealand performed better for these speakers than for participants from Tyneside. Overall, key press accuracy was best for stories that did not include a change.

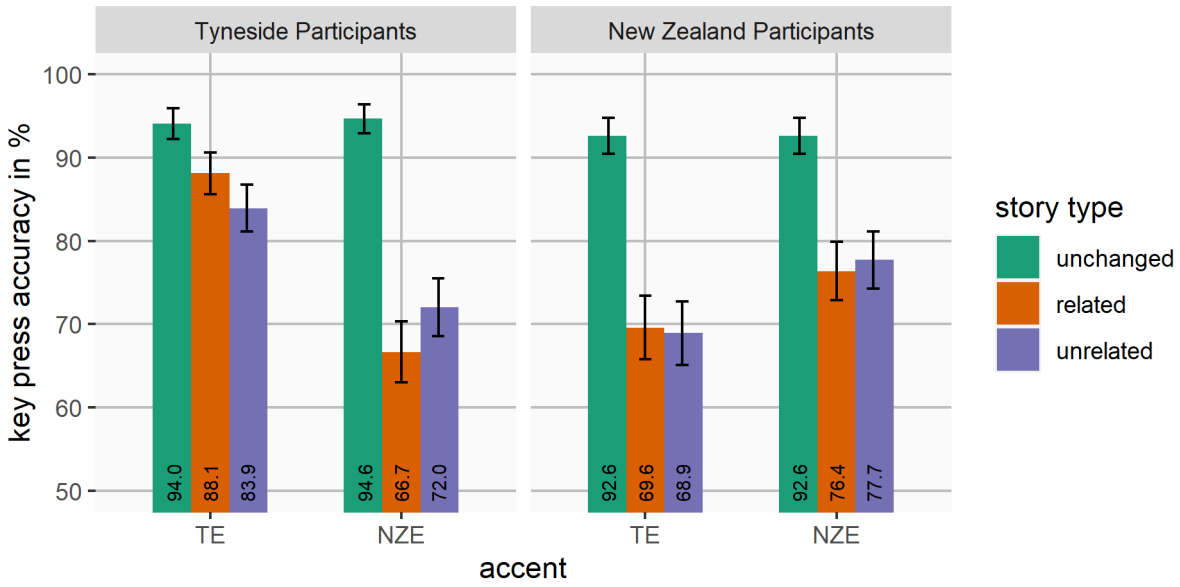


Figure 4.17: Key press accuracy as mediated by accent, story type and participant origin (bars: mean, error bars: ±one SE, colour: story type; noise data)

The model fitted to the key press data is provided below, along with its output in Table 4.12.

$$\text{glmer}(\ \text{accuracy\_key\_press} \sim \text{accent} \times \text{story\_type} \times \text{partic\_origin} + \text{block} + \text{speech\_rate\_z} + (1 + \text{accent}|\text{participant}) + (1|\text{item})\ )$$

Predictor	Estimate	Standard Error
(Intercept)	3.44	0.44
Accent (NZE)	-0.28	0.60
Story Type (related)	-0.91	0.41
Story Type (unrelated)	-1.24	0.40
Participant Origin (New Zealand)	-0.16	0.52
Accent (NZE) : Story Type (related)	-1.60	0.57
Accent (NZE) : Story Type (unrelated)	-0.92	0.56
Accent (NZE) : Participant Origin (New Zealand)	-0.16	0.68
Story Type (related) : Participant Origin (New Zealand)	-1.14	0.56
Story Type (unrelated) : Participant Origin (New Zealand)	-0.83	0.55
Accent (NZE) : Story Type (related) : Participant Origin (New Zealand)	2.07	0.79
Accent (NZE) : Story Type (unrelated) : Participant Origin (New Zealand)	1.58	0.78
Block (second half)	0.02	0.14
Speech Rate (z-scored)	0.26	0.15

Table 4.12: Model output: key press accuracy (noise data)

The model comparisons showed a main effect of story type, $\chi^2(2) = 108.00$, $p < .001$. Two interaction terms further reached significance: the two-way interaction between accent and participant origin, $\chi^2(1) = 9.54$, $p = .002$; and the three-way interaction between accent, story type and participant origin, $\chi^2(2) = 6.62$, $p = .037$. The other model components did not emerge as significant: accent, $\chi^2(1) = 2.32$, $p = .127$; participant origin, $\chi^2(1) = 1.26$, $p = .262$; block, $\chi^2(1) = 0.02$, $p = .879$; speech rate, $\chi^2(1) = 3.17$, $p = .075$; the interaction between accent and story type, $\chi^2(2) = 2.69$, $p = .260$; and the interaction between story type and participant origin, $\chi^2(2) = .08$, $p = .963$.

Key press accuracy rates per story type are displayed in Figure 4.18. Pairwise comparisons between the story types showed significant accuracy differences between unchanged and related stories (higher key press accuracy for unchanged stories), $z = 8.70$, $p < .001$; as well as unchanged and unrelated stories (higher key press accuracy for unchanged stories), $z = 8.57$, $p < .001$; but not between related and unrelated stories, $t = -0.27$, $p > .999$. As can be seen in the graph, accuracy was close to 20 percentage points higher for the unchanged stories as compared to the other story types.

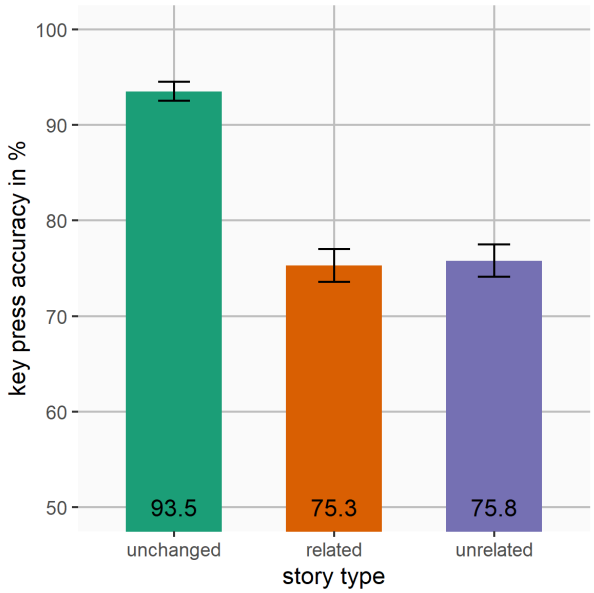


Figure 4.18: Key press accuracy as mediated by story type (bars: mean, error bars: ±one SE, colour: story type; noise data)

Figure 4.19 helps to interpret the significant interaction between accent and participant origin. Through pairwise comparisons, significant differences between TE and NZE trials were found for Tyneside participants (higher key press accuracy for TE speakers), $z = 2.70$, $p = .007$; but not for New Zealand participants, $z = 0.15$, $p = .879$. For participants from Tyneside, the accuracy was 10.9 percentage points higher when they heard their own accent. While it did not reach significance, the data from New Zealand participants show an accent advantage of 5.2 percentage points.

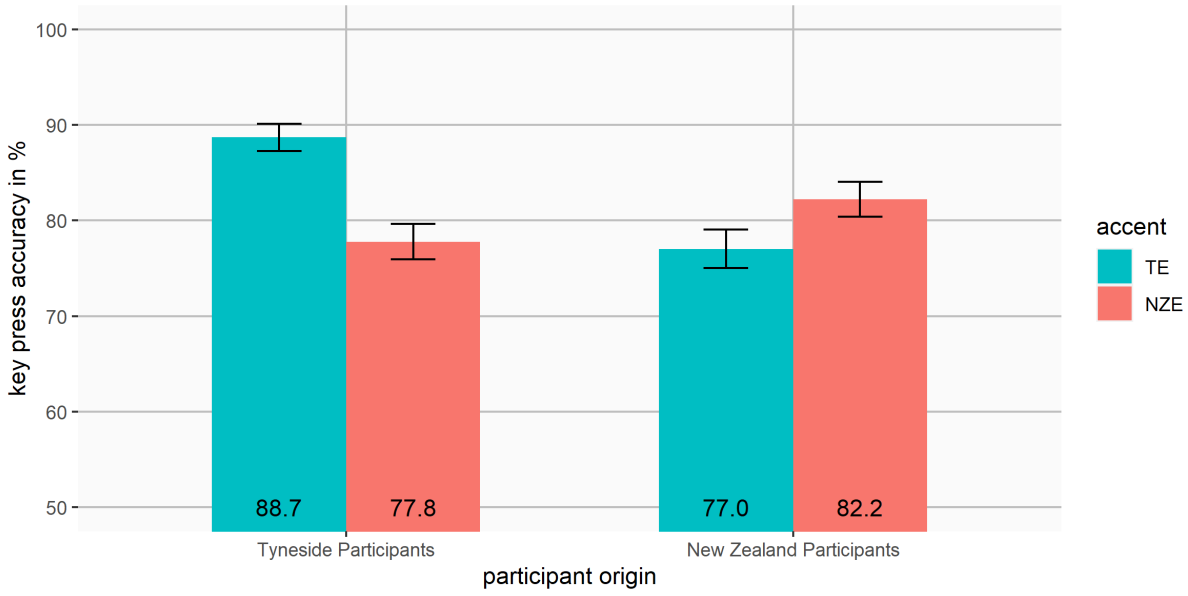


Figure 4.19: Key press accuracy as mediated by accent and participant origin (bars: mean, error bars: ±one SE, colour: accent; noise data)

The three-way interaction between accent, story type and participant origin is shown in Figure 4.17 above. Here, pairwise comparisons between the story types showed the following significant differences:[14]

- TE trials presented to Tyneside participants: unchanged versus unrelated (higher key press accuracy for unchanged stories), $z = 3.11$, $p = .006$.
- NZE trials presented to Tyneside participants: unchanged versus related (higher key press accuracy for unchanged stories), $z = 6.26$, $p < .001$; and unchanged versus unrelated (higher key press accuracy for unchanged stories), $z = 5.38$, $p < .001$.
- TE trials presented to New Zealand participants: unchanged versus related (higher key press accuracy for unchanged stories), $z = 5.23$, $p < .001$; and unchanged versus unrelated (higher key press accuracy for unchanged stories), $z = 5.28$, $p < .001$.
- NZE trials presented to New Zealand participants: unchanged versus related (higher key press accuracy for unchanged stories), $z = 3.99$, $p < .001$; and unchanged versus unrelated (higher key press accuracy for unchanged stories), $z = 3.57$, $p = .001$.

As can be seen, significant differences emerge because of the much higher key press accuracy for unchanged stories. Importantly, the differences in accuracy between unchanged versus

[14] To maintain claritiy, only pairwise comparisons that reached significance are reported here and in the following for three-way interactions in the recall task.

changed stories are mediated by the specific combination of accent and participant origin. For example, the difference was at approximately 25 percentage points for NZE trials presented to Tyneside participants. For NZE trials presented to New Zealand participants, on the other hand, the difference was less than 20 percentage points.

4.8.1.2.2 Correction Accuracy

For the correction accuracy measure, only those trials were relevant for which the participants correctly indicated via key press that there was a difference between the two stories. Thus, unchanged stories are not considered here and the dataset consisted of 955 responses. 797 (83.5%) of them were correct and 158 (16.5%) were incorrect. Accuracy rates for each combination of accent, story type and participant origin are presented in Figure 4.20. Overall, correction accuracy was fairly similar for both story types. However, there was a noticeable decrease in accuracy for unrelated stories presented in TE to participants from Tyneside.

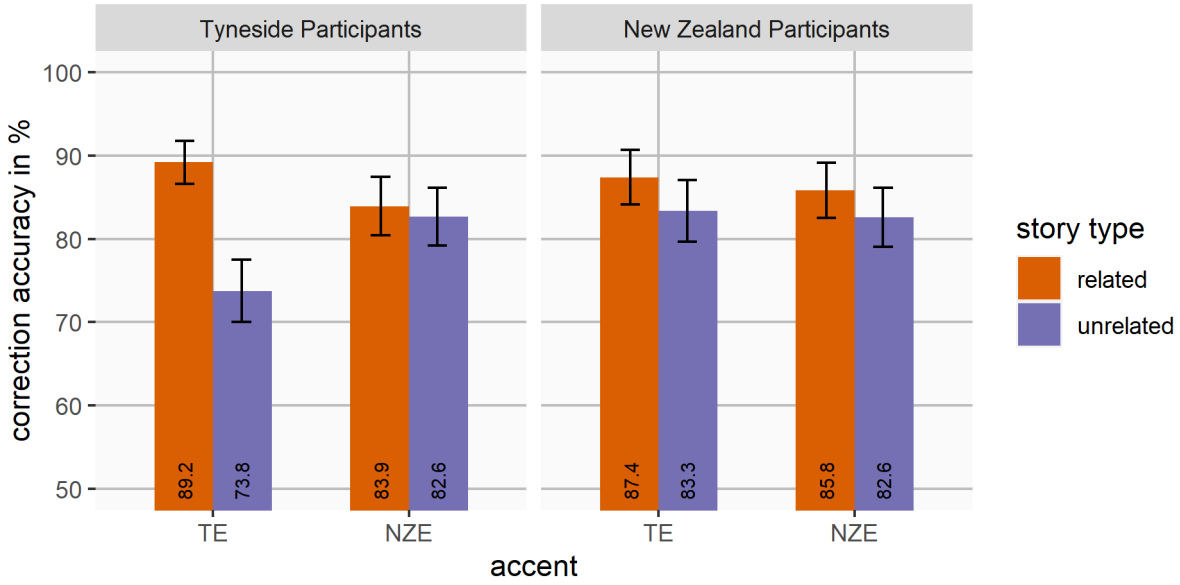


Figure 4.20: Correction accuracy as by mediated accent, story type and participant origin (bars: mean, error bars: ±one SE, colour: story type; noise data)

The below model was used to analyse the correction accuracy data. Since unchanged stories were not relevant for this measure, the story type predictor only had two levels: 'related' (reference level) and 'unrelated'. Table 4.13 provides the output of the model.

$$\text{glmer(correction\_accuracy} \sim \text{accent x story\_type x partic\_origin + block + speech\_rate\_z +}$$
$$\text{(1 + accent|participant) + (1|item))}$$

Predictor	Estimate	Standard Error
(Intercept)	2.60	0.51
Accent (NZE)	-0.28	0.68
Story Type (unrelated)	-1.06	0.40
Participant Origin (New Zealand)	-0.43	0.54
Accent (NZE) : Story Type (unrelated)	0.87	0.56
Accent (NZE) : Participant Origin (New Zealand)	0.66	0.63
Story Type (unrelated) : Participant Origin (New Zealand)	1.08	0.61
Accent (NZE) : Story Type (unrelated) : Participant Origin (New Zealand)	-1.40	0.83
Block (second half)	-0.24	0.21
Speech Rate (z-scored)	-0.26	0.22

Table 4.13: Model output: correction accuracy (noise data)

The only predictor to emerge as significant from the model comparisons was story type, $\chi^2(1) = 4.36$, $p = .037$. The other predictors and interactions were not significant: accent, $\chi^2(1) = 0.06$, $p = .804$; participant origin, $\chi^2(1) = 0.08$, $p = .783$; block, $\chi^2(1) = 1.43$, $p = .232$; speech rate, $\chi^2(1) = 1.36$, $p = .244$; the interaction between accent and story type, $\chi^2(1) = 0.15$, $p = .694$; the interaction between accent and participant origin, $\chi^2(1) = 0.01$, $p = .923$; the interaction between story type and participant origin; $\chi^2(1) = 0.84$, $p = .358$; and the interaction between accent, story type and participant origin, $\chi^2(1) = 2.87$, $p = .090$.

Figure 4.21 shows correction accuracy rates per story type. Correction accuracy was generally lower for unrelated stories, β = -1.15, SE = 0.55. As the graph shows, the correction accuracy was at 86.8% for related and 80.2% for unrelated stories. Thus, semantically unrelated changes were corrected with 6.6 percentage points less accuracy.

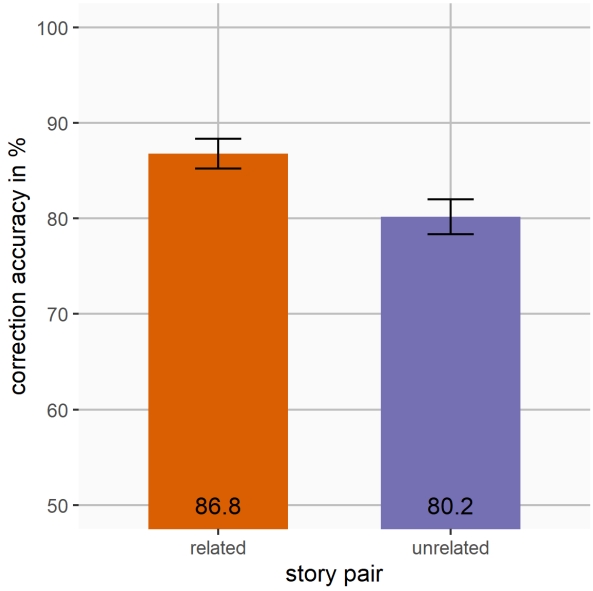


Figure 4.21: Correction accuracy as mediated by story type (bars: mean, error bars: ±one SE, colour: story type; noise data)

4.8.2 Experiment 1 versus Experiment 2 (Tynesider Data)

As explained in section 4.7, the comparison of the results from Experiments 1 and 2 allows for an analysis of how manipulations of intelligibility affect the processing of an (un)familiar accent. Specifically, the lexical processing and recall of TE and NZE by Tyneside participants is analysed here. The analysis does not include the data from the New Zealand participants. Altogether, the dataset used in this subsection contains 85 Tyneside participants: 43 from Experiment 1 and 42 from Experiment 2.

4.8.2.1 Lexical Decision Task

4.8.2.1.1 Accuracy

Following the analysis procedure from subsection 4.7, 433 outlier responses were excluded. The resulting dataset consisted of 4327 reactions to words in the lexical decision task. The mean accuracy rates, broken down by accent and condition, are shown in Figure 4.22. The results for the trials in quiet have been analysed in depth in subsection 3.7.1. The main concern here is to compare the accuracy of the lexical decisions between the quiet and the noise condition. The bar plot shows that accuracy decreased in noise for both accents. Importantly, this decrease was much higher for NZE. For the unfamiliar accent, there was an accuracy

difference of 34.2 percentage points between the two conditions. For TE, this difference was at 10.1 percentage points. In both conditions, accuracy was lower for NZE than TE.

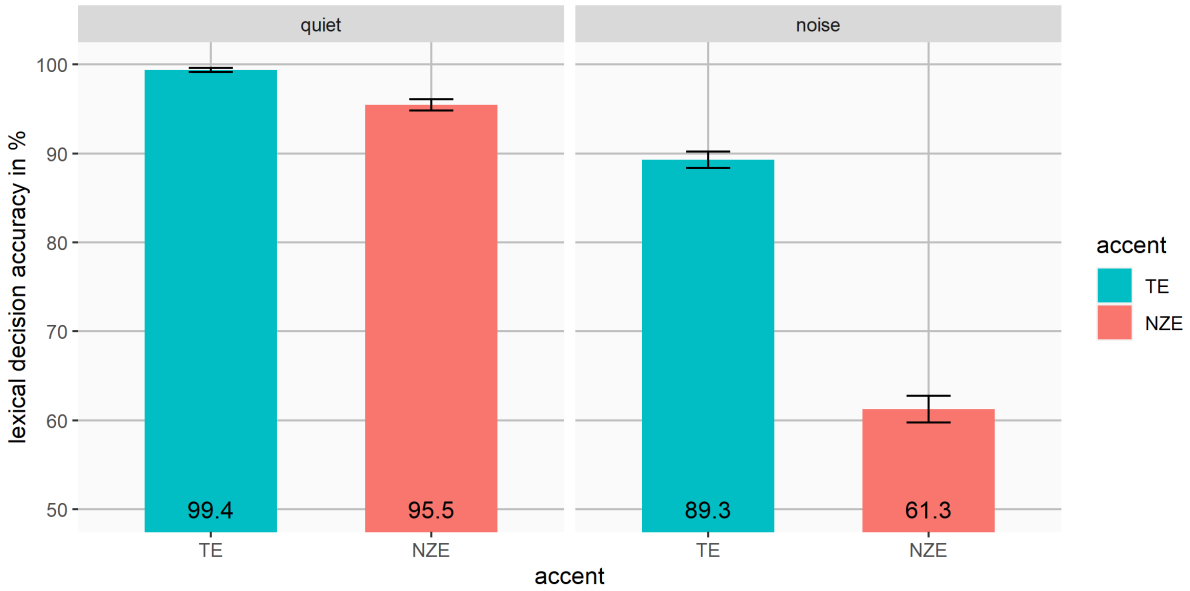


Figure 4.22: Lexical decision accuracy as mediated by accent and condition (bars: mean, error bars: ±one SE, colour: accent; Tynesider data)

The accuracy data were analysed by means of the following model. The model output is shown in Table 4.14.

$$\text{glmer}(\ \text{accuracy\_lex\_dec} \sim \text{accent x condition x block} + \text{speech\_rate\_z} + \text{freq\_log10\_z}$$
$$(1 + \text{accent}|\text{participant}) + (1|\text{item})\ )$$

Predictor	Estimate	Standard Error
(Intercept)	5.61	0.57
Accent (NZE)	-1.64	0.67
Condition (noise)	-2.61	0.54
Block (second half)	1.33	0.85
Accent (NZE) : Condition (noise)	-0.90	0.62
Accent (NZE) : Block (second half)	-1.97	0.91
Condition (noise) : Block (second half)	-1.49	0.88
Accent (NZE) : Condition (noise) : Block (second half)	2.25	0.95
Speech Rate (z-scored)	0.13	0.10
Word Frequency (log10, z-scored)	0.16	0.16

Table 4.14: Model output: lexical decision accuracy (Tynesider data)

The model comparisons showed that lexical decision accuracy was significantly affected by accent, $\chi^2(1) = 35.21$, $p < .001$; condition, $\chi^2(1) = 110.66$, $p < .001$; and the three-way interaction between accent, condition and block, $\chi^2(1) = 6.19$, $p = .013$. The other model components did not emerge as significant: block, $\chi^2(1) = 0.48$, $p = .490$; speech rate, $\chi^2(1) = 1.63$, $p = .202$; word frequency, $\chi^2(1) = 0.93$, $p = .334$; the interaction between accent and condition, $\chi^2(1) = 0.19$, $p = .663$; the interaction between accent and block, $\chi^2(1) = 3.47$, $p = .062$; and the interaction between condition and block, $\chi^2(1) = 0.62$, $p = .433$.

The mean accuracy rates for each accent are shown in Figure 4.23, which demonstrates that accuracy was almost at ceiling for the TE trials (94.4%) but significantly lower for the NZE trials (78.8%), β = -1.42, $SE = 0.66$ (see footnote 13).

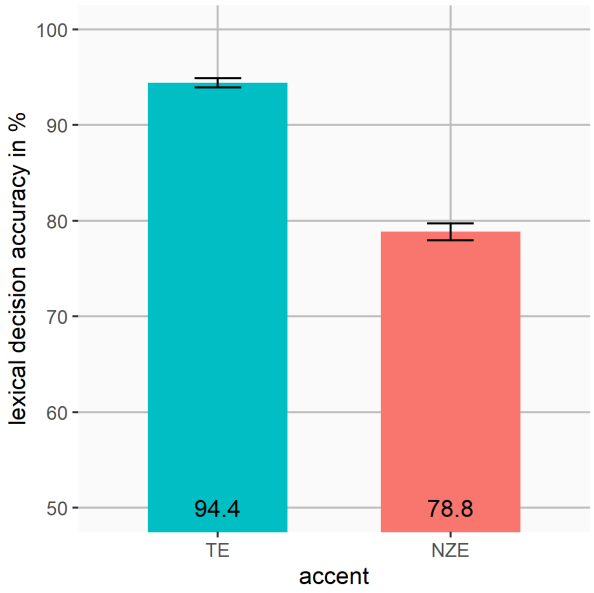


Figure 4.23: Lexical decision accuracy as mediated by accent (bars: mean, error bars: ±one SE, colour: accent; Tynesider data)

With regards to condition, lexical decision accuracy was significantly lower in noise than in quiet, β = -2.15, SE = 0.66 (see footnote 13). As can be seen in Figure 4.24, the accuracy averaged at 97.4% in quiet and at 75.5% in noise. Thus, there was a difference of 21.9 percentage points between the two conditions and performance was practically at ceiling when the stimuli were presented in quiet.

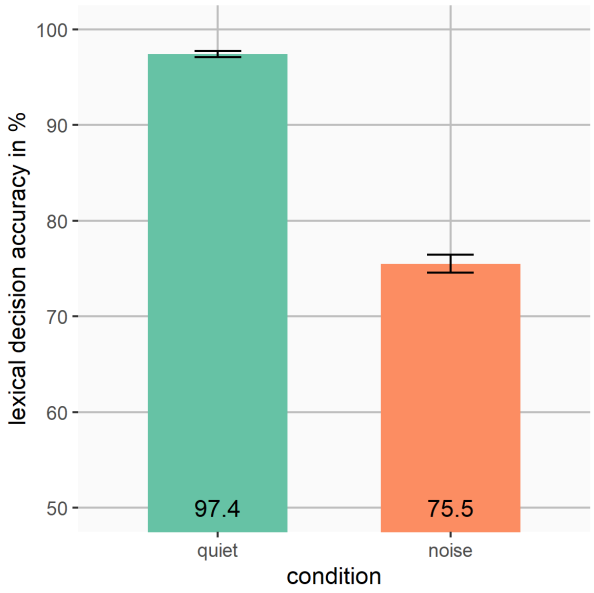


Figure 4.24: Lexical decision accuracy as mediated by condition (bars: mean, error bars: ±one SE, colour: condition; Tynesider data)

Figure 4.25 visualises the interaction between accent, condition and block. The main effects of accent and condition are apparent here in that accuracy rates were higher for TE trials and in quiet. Importantly, there was a significant drop in accuracy for NZE trials presented in noise. Pairwise comparisons between TE and NZE show significant differences between the accents for all combinations of condition and block: higher accuracy for quiet TE stimuli in the first half, $z = 2.45$, $p = .014$; higher accuracy for quiet TE trials in second half, $z = 4.31$, $p < .001$; higher accuracy for noise-masked TE trials in first half, $z = 6.30$, $p < .001$; and higher accuracy for noise-masked TE trials in second half, $z = 5.67$, $p < .001$. The noise-masking was especially detrimental for the unfamiliar accent. Here, the accuracy rates were around 60%, compared to the near-ceiling performance for both accents in quiet.

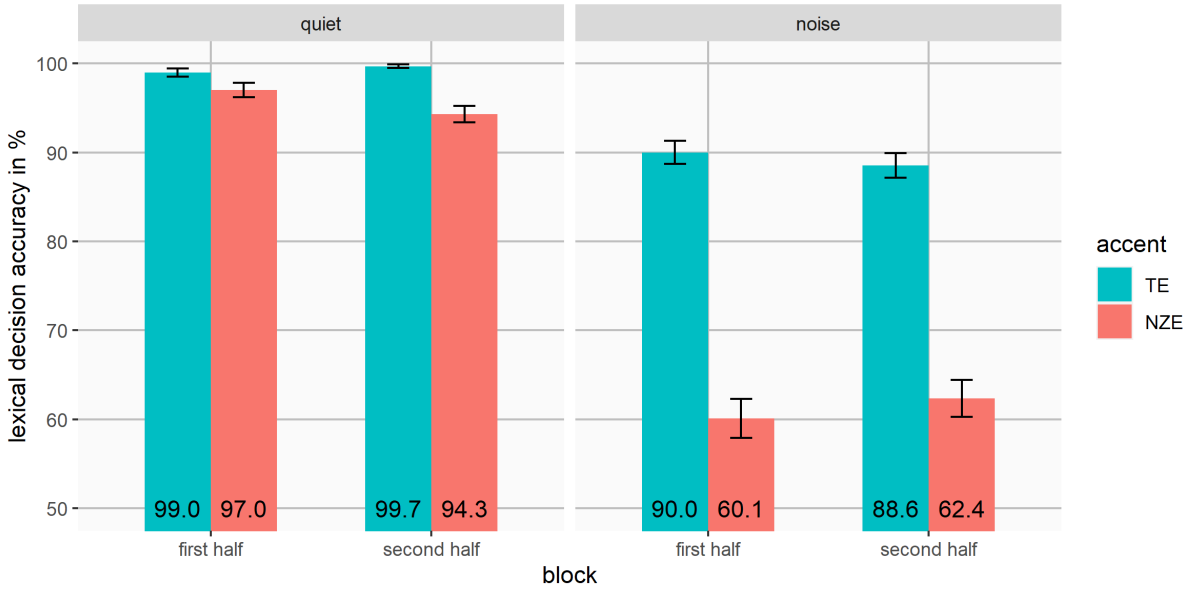
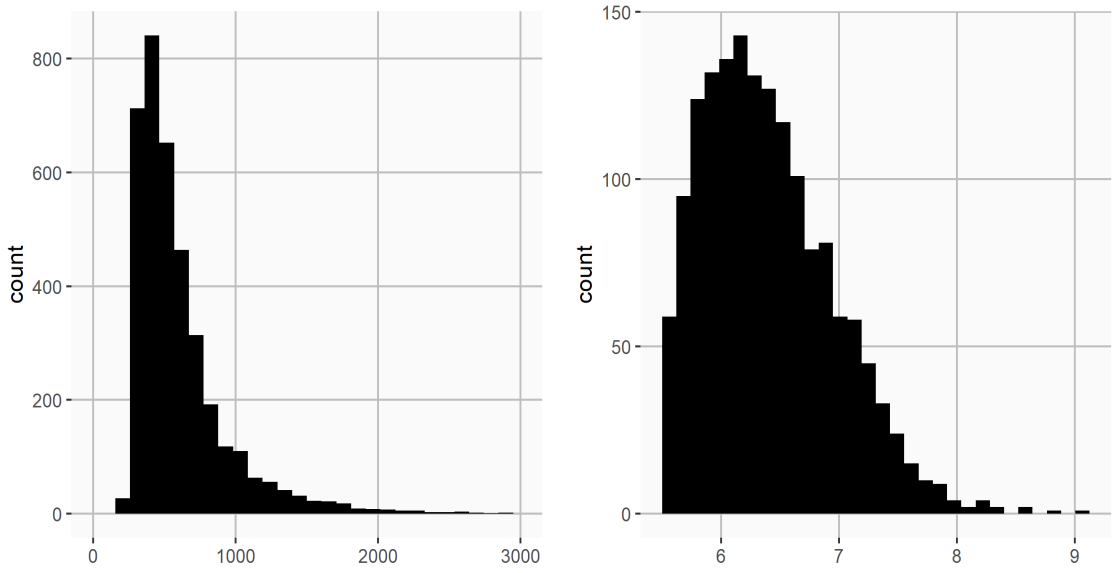


Figure 4.25: Lexical decision accuracy as mediated by accent, condition and block (bars: mean, error bars: ±one SE, colour: accent; Tynesider data)

4.8.2.1.2 Lexical Decision Latencies

The dataset analysed in this subsection included latencies from correct lexical decisions made by Tyneside participants in Experiments 1 and 2. Out of these 4107 latency measurements, 354 were excluded as outliers, which resulted in a final dataset of 3753 observations. The raw latencies were log-transformed. The results of this non-linear transformation are presented in Figure 4.26. The histograms demonstrate that the transformation resulted in

a distribution closer to the bell shape of the normal distribution. The distribution remained slightly positively skewed, which is practically inevitable when working with latencies.



(a) raw lexical decision latencies *(b) log-transformed lexical decision latencies*

Figure 4.26: Histograms for LDLs (Tynesider data)

The boxplots in Figure 4.27 provide an overview of the log-transformed latencies, broken down by condition and accent. As can be seen, latencies were overall longer in the noise condition, especially for the NZE trials. In the quiet condition, there seems to be little variation between the two accents, granted that the logarithmic transformation resulted in a compression of the data. The interquartile ranges of the boxplots demonstrates that variance was generally higher when the stimuli were masked with noise.

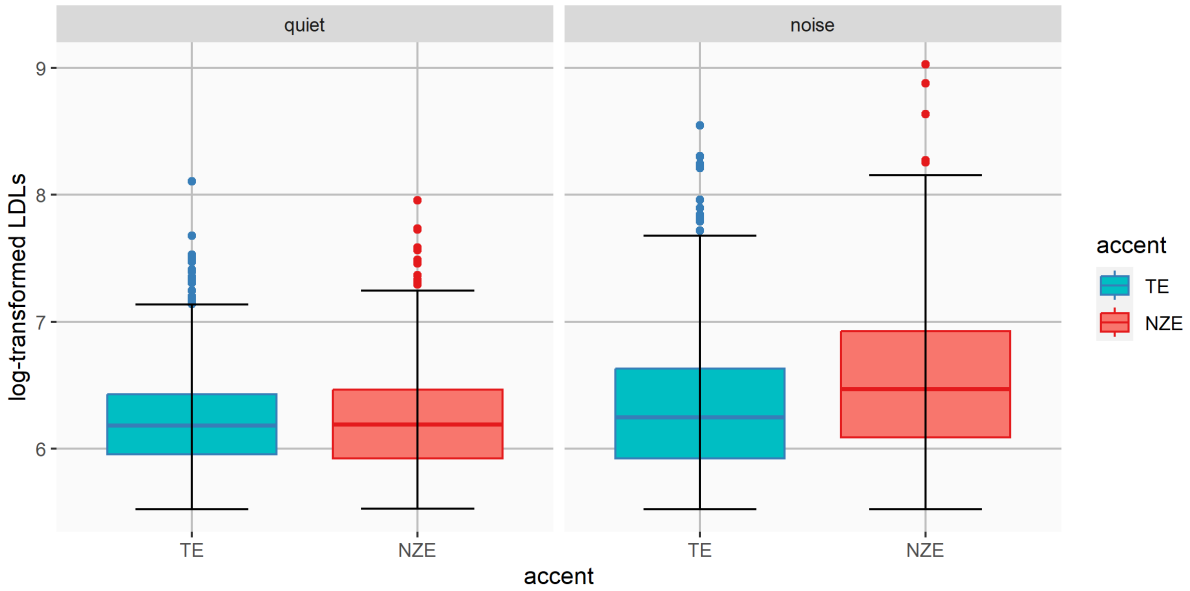


Figure 4.27: Log-transformed LDLs as mediated by accent and condition (colour: accent; Tynesider data)

The LDL data were fed into the below model, the output of which is shown in Table 4.15.

glmer(LDL_log ~ accent x condition x block + speech_rate_z + freq_log10_z + (1 + accent|participant) + (1|item))

Predictor	Estimate	Standard Error
(Intercept)	6.23	0.05
Accent (NZE)	0.00	0.05
Condition (noise)	0.14	0.06
Block (second half)	-0.02	0.02
Accent (NZE) : Condition (noise)	0.14	0.04
Accent (NZE) : Block (second half)	-0.03	0.03
Condition (noise) : Block (second half)	-0.09	0.03
Accent (NZE) : Condition (noise) : Block (second half)	0.12	0.05
Speech Rate (z-scored)	0.00	0.01
Word Frequency (log10, z-scored)	-0.00	0.02

Table 4.15: Model output: LDLs (Tynesider data)

Three predictors emerged as significant from model comparisons: first, accent, $\chi^2(1) = 5.71$, $p = .017$; second, condition, $\chi^2(1) = 11.66$, $p < .001$; and third, block,

$\chi^2(1) = 15.47$, $p < .001$. Additionally, the interaction between accent and condition, $\chi^2(1) = 27.46$, $p < .001$; and the interaction between accent, condition and block, $\chi^2(1) = 6.03$, $p = .014$, were significant. The other model components were not significant: speech rate, $\chi^2(1) = 0.07$, $p = .785$; word frequency, $\chi^2(1) = 0.05$, $p = .815$; the interaction between accent and block, $\chi^2(1) = 1.76$, $p = .185$; and the interaction between condition and block, $\chi^2(1) = 1.27$, $p = .260$.

The main effect of accent showed longer LDLs for the NZE trials, $\beta = 0.09$, $SE = 0.04$ (see footnote 13). As can be seen in Figure 4.28, participants' reactions were 57 ms slower when they heard the unfamiliar rather than familiar accent.

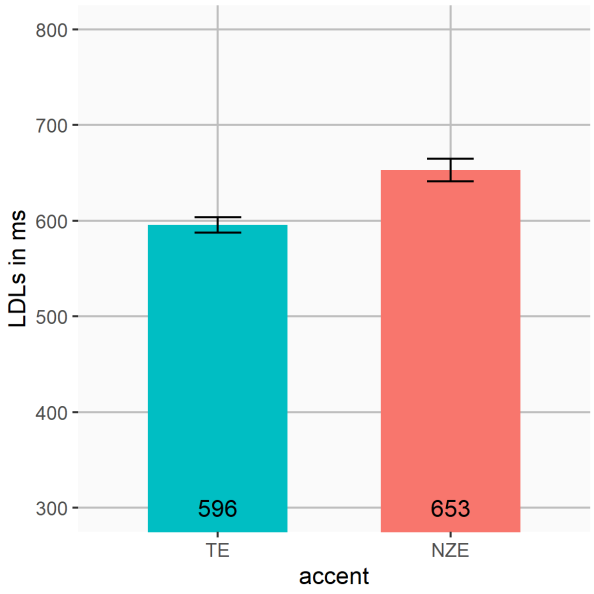


Figure 4.28: LDLs as mediated by accent
(bars: mean, error bars: ±one SE, colour: accent; Tynesider data)

Mean LDLs per condition are shown in Figure 4.29. As can be seen, the participants' lexical decisions took, on average, 546 ms in the quiet condition and 723 ms in the noise condition, $\beta = 0.19$, $SE = 0.04$ (see footnote 13). There was a difference of 177 ms between the two conditions, which surpassed the main effect of accent above.

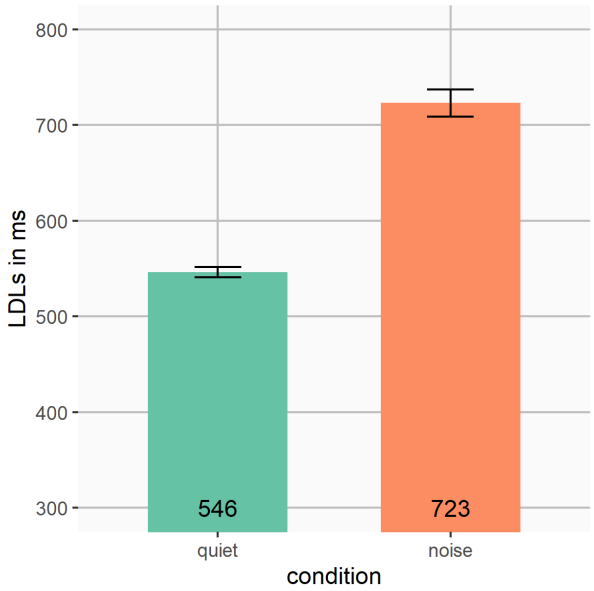


Figure 4.29: Latencies as mediated by condition
(bars: mean, error bars: ±one SE, colour: condition; Tynesider data)

With regards to block, the latencies decreased when moving from the first to the second half of trials produced by a specific speaker, β = -0.05, SE = 0.04 (see footnote 13). Figure 4.30 shows these decreasing latencies. On average, reactions during the second half of trials were 29 ms faster, which suggests an overall adaptation effect for the lexical decision task.

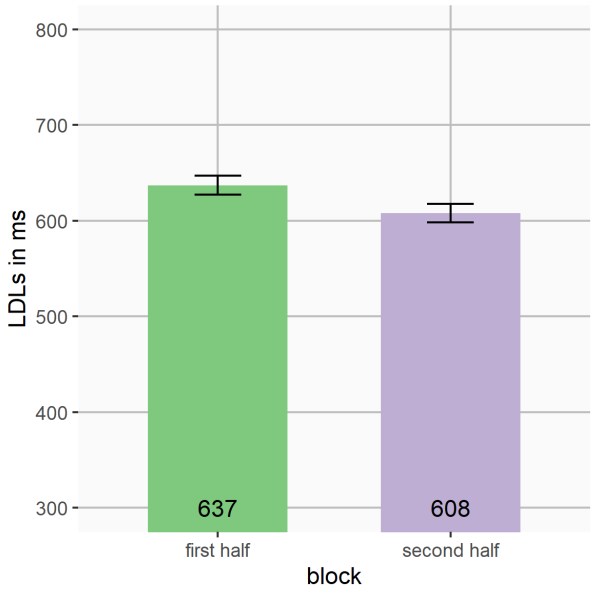


Figure 4.30: LDLs as mediated by block
(bars: mean, error bars: ±one SE, colour: block; Tynesider data)

The interaction between accent and condition is already shown in Figure 4.27 above. For ease of interpretation and to show the differences more clearly, LDLs are given in milliseconds

in Figure 4.31. Pairwise comparisons showed significant differences between TE and NZE trials in noise, z = -4.49, $p < .001$ (longer LDLs for NZE speakers); but not in quiet, z = 0.23, p = .820. This demonstrates that the main effect of accent (see Figure 4.28) was due to the results in the noise condition. In fact, average latencies in the quiet condition were equal across both accents. Under adverse listening conditions, however, Tyneside participants' lexical decisions were faster for the TE trials. The average difference in latencies between the TE and the NZE speakers in this condition was 176 ms.

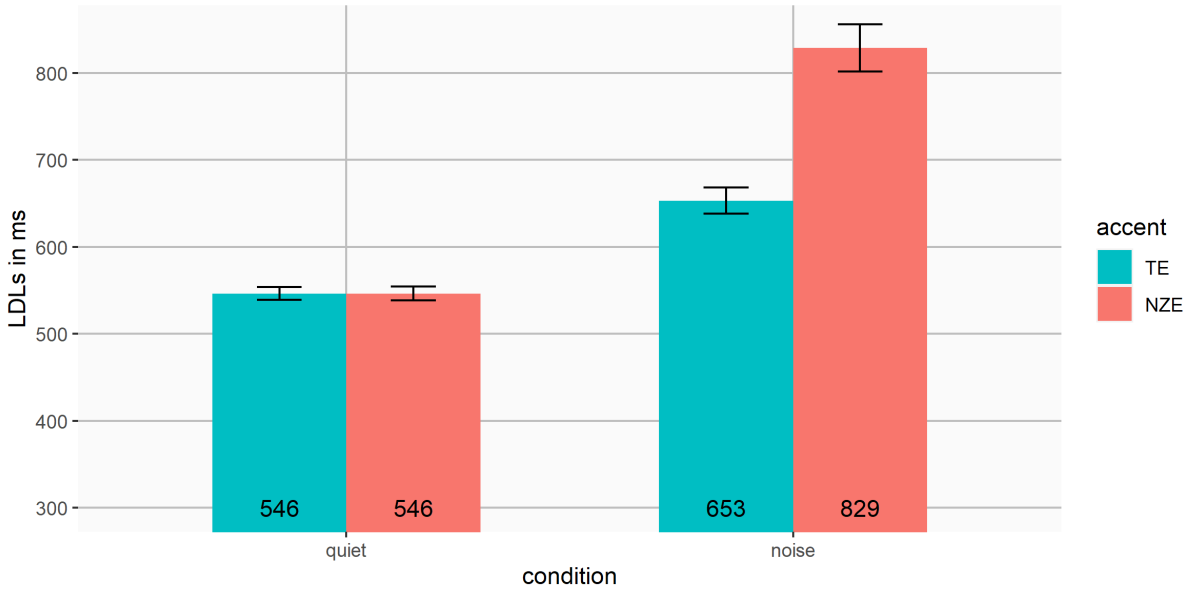


Figure 4.31: LDLs as mediated by accent and condition
(bars: mean, error bars: ±one SE, colour: accent; Tynesider data)

To interpret the significant three-way interaction, Figure 4.32 depicts average latencies, as mediated by accent, block and condition. All main effects from above are apparent in the data in that, overall, latencies were shorter for TE trials, in the quiet condition and during the second half of trials. Pairwise comparisons between TE and NZE were significant for the noise trials (longer LDLs for NZE speakers in the first half, z = -3.08, p = .002; and longer LDLs for NZE speakers in the second half, z = -5.12, $p < .001$) but not for the quiet trials (first half, z = -0.11, p = .916; and second half, z = 0.54, p = .588). Thus, the three-way interaction has to do with TE-NZE difference in noise, which was larger during the second half of trials. This is because participants' lexical decisions became faster for TE trials in noise (decrease from 690 ms to 615 ms) but slower for NZE trials in noise (increase from 802 ms to 853 ms). Taken together with the accuracy results, this suggests less successful adaptation of lexical processing to the unfamiliar accent in noise.

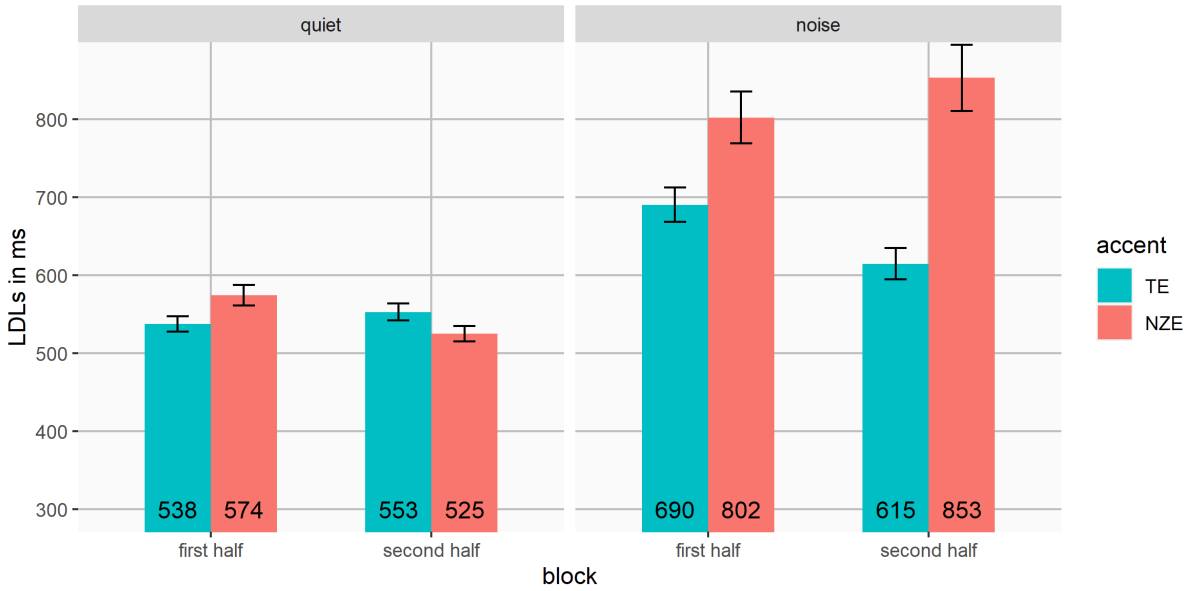


Figure 4.32: LDLs as mediated by accent, condition and block (bars: mean, error bars: ±one SE, colour: accent; Tynesider data)

4.8.2.2 Recall Task

4.8.2.2.1 Key Press Accuracy

The dataset for the accuracy of the key press included all experimental trials completed by participants from Tyneside in Experiments 1 and 2. Out of the total 2040 key presses, 1761 (86.3%) were correct and 279 (13.7%) were incorrect. Mean accuracy rates across the accents, conditions and story types are provided in Figure 4.33. For the TE trials, the adverse listening conditions did not appear to have a strong effect on the participants' accuracy. For the NZE speakers, however, changes between the stories were not detected as reliably in noise. Accuracy was best for unchanged stories in most cases.

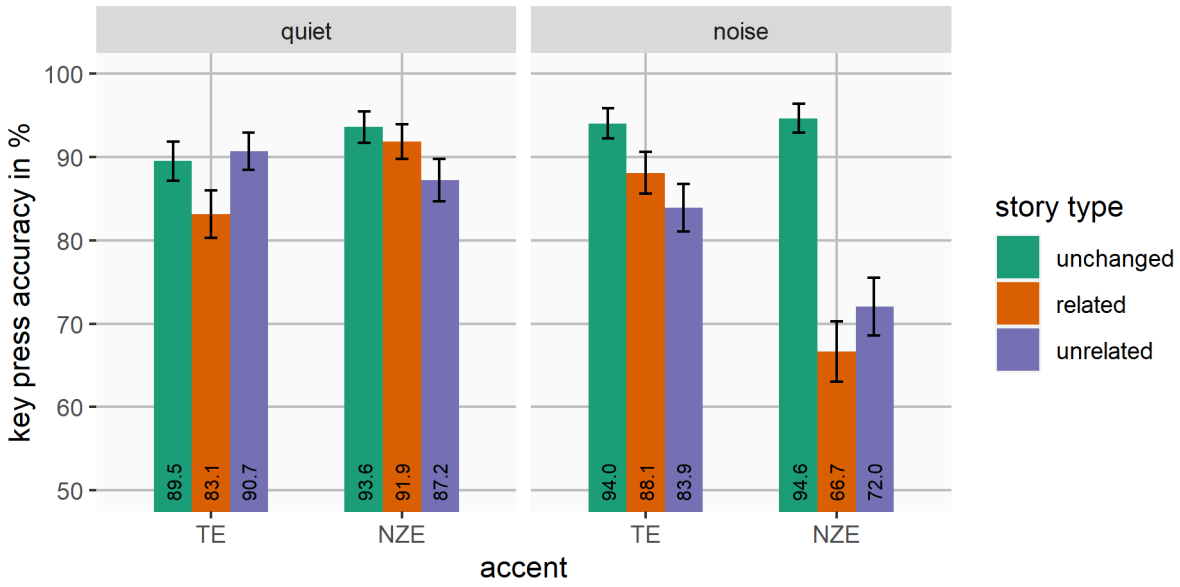


Figure 4.33: Key press accuracy as mediated by accent, story type and condition (bars: mean, error bars: ±one SE, colour: story type; Tynesider data)

The below model was used to analyse the key press accuracy data. The output of the model is provided in Table 4.16.

glmer(accuracy_key_press ~ accent x story_type x condition + block + speech_rate_z + (1|participant) + (1|item))

Predictor	Estimate	Standard Error
(Intercept)	2.64	0.35
Accent (NZE)	0.35	0.50
Story Type (related)	-0.60	0.33
Story Type (unrelated)	0.16	0.37
Condition (noise)	0.59	0.46
Accent (NZE) : Story Type (related)	0.37	0.54
Accent (NZE) : Story Type (unrelated)	-0.87	0.54
Accent (NZE) : Condition (noise)	-0.44	0.62
Story Type (related) : Condition (noise)	-0.24	0.53
Story Type (unrelated) : Condition (noise)	-1.35	0.54
Accent (NZE) : Story Type (related) : Condition (noise)	-1.91	0.78
Accent (NZE) : Story Type (unrelated) : Condition (noise)	-0.02	0.78
Block (second half)	-0.05	0.14
Speech Rate (z-scored)	0.13	0.14

Table 4.16: Model output: key press accuracy (Tynesider data)

The model comparisons showed main effects of story type, $\chi^2(2) = 33.57$, $p < .001$; and condition, $\chi^2(1) = 3.91$, $p < .048$. Furthermore, there were three significant interactions: the interaction between accent and condition, $\chi^2(1) = 12.37$, $p < .001$; the interaction between story type and condition, $\chi^2(2) = 12.83$, $p = .002$; and the interaction between accent, story type and condition, $\chi^2(2) = 9.71$, $p = .008$. The remaining model components did not emerge as significant: accent, $\chi^2(1) = 1.19$, $p = .276$; block, $\chi^2(1) = 0.10$, $p = .753$; speech rate, $\chi^2(1) = 0.93$, $p = .335$; and the interaction between accent and story type, $\chi^2(1) = 4.90$, $p = .086$.

Key press accuracy per story type is displayed in Figure 4.34. The results of pairwise comparisons between story types were significant differences between unchanged and related stories (higher key press accuracy for unchanged stories), $z = 5.19$, $p < .001$; as well as unchanged and unrelated stories (higher key press accuracy for unchanged stories), $z = 4.91$, $p < .001$; but not between related and unrelated stories, $z = -0.36$, $p > .999$. On average, 82.5 % and 83.5 % of the participants made the correct decision for related and unrelated stories, respectively, compared to an accuracy rate of 92.9% if the stories were identical.

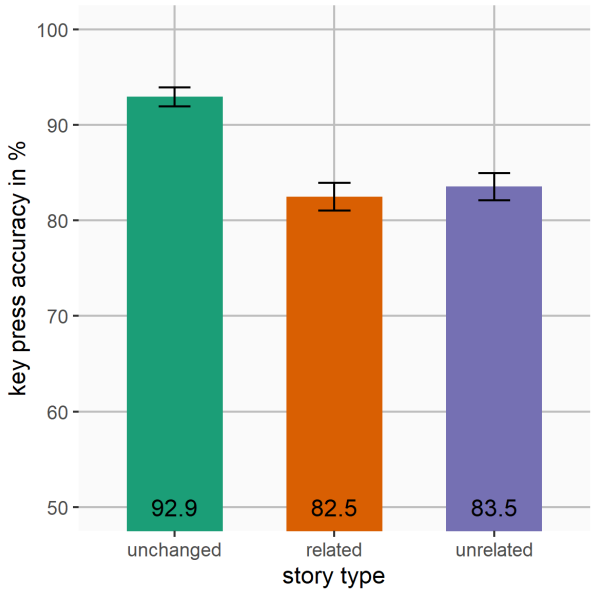


Figure 4.34: Key press accuracy as mediated by story type (bars: mean, error bars: ±one SE, colour: story type; Tynesider data)

In terms of condition, accuracy was lower under adverse listening conditions, $\beta = 0.01$, $SE = 0.43$. Figure 4.35 shows that the average accuracy in noise was 6.1 percentage points lower than in quiet. Importantly, the condition predictor was also part of three interactions, as addressed below.

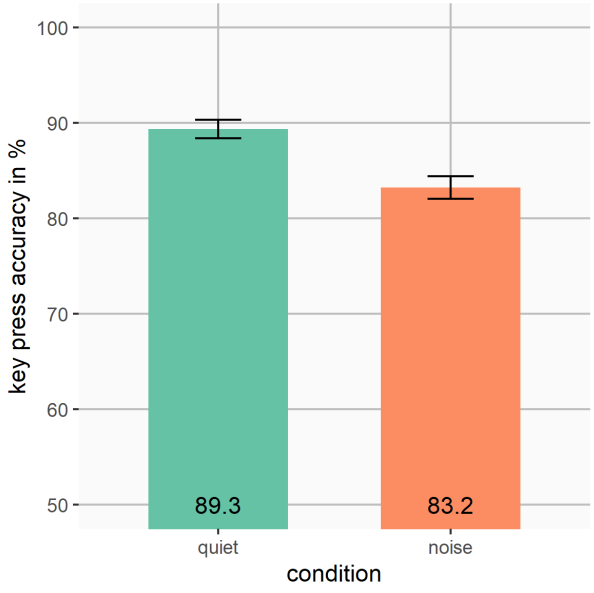


Figure 4.35: Key press accuracy as mediated by condition (bars: mean, error bars: ±one SE, colour: condition; Tynesider data)

Figure 4.36 visualises the interaction between accent and condition. Pairwise comparisons between TE and NZE showed significant differences for trials in noise (higher key press accuracy

for TE speakers), z = 2.54, p = .011; but not in quiet, z = -0.51, p = .611. In noise, the participants considered here were more accurate when they heard a speaker of TE (88.7%) than NZE (77.8%), which suggests a familiarity benefit for this task under adverse listening conditions.

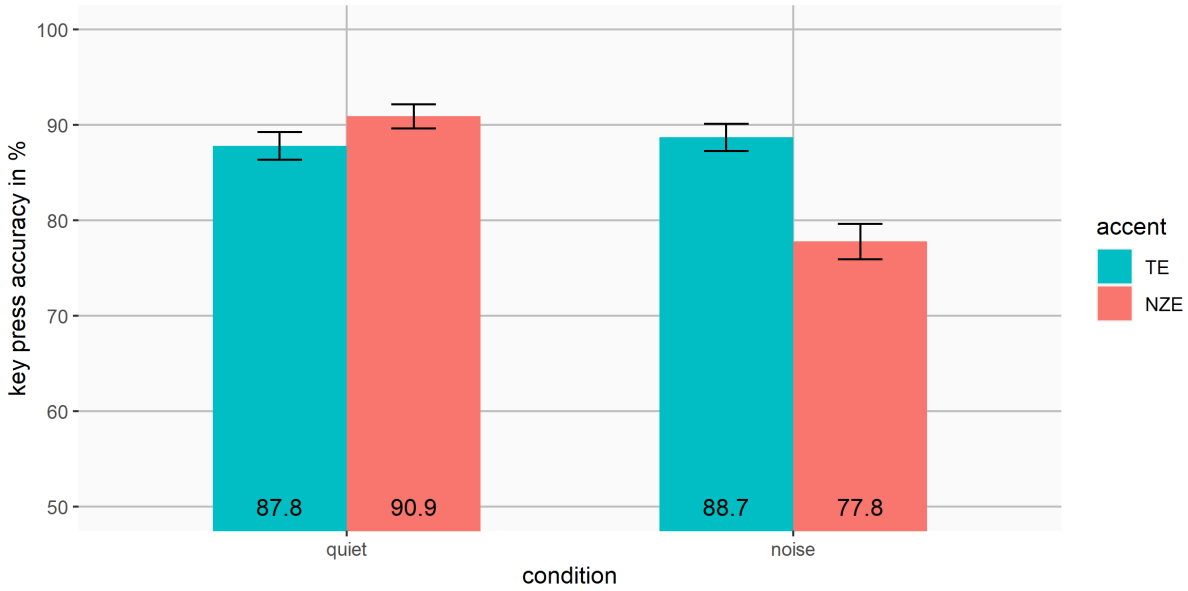


Figure 4.36: Key press accuracy as mediated by accent and condition (bars: mean, error bars: ±one SE, colour: accent; Tynesider data)

The interaction between story type and condition is shown in Figure 4.37. Pairwise comparisons between the two conditions by story type showed significant differences between trials in noise versus trials in quiet for related stories (higher key press accuracy for quiet trials), z = 2.73, p = .006; and unrelated stories (higher key press accuracy for quiet trials), z = 3.34, p = .001; but not for unchanged stories, z = -1.01 p = .314. Under adverse listening conditions, there was a clear drop in accuracy for the related and unrelated stories. This decrease in accuracy also caused the main effect of story type reported on above. The highest accuracy rate in Figure 4.37 came from unchanged stories in noise, which participants responded to correctly in 94.3% of the cases.

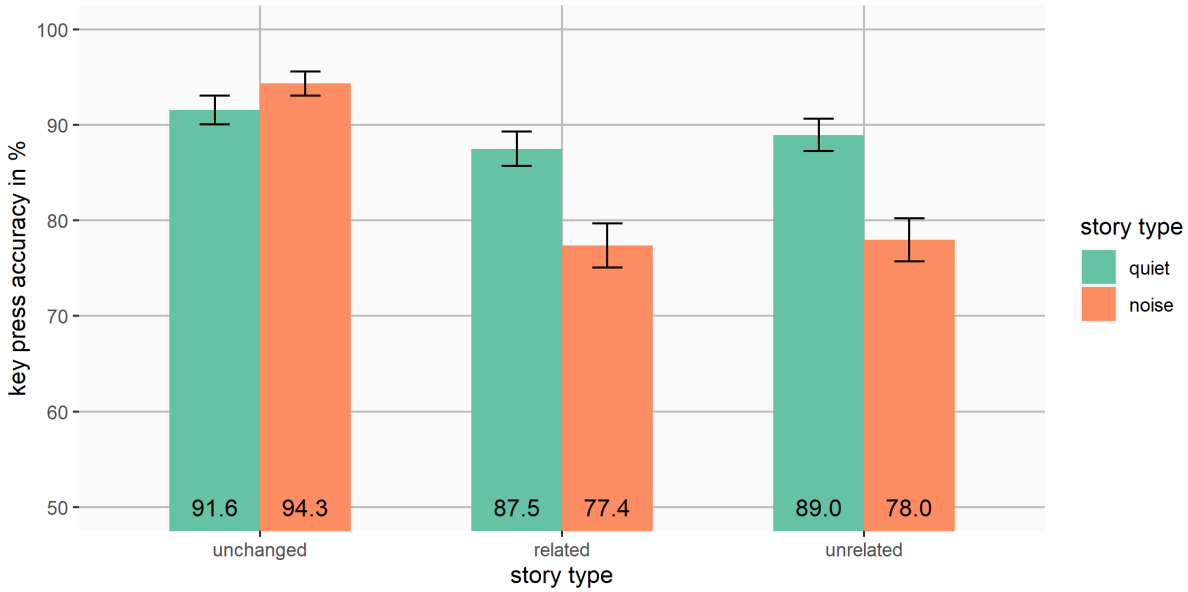


Figure 4.37: Key press accuracy as mediated by story type and condition (bars: mean, error bars: ±one SE, colour: condition; Tynesider data)

Finally, the three-way interaction between accent, story type and condition is displayed in Figure 4.33 above. Pairwise comparisons were conducted between story types for the four combinations of accent and condition. The following comparisons reached significance (see footnote 14):

- TE trials presented in noise: unchanged versus unrelated (higher key press accuracy for unchanged stories), $z = 3.02$, $p = .008$.
- NZE trials presented in noise: unchanged versus related (higher key press accuracy for unchanged stories), $z = 6.10$, $p < .001$; and unchanged versus unrelated (higher key press accuracy for unchanged stories), $z = 5.31$, $p < .001$.

Thus, significantly lower accuracy rates resulted from a combination of NZE trials, (un)-related stories and adverse listening conditions. For example, the lowest accuracy rate of 66.7% occurred when the participants had to decide if two related stories read by a NZE speaker in noise were identical or different. For unrelated NZE stories in noise, the rate was at 72.0%. The accuracy for the three story types also varied for the TE speakers in noise but was overall higher than for the NZE speakers, as evidenced by the interaction between accent and condition addressed above.

4.8.2.2.2 Correction Accuracy

Since unchanged stories and stories for which the participants made an incorrect key press were not relevant for the correction accuracy, the dataset included 1129 observations. There were 963 (85.3%) accurate and 166 (14.7%) inaccurate corrections. The bar plots in Figure 4.38 demonstrate that inaccurate corrections were more concentrated among the trials presented in noise. In quiet, there was a drop in correction accuracy for unrelated stories in TE. This corresponded to the lowest correction accuracy of 73.8% found for unrelated stories presented in TE in noise.

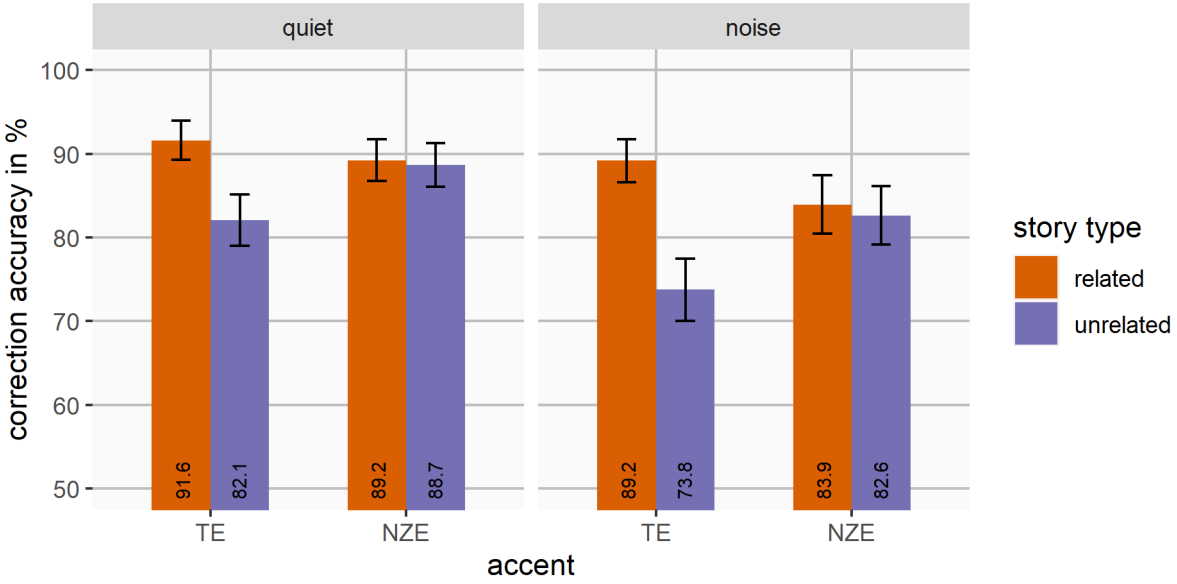


Figure 4.38: Correction accuracy as mediated by accent, story type and condition (bars: mean, error bars: ±one SE, colour: story type; Tynesider data)

The correction accuracy data were fed into the following model, the output of which is shown in Table 4.17.

$$\text{glmer(accuracy\_correction} \sim \text{accent x story\_type x condition} + \text{block} + \text{speech\_rate\_z} + (1|\text{participant}) + (1|\text{item}) \text{)}$$

Predictor	Estimate	Standard Error
(Intercept)	2.87	0.46
Accent (NZE)	-0.06	0.60
Story Type (unrelated)	-1.08	0.40
Condition (noise)	-0.51	0.48
Accent (NZE) : Story Type (unrelated)	0.96	0.55
Accent (NZE) : Condition (noise)	-0.00	0.58
Story Type (unrelated) : Condition (noise)	0.20	0.54
Accent (NZE) : Story Type (unrelated) : Condition (noise)	-0.27	0.76
Block (second half)	-0.22	0.19
Speech Rate (z-scored)	-0.24	0.20

Table 4.17: Model output: correction accuracy (Tynesider data)

The model comparisons showed a significant effect of story type, $\chi^2(1) = 8.65$, $p = .003$; and of the interaction between accent and story type, $\chi^2(1) = 4.55$, $p = .033$. The other predictors and interactions did not reach significance in the model comparisons: accent, $\chi^2(1) = 0.54$, $p = .463$; condition, $\chi^2(1) = 3.07$, $p = .080$; block; $\chi^2(1) = 1.32$, $p = .251$; speech rate, $\chi^2(1) = 1.28$, $p = .258$; the interaction between accent and condition, $\chi^2(1) = 0.13$, $p = .722$; the interaction between story type and condition, $\chi^2(1) = 0.03$, $p = .874$; and the interaction between accent, story type and condition, $\chi^2(1) = 0.12$, $p = .729$.

With regards to story type, correction accuracy was lower for unrelated stories, β = -0.86, SE = 0.49 (see footnote 13). The difference in accuracy between stories with semantically related versus unrelated changes is shown in Figure 4.39. There was an accuracy difference of 6.9 percentage points between the two story types.

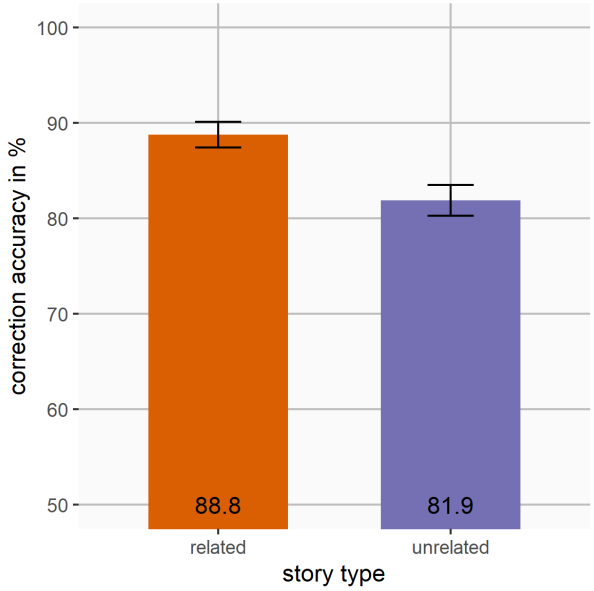


Figure 4.39: Correction accuracy as mediated by story type (bars: mean, error bars: ±one SE, colour: story type; Tynesider data)

Importantly, there was also an interaction between accent and story type, which is visualised in Figure 4.40. Participants from Tyneside corrected changes between stories in their own accent more reliably when these changes were semantically related. For unrelated stories, correction rates were higher for the unfamiliar accent NZE. However, these differences did not reach significance in pairwise comparisons: TE versus NZE for related stories, z = 0.12, p = .906; and TE versus NZE for unrelated stories, z = -1.61, p = .108.

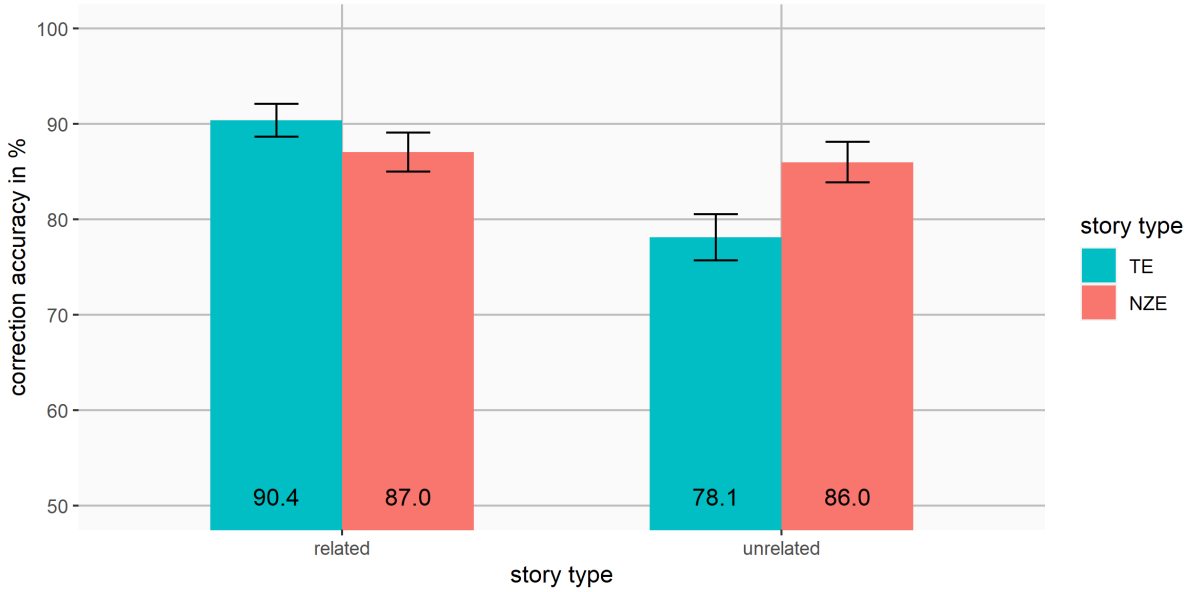


Figure 4.40: Correction accuracy as mediated by accent and story type (bars: mean, error bars: ±one SE, colour: accent; Tynesider data)

4.9 Summary of Findings

Experiment 1 (see section 3.7) found an effect of word frequency on the accuracy of lexical decisions in that participants' accuracy was higher for more frequent words (see McNamara, 2005; Perea & Rosa, 2002). This effect was not replicated for the noise data. However, there was a main effect of accent, with participants generally performing better for the TE speakers under adverse listening situations. This effect was qualified by an interaction with participant origin. The difference in accuracy between TE and NZE trials was larger for the Tyneside than the New Zealand participants. Evidence for a processing cost for the unfamiliar accent under adverse listening conditions came from a comparison of the Tynesider data in Experiments 1 and 2. The Tyneside participants' accuracy decreased much more for the unfamiliar NZE when noise was added to the stimuli than for the familiar TE (see Adank & Janse, 2010; Adank & McQueen, 2007; Stringer & Iverson, 2019). Interestingly, even the participants from New Zealand made more accurate decisions for the TE rather than the NZE trials, which links back to the potential confound of individual voice characteristics and noise-masking (see subsection 4.2.2). However, there was an evident effect of familiarity in that participants from New Zealand did better for the NZE trials in noise than participants from Tyneside.

Lexical decision latencies for the noise data were mediated by interactions between accent and participant origin as well as accent and block. Only for participants from Tyneside was there an evident difference in latencies. Their lexical decisions were, overall, faster for the TE than the NZE trials, suggesting that familiarity with an accent plays an important role under adverse listening conditions, especially because average latencies were identical for TE and NZE trials in quiet. In addition, there was evidence for adaptation in noise only for the TE speakers. For these speakers, latencies decreased from the first to the second half of trials. In quiet, adaptation by Tyneside participants was observed for the NZE speakers, although the difference between the first and second half of trials did not reach significance. Adaptation in terms of latencies was accompanied by a slight decrease in accuracy, suggesting a speed accuracy tradeoff. Returning to the discussion of latencies, the data suggest that there are both upper and lower limits to adaptation. If listeners are highly familiar with an accent, there might be little adaptation in quiet because the lower end of latencies has already been reached. Once noise is added to the signal, the task becomes more challenging, which allows for adaptation even for a highly familiar accent. The unfamiliar accent is more difficult to process than the familiar one even in quiet. This makes room for adaptation effects, accompanied by a speed

accuracy tradeoff. However, under adverse listening conditions the unfamiliar accent might become too challenging for adaptation to take place.

A familiarity benefit in noise also emerged for the recall data. Participants' initial decision via key press as to whether the two stories they heard were identical or different was mediated by an interaction between accent, story type and participant origin. Participants from Tyneside made more accurate decisions when they heard their own accent. This was mediated by story type in that the familiarity benefit was only apparent in story pairs that included a change. For identical stories, key press accuracy was practically identical across the two accents. A comparison of the Tynesider data in quiet versus in noise further substantiated the familiarity benefit. Tyneside participants generally performed worse in noise, especially for the unfamiliar NZE. Again, this effect was apparent in stories that did include a change rather than unchanged stories. These findings were telling in that the familiarity benefit did not emerge in Experiment 1, when TE was presented to Tyneside participants in quiet. Apparently, the added difficulty was necessary to elicit the effect (see Hällgren et al., 2001; Kjellberg et al., 2008; Ljung et al., 2009; Marsh et al., 2015).

One effect that was replicated from Experiment 1 is the main effect of story type on correction accuracy. Both in noise and in quiet, participants were better at correcting semantically related rather than semantically unrelated changes. This suggests that semantic proximity aids recall, potentially by means of a priming effect (see McNamara, 2005; Perea & Rosa, 2002). While there was an interaction between story type and accent, pairwise comparisons did not show any significant effects. This indicates that semantic proximity constitutes a general proxy for the recall of L1 accents.

In terms of the research questions from subsection 4.1.3, the main findings from Experiment 2 are the following:

(1) How do adverse listening conditions and accent familiarity affect lexical processing?
For participants from Tyneside, there was a clear accent familiarity benefit in terms of both accuracy and latencies. New Zealand participants also performed better for TE than for their own accent. However, their lexical decisions for NZE were more accurate than the decisions made by participants from Tyneside, although not to a significant extent.

(2) Is there evidence for adaptation in terms of lexical processing?
Adaptation occurred for Tyneside participants when they heard TE speakers in noise.

Additionally, their reactions became faster (although not to a significant degree) for NZE trials in quiet.

(3) What is the effect of word frequency on lexical processing?
There was no effect of word frequency on lexical processing in Experiment 2.

(4) How do adverse listening conditions, accent familiarity and semantic proximity of the change affect recall?
Under adverse listening conditions, Tyneside participants' key press accuracy was higher for their own accent if they heard stories that included a change.
In quiet and in noise, there was a general recall benefit for semantically related changes. The target in the first story might prime the one in the second story, which aids the subsequent recall of the change.

(5) Overall, is there evidence for less-detailed processing in quiet versus in noise and/or for the unfamiliar versus familiar accent?
There is some evidence for less-detailed processing under adverse listening conditions. Tyneside participants' key press accuracy was higher for their own accent, which might suggest that they processed the unfamiliar accent in less detail, resulting in poorer change detection. However, if less-detailed processing was the key mechanism behind this effect, performance in the recall task should be especially low for semantically related changes. These changes would be more difficult to detect and correct if encoding is less detailed. This pattern did not emerge from the results. In fact, as shown in (3), semantic proximity aided recall.

While the effects of intelligibility and familiarity on speech processing have received substantial attention in previous studies, research on the potential effects of attitude is very sparse. Attitude is the final factor of interest to be considered within this thesis and will be presented in the following chapter to shed light on how positive and negative attitudes towards a speaker/accent might affect speech processing.

Chapter 5

Experiment 3: Attitude

5.1 Introduction

Experiments 1 and 2 investigated how lexical access and recall are mediated by someone's familiarity with an accent and by the presentation of the accent in quiet versus in noise. Familiarity was defined as long-term familiarity with an accent and determined based on geographical criteria as well as self-reports. Intelligibility was manipulated by overlaying the stimuli with multi-speaker babble noise. The key insights from these experiments were the following: First, there was a clear familiarity benefit for lexical access under adverse listening conditions. Accuracy was generally higher and lexical decision latencies were generally shorter when Tyneside participants did the lexical decision task for their own accent in noise. Second, the findings for the Tyneside participants suggested upper and lower boundaries for adaptation to accents. In quiet, Tynesiders' latencies decreased during the second half as compared to the first half of New Zealand English (NZE) trials, albeit not to a significant extent. In noise, latencies decreased for the Tyneside English (TE) but not the NZE trials. Thus, there was some evidence for adaptation for the unfamiliar accent in quiet and the familiar accent in noise. Conversely, the familiar accent in quiet might have been too easy and the unfamiliar accent in noise might have been too difficult for any adaptation effects to emerge. Third, accent familiarity aided recall in that the key press accuracy was higher when the participants heard stories with changes in their own accent in noise. Finally, semantic proximity aided the recall of the changes between the two stories. The participants' correction accuracy was higher when the changes between the stories were semantically related rather than unrelated, potentially due to a priming effect.

Experiment 3 expanded on these results by examining the potential influence of attitude on lexical processing and recall. As explained in more detail in subsection 2.5.3.3, language attitudes are a complex construct that can be divided into implicit and explicit attitudes. Typical language attitude questionnaires focus on explicit attitudes and include items along the constructs of status, solidarity and, sometimes, dynamism (Montgomery, 2018). Previous research has shown that implicit and explicit attitudes need not converge (e.g. Pantos & Perkins, 2012). Furthermore, a myriad of factors were found to mitigate explicit attitudes, including education, tolerance of ambiguity, emotional stability and extraversion (e.g. Dewaele & McCloskey, 2015).

The following subsection will recap research on how attitudes affect linguistic processing specifically and cognitive processing more generally. Next, the research questions for Experiment 3 are listed, before an overview of the remainder of this chapter is provided.

5.1.1 Attitudes, Linguistic Processing and Cognitive Processing

The research in this subsection has been discussed in more detail in subsections 2.5.3.4 and 2.5.3.5. Here, the main findings are revisited to show the research gap in the literature and, thus, motivate the research questions in subsection 5.1.2.

Studies into the effects of implicit/explicit attitudes on linguistic processes such as lexical access are very scarce. Derwing et al. (2002) did not find an effect of more positive attitudes towards L2 English on the participants' comprehension and transcription thereof. Lev-Ari et al. (2019) found that lexical access was mitigated by an interaction between implicit attitudes and information provided on a speaker. Their participants were more likely to apply a Southern American rather than a General American phonology for the identification of an ambiguous lexical item if the speaker was presented as Southern and if they had a negative implicit bias towards the American South. This finding suggested that the integration of indexical information might be affected by implicit attitudes.

More generally, a link between attitude and linguistic or cognitive processing could be established via working memory and processing mode. It is plausible to assume that negative attitudes towards an accent are associated with negative affect. Negative affect, in turn, has been shown to reduce working memory capacity and to elicit item-specific rather than relational encoding. With regards to working memory, Figueira et al. (2017) found electroencephalographic evidence that participants' working memory capacity was compro-

mised after exposure to pictures that elicited negative affect. Reduced working memory capacity might negatively affect lexical processing and recall performance.

In terms of processing mode, Storbeck and Clore's (2005) results showed that participants employed item-specific rather than relational processing after listening to music that elicited negative affect. This was shown by the recall of fewer critical lures in a task based on the Deese-Roediger-McDermott paradigm. This paradigm involves the presentation of word lists with items that are all semantically related to a lure. The lure is not part of the word list but frequently named by participants during the subsequent free recall. The recall of fewer lures suggests more item-specific processing, which "involves encoding items by their features, elements, and distinctive qualities [...] [rather than] in relation to other concepts in memory" (Storbeck & Clore, 2005, 786). The results in Storbeck and Clore (2005) were replicated by Storbeck (2013), who again found fewer critical lures in the recall of participants after they listened to music or saw pictures that elicited negative affect. Storbeck's (2013) results also showed significantly better spatial recall in the negative affect condition, which further indicated a tilt of the processing mode towards item-specific processing. Correspondingly, long-term memory presentations were found to be more specific and granular if the event to be remembered was associated with negative emotions (Kensinger, 2009; Levine & Bluck, 2004).

In summary, negative attitudes towards an accent could affect lexical access and recall in that negative affect might compromise working memory capacity and induce item-specific processing.

5.1.2 Research Questions

Integrating language attitudes into experimental designs and statistical models comes with a methodological challenge, which could explain the relatively small number of studies in the area. If language attitudes are conceptualised as a categorical independent variable (positive versus negative), how can it be ensured that the participants actually hold such attitudes to the speakers and their accents? While previous research into language attitudes might help with the selection of an accent, it is not guaranteed that the participants' attitudes will follow these patterns. Additionally, introducing speakers of various accents into the design opens a large space for potential confound variables. As the noise masking procedure from Experiment 2 (see subsection 4.2.2) has shown, individual voice characteristics might lead to patterns in the data that can easily be misattributed.

To avoid these problems as much as possible, the current experiment took a novel approach. The main idea was to use the same speakers and to then manipulate the participants' attitudes deliberately. Specifically, the Tyneside speakers from Experiments 1 and 2 were included in the study and the attitude manipulation took place before the participants completed the lexical decision and recall task. Accent familiarity was not manipulated here to keep the online experiment at a reasonable length for the participants. As described in more detail below, this manipulation consisted of overt speaker information, a judgement task and an investment task. It was hypothesised that negative attitudes towards a speaker would result in a negative emotional state while listening to this speaker. This would, in turn, reduce working memory capacity and promote item-specific processing.

Based on these considerations, the results from Experiments 1 and 2 and previous research, the following research questions were addressed in Experiment 3:

(1) How do language attitudes affect lexical processing?

(2) Is there evidence for adaptation in terms of lexical processing?

(3) What is the effect of word frequency on lexical processing?

(4) How do language attitudes and the semantic proximity of the change affect recall?

(5) Overall, is there evidence for less-detailed processing when the speaker is evaluated positively versus negatively?

The pilot study for Experiment 3 is the concern of the following section. Next, the participants, stimuli, procedure and data analysis steps of the main experiment are addressed. After the presentation of results, key insights from this chapter are recapped.

5.2 Pilot Study

The online pilot study was conducted to evaluate how effective the attitude manipulations were that would be used in the main experiment. Altogether, there were three different manipulations, which were meant to elicit negative responses for one of the TE speakers and positive responses for the respective other speaker. Importantly, all three manipulations contained speech stimuli so that the participants' evaluation of the speakers would be as closely tied to the features of their speech as possible. The manipulations were targeted at the status

and the solidarity dimensions of language attitudes as they are the most commonly used ones in the literature.

5.2.1 Participants

The participants for the pilot study were recruited via Prolific and received payment for completing the study. The sampling criteria on the platform were set to include Prolific users who were born in the North East of England and resided there at the time of the study. Any user who had taken part in Experiments 1 or 2 or the corresponding pilot studies was not eligible. The sampling criteria for this pilot study and the main experiment on attitude were looser than for the experiments on familiarity and intelligibility. This was because familiarity was not a factor that was manipulated but rather one that had to be controlled for to the extent that it did not become a confounding variable in the analysis. The chief concern here, as stated before, was to evaluate the effectiveness of the attitude manipulation and, later on, how lexical processing and recall were affected by it.

In total, there were 21 complete submissions. The data from three participants was excluded because, based on their own reports in the demographic questionnaire, they did not fulfil the geographic sampling criterion. As a result, the dataset used in the analyses below consisted of the ratings from 18 participants. On average, they were 45.4 years old ($sd = 10.3$ years)[1] and, thus, slightly older than the participants from the previous experiments. When the study was conducted, there was an even split between male and female participants. The L1 of all participants was English and none of them reported speech, language, hearing or learning difficulties.

5.2.2 Stimuli

The aim of the attitude study was to manipulate the participants' attitudes to the two TE speakers, such that they would have a positive attitude to TE_1 and a negative attitude to TE_2 or vice versa. This is referred to in the following as 'speaker manipulation'. To achieve this, additional tasks were developed that required the recording of stimuli, namely a judgment task and an investment task. Further details on the tasks will be provided in subsection 5.2.3. Briefly, the judgement task included participants judging the speakers' responses to trivia

[1] These numbers are based on the responses from 17 participants as age data were not available for one participant.

questions. During the investment task, the participants interacted with the speakers, invested virtual money and received virtual money back from the speakers.

For the judgement task, ten trivia questions were developed, such as *What nut is used to make marzipan?* For each question, one right and one wrong answer was created. For the example question, the right answer was *You make marzipan with almonds*, while the wrong answer was *You make marzipan with peanuts.* The answers only differed in a single word or phrase. They were deliberately devised as full sentences rather than single word responses so that the participants would be exposed to more speech input from TE_1 and TE_2. Table 5.1 shows the ten questions and corresponding responses. Only the responses were recorded with the speakers as the questions would be shown visually during the pilot study and main experiment.

Question	**Correct Response**	**Incorrect Response**
What nut is used to make marzipan?	You make marzipan with almonds.	You make marzipan with peanuts.
What is the highest mountain in the world?	Mount Everest is the highest mountain in the world.	Ben Nevis is the highest mountain in the world.
What is the capital of Portugal?	The capital of Portugal is Lisbon.	The capital of Portugal is Madrid.
Which English king married six times?	Henry VIII married six times.	James III married six times.
Where do you find the Golden Gate Bridge?	You find the Golden Gate Bridge in San Francisco.	You find the Golden Gate Bridge in Japan.
What is the main fruit in Ribena?	Blackcurrant is the main fruit in Ribena.	Orange is the main fruit in Ribena.
What does 'GP' stand for?	GP stands for general practitioner.	GP stands for ghost place.
Brie and Camembert are types of which food?	Brie and Camembert are types of cheese.	Brie and Camembert are types of meat.
A pug is a breed of which domestic pet?	A pug is a breed of dog.	A pug is a breed of cat.
What is the largest country in the world?	The largest country in the world is Russia.	The largest country in the world is Spain.

Table 5.1: Questions and correct/incorrect responses for the judgement task

For the investment task, sentences were used that were supposed to encourage the participants to invest in their virtual partner (e.g. *My strategy is clear: always return part of your investment*). Some introductory phrases were also selected (e.g. *Hello, nice to meet you. Let's get started with the investment game*). In total, four introductory phrases and 36 encouragement phrases became part of the stimulus inventory. All sentences were taken from Torre (2017), who used the investment task to analyse how trust attribution is mediated by voice characteristics. In the current design, trust is not the main concern. However, the investment task was the ideal task to manipulate the participants' attitudes. It was interactive, included speech samples and allowed for a positive or negative treatment of the participants so as to influence their attitudes towards TE_1 and TE_2. A full list of the introductory phrases and the encouragement phrases can be found in Appendix B.5. Generally, the phrases were similar in length to the above examples and were meant to increase participants' investments.

Since the main experiment included the lexical decision and recall task with TE_1 and TE_2, it was imperative that these speakers would produce the stimuli for all the tasks described here as well. To that end, the two TE speakers (middle-aged, female) returned to the School of Education, Communication and Language Sciences (ECLS) at Newcastle University in March 2022. The recordings were conducted in a sound-attenuated booth in the new phonetics lab in ECLS with an Edirol R09HR digital recorder and a Sennheiser radio microphone. The recordings were made in the wav format at a sampling and digitisation rate of 44.1 kHz and 16 bit, respectively. As can be seen in Figure 5.1, the recording setup was slightly different from the first recording sessions in 2020 because no such booth was available then. Importantly, the quality of the recording was high in both setups and, thus, sufficient for the purposes of the pilot study and the main experiment. Similar to the recordings in 2020, the participants went through the recording session in a self-paced manner. The sentences to be recorded were presented via the PennController for IBEX (PCIbex) on a laptop in the recording booth. The researcher sat outside the booth and checked the accuracy of what the participants read. The recording session consisted of two blocks with an intermittent break. After the second block, sentences with inaccurate readings were rerecorded with the researcher present in the booth.

The recording sessions lasted no longer than 45 minutes. The two speakers were given a token of appreciation for their efforts.

Figure 5.1: Recording setup for TE speakers: attitude stimuli

The recordings were processed in Praat. The individual stimuli were segmented on a tier and then extracted as separate wav files with a script by Lennes (2002). Using a different script, their amplitude was then normalised to 65db and they were uploaded to LabVanced.

5.2.3 Procedure

After the participants were informed about the purpose of the pilot study and informed consent had been secured, they completed five tasks: a headphone check, a speaker information task, a judgement task, an investment task and a rating task. They also filled in a demographic and language background questionnaire. Further details on the five tasks and the questionnaire will be provided in the following.

5.2.3.1 Headphone Check

Since the pilot study was run online via LabVanced, a headphone check was implemented to ensure that the participants heard the stimuli properly and focused on the tasks. Within sequences of three tones, the participants had to identify the softest tone, which, due to the

phase difference in the tones (see Woods et al., 2017 and subsection 3.2.3.1), was only reliably possible when headphones rather than computer loudspeakers were used. In order to pass the headphone check, at least five out of six trials had to be completed correctly.

5.2.3.2 Speaker Information

TE_1 and TE_2 were included in Experiment 3. The participants were told that they would hear and interact with two different speakers throughout the online experiment. The first step to actively manipulate their attitudes was taken when they were given information about the two speakers, who were introduced to them as 'Speaker A' and 'Speaker B'.

Listeners' expectations within linguistic studies have been manipulated before by providing additional information on the speaker. In many studies, photographs of the alleged speakers were used. For example, Rubin (1992) found that American students rated the same speaker from Ohio as more foreign-accented when the voice sample was preceded by a photograph of a Chinese rather than a white woman. Vaughn (2019) used introductory text passages rather than photographs to lead participants to believe that the speaker was an L1 or L2 speaker of English. For the current study, the participants were provided with such text passages rather than photographs because the latter could induce further bias and because it would be rather difficult to select photographs that reliably result in a positive or negative reaction from the participants.

The text passages, which were aimed at both the status and the solidarity dimension of language attitudes, are provided in Table 5.2. They were positive for Speaker A and negative for Speaker B, or vice versa. In addition to reading the passages, the participants listened to a short sound file so that they could get an idea of the speaker's voice. This sound file included three short and unrelated sentences, which had originally been recorded for the lexical decision task.[2] The audio stimulus was included so that the participants could link this information immediately to a specific voice rather than creating an abstract image of Speaker A and Speaker B.

[2] In the main experiment, it was ensured that the participants did not hear these sentences later on during the lexical decision task. The three sentences used here were: *He keeps his stuff in the garage. For breakfast, children eat toast or cereal. The man in the corner is the captain.*

Positive	Speaker A/B was a well-behaved student. She did a degree in education and now works as a teacher. After work, she enjoys spending time with her family and doing volunteering work.
Negative	Since she was very young, Speaker A/B has been getting into trouble. She dropped out of school early and struggles to maintain a job. She sleeps in most days and then goes out with her friends at her parents' expense.

Table 5.2: Positive/Negative speaker information

5.2.3.3 Judgement Task

The intention of the judgement task was to influence the listeners' attitudes towards the status dimension of the speakers. More specifically, one speaker was presented as more knowledgeable than the other. During the judgement task, the participants first saw one of the questions from Table 5.1 on their screen. They then heard either the correct or the incorrect response from Speaker A/B and had to decide if the speaker's response to the question was accurate. The first five questions from Table 5.1 were used for Speaker A while responses from Speaker B were presented for the second half of questions. This resulted in a total of 10 trials. In the positive speaker manipulation, the speaker's responses to the five questions were all correct, regardless whether it was Speaker A or B that was presented positively. In the negative manipulation, her answers were incorrect.

5.2.3.4 Investment Task

While the judgement task was directed at the status dimension of language attitudes, the investment task was included to manipulate the solidarity dimension. The investment task or "investment game" was first developed by Berg, Dickhaut, and McCabe (1995). The task used here is largely adapted from Torre (2017). The participants were told that they would interact with Speaker A and Speaker B during this task. For each speaker, there were eight rounds and each round consisted of five steps:

(1) The participants received £10 of virtual money.

(2) The participants listened to a short recording of the speaker. In the first round, this was an introductory statement (e.g. *Welcome to the investment game. I hope we will enjoy playing it*). In the following rounds, the statements were supposed to positively

influence the participants' investment decisions in the next step (e.g. *I am expecting you to share because that's exactly what I am doing*). As mentioned in subsection 5.2.2, the statements used in this step were sourced from Torre (2017).[3]

(3) The participants decided how much of their virtual money they would like to invest into the speaker. They could choose any amount in full pounds from £0 to £10.

(4) The invested amount was tripled and 'transferred' to the speaker. For example, if a participant invested £7, they would be left with £10 - £7 = £3 at this stage. The speaker would now have a balance of £7 x 3 = £21.

(5) The speaker returned a predetermined percentage of their balance in this round to the participant. For example, if the percentage was set to 70% for this round, the speaker would return 0.7 x £21 ≈ £15. Thus, the participant would end this round with £3 + £15 = £18. The speaker's final balance for this round would be £21 - £15 = £6. The participants could track their cumulative balance over the rounds in the top left of the screen and their final balance at the end of the eight rounds per speaker was shown to them as well.

Importantly, all participants were exposed to two conditions during the investment task. The conditions differed with regards to the return behaviour of the two speakers. If Speaker A had been presented positively during the speaker information and judgement task, the participants completed the investment task with Speaker A in the generous condition. If the overall presentation of Speaker A was negative, the mean investment condition was applied. Speaker B behaved in the correspondingly opposite manner. Table 5.3 includes the return rates across the eight rounds in the two conditions. In the generous condition, it ranged from 65% to 80% while it never exceeded 40% in the mean condition. The investment task was the final component in the current attempt to manipulate the participants' attitudes towards the two speakers.

Round		1	2	3	4	5	6	7	8
Condition	*generous*	70%	80%	75%	65%	60%	80%	70%	75%
	mean	25%	40%	15%	20%	30%	10%	5%	0%

Table 5.3: Return rates per condition in the investment task

[3] See Appendix B.5 for a complete list of the statements used.

The preceding subsections deliberately referred to Speaker A and Speaker B rather than TE_1 and TE_2. During the pilot study (and the subsequent main experiment), the participants did each task first for Speaker A and then for Speaker B. The identity of Speaker A changed between groups to avoid order effects. Specifically, the participants were iteratively allocated to four groups that differed in two respects: first, in terms of whether TE_1 or TE_2 was presented first (i.e. took the spot of Speaker A) and; second, in terms of whether TE_1 or TE_2 was presented positively. Table 5.4 provides an overview of the four groups and their progression through the conditions of the speaker information, judgement task and investment task. The second row in Table 5.4 shows the number of participants per group. As can be seen, the distribution was fairly uniform, with fewer participants only in Group B.

Participant Group		A	B	C	D
Number of Participants		5	3	5	5
Task	*Speaker Information: Speaker A*	TE_1 negative	TE_1 positive	TE_2 negative	TE_2 positive
	Speaker Information: Speaker B	TE_2 positive	TE_2 negative	TE_1 positive	TE_1 negative
	Judgement Task: Speaker A	TE_1 incorrect	TE_1 correct	TE_2 incorrect	TE_2 correct
	Judgement Task: Speaker B	TE_2 correct	TE_2 incorrect	TE_1 correct	TE_1 incorrect
	Investment Task: Speaker A	TE_1 mean	TE_1 generous	TE_2 mean	TE_2 generous
	Investment Task: Speaker B	TE_2 generous	TE_2 mean	TE_1 generous	TE_1 mean

Table 5.4: Task order, content and number of participants for the four groups in the attitude pilot study

5.2.3.5 Rating Task

The previous three tasks were used to manipulate the participants' attitudes. The rating task was implemented to evaluate if this manipulation was successful. As is common in language attitude research, Likert scales were used to elicit the participants' evaluations of the two speakers. Specifically, the participants rated both speakers in terms of eleven attributes on seven-point scales (1 = *strongly disagree*, 4 = *neutral* and 7 = *strongly agree*). The attributes

were categorised into the three dimensions often considered in attitude research (Kristiansen, Garrett, and Coupland, 2005, 16; Montgomery, 2018, 131):

(1) Status: Speaker A/B is *professional / educated / intelligent / respected by their peers*.

(2) Solidarity: Speaker A/B is *approachable / friendly / sociable / trustworthy*.

(3) Dynamism: Speaker A/B is *energetic / enthusiastic / confident*.

The participants rated Speaker A first, followed by Speaker B (see Table 5.4 for information on the identity and portrayal of each speaker). To remind them who each speaker was and what she sounded like, they could listen again to the sound files from the speaker information component (see subsection 5.2.3.2) before rating.

5.2.3.6 Demographic Questionnaire

After the main tasks of the experiment, the participants provided some demographic information. This included their age, gender (on a voluntary basis), whether their first language was English and some information on their residence history. The questionnaire also included a catch trial to filter out participants who did not pay full attention to the study. In addition, the participants were asked for feedback on the tasks.

5.2.4 Data Analysis

Since the aim of the pilot study was to test whether the participants' attitudes towards the two speakers could be manipulated, the results from the rating task (see subsection 5.2.3.5) were of chief interest. Responses were also collected for the other tasks and, in fact, have been used in previous research as dependent variables. Torre (2017), for example, used the amount of virtual money invested by the participants during the investment task as a proxy for trust attributions to different speakers. Within the framework of this experiment, it was not the participants' behaviour during the investment task itself that mattered but their subsequent evaluation of the two speakers.

All data processing, visualisation and analysis took place in R. The participants' ratings were treated as linear scales and fed into linear mixed effects models. Technically, the data obtained from Likert scales is ordinal, which is why the use of parametric tests, such as linear

models, is contested. However, as research by Kizach (2014) and Norman (2010) showed, these models can be used on Likert scale data without significant disadvantages.

The following two fixed effects were included in the models:

- **Speaker manipulation** coded how the speaker was portrayed during the pilot study. It had two levels: 'positive' and 'negative'. The reference level was set to 'positive'. This predictor was included to test how successfully the participants' attitudes were manipulated.
- **Rating category** was also categorical and had three levels: 'dynamism' (reference level), 'solidarity' and 'status'. The purpose of this predictor was to evaluate potentially different rating behaviour across the three categories.
- There was no theoretical motivation that the speaker manipulation would affect any of three rating categories more than the respective other two. Therefore, no interaction between **speaker manipulation** and **rating category** was included.

Two random effects were added to the models:

- Random intercepts of **participants** were included to account for the repeated measures per participant. Given the small sample size (see subsection 5.2.1), random slopes for speaker manipulation by participant could not be included as they would have resulted in issues of singular fit.
- Random intercepts for **speaker** were included to mitigate differences in the evaluation of TE_1 versus TE_2 beyond the manipulation of the study.

Following the procedure of the experiments presented so far, significant predictors were identified by log-likelihood comparisons between nested models, starting with the maximum model. The comparisons were conducted automatically in R with the 'afex' package. This model is provided in the following subsection, when the results from the pilot study are presented.

5.2.5 Results

The participants' average ratings, as mediated by rating scale and speaker manipulation, are displayed in Figure 5.2. Within each scale, the manipulation was successful in that ratings

were around one rating unit higher when the speaker was presented positively. The highest mean rating was found for a positive speaker manipulation in the solidarity category (5.5) while the lowest rating was found for a negative speaker manipulation in the status category (3.8).

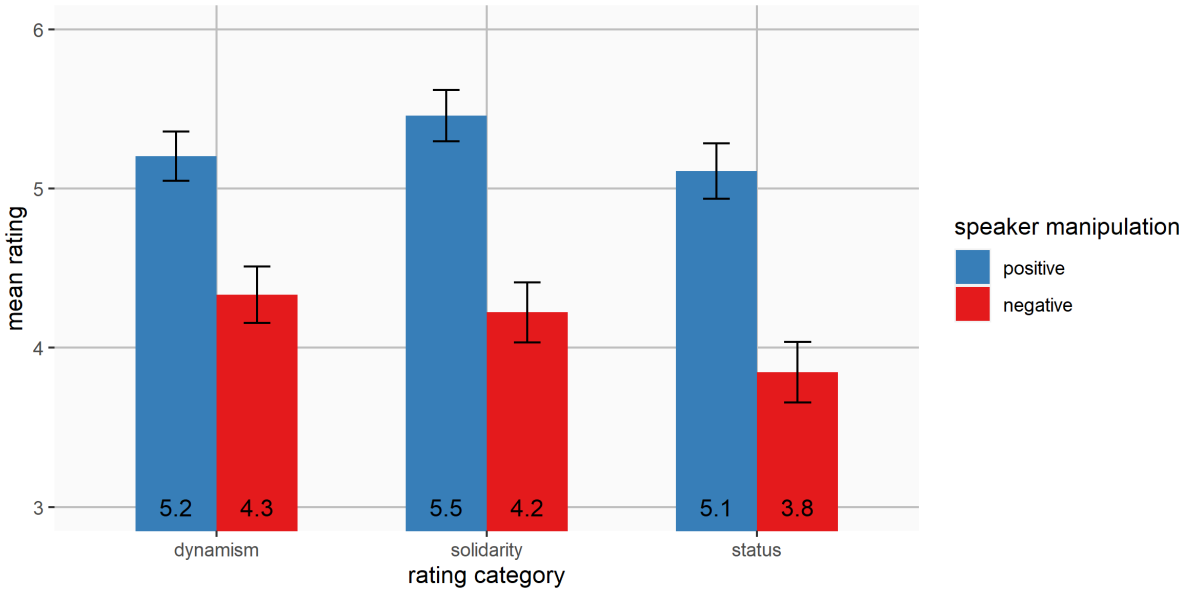


Figure 5.2: Ratings as mediated by rating scale and speaker manipulation (bars: mean, error bars: ±one SE, colour: speaker manipulation)

The ratings were fed into the model specified below. The model output is shown in Table 5.5.

$$\text{lmer(rating} \sim \text{speaker\_manipulation} + \text{rating\_category} + (1|\text{participant}) + (1|\text{speaker}) \text{)}$$

Predictor	Estimate	Standard Error
(Intercept)	5.38	0.41
Speaker Manipulation (negative)	-1.23	0.14
Rating Category (solidarity)	0.07	0.17
Rating Category (status)	-0.29	0.17

Table 5.5: Model output: ratings

The model comparisons showed a main effect of speaker manipulation on rating, $\chi^2(1) = 71.41$, $p < .001$. Rating category did not emerge as significant, $\chi^2(2) = 5.48$, $p = .065$.

With regards to speaker manipulation, the participants' ratings were significantly lower in the negative condition, β = -1.23, SE = 0.14. This has already been commented on above and is shown again in Figure 5.3, which plots mean ratings per speaker manipulation. The ratings were 1.2 units higher when the speaker was presented positively.

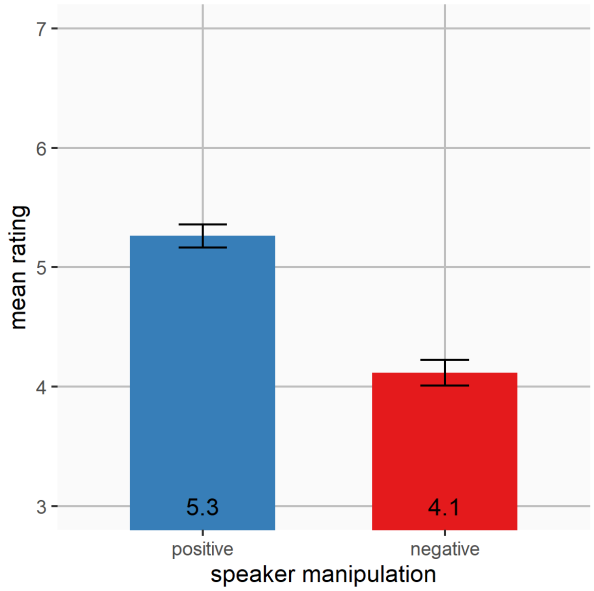


Figure 5.3: Ratings as mediated by speaker manipulation (bars: mean, error bars: ±one SE, colour: speaker manipulation)

5.2.6 Summary

The pilot study for the experiment on attitude was conducted to determine if participants' attitudes towards the two Tyneside speakers could be manipulated such that they were positively biased towards one and negatively biased towards the other speakers. The manipulation was achieved through information on the speakers, a judgement task and an investment task. The participants' attitudes after the manipulation were measured via Likert scales. The responses from 18 middle-aged participants from North East England were analysed. The results showed that the manipulation was successful in that ratings were generally higher when the speakers were presented positively. Based on this finding, it was concluded that the manipulation procedure could be used for the main experiment. The next sections include details on the participants, stimuli, procedure, data analysis steps and results thereof.

5.3 Participants

Two sources were used to recruit participants for the main experiment on attitude, which was conducted online via LabVanced: the SONA participant pool in ECLS at Newcastle University and Prolific. The implementation of a screening procedure by means of a survey was not as easy on these platforms. Hence, participants were identified as born and raised in the North East of England through their self-reports in the questionnaire at the end of the experiment (see section 5.5). For the approximately 45 minutes they spent on the study, the participants received SONA credits or payment.

The study was completed by 60 participants in total. Across both groups, 19 participants were excluded: One of them indicated that they had diagnosed speech, language, hearing or learning difficulties. A further 18 participants were not originally from the North East. Consequently, the results stem from the analysis of the responses of 41 participants. With regards to gender, 25 participants identified as female and 15 as male when the study was conducted. One participant preferred not to disclose this information. The mean age of the participants was 35.3 years (sd = 10.7 years), which was slightly higher than the average age of the participants in Experiments 1 and 2.

As explained above, familiarity was less of a concern in this experiment and only had to be controlled for so as not to become a confounding variable. Since the participants only heard their own accent or one they were assumed to be highly familiar with, the accent matching task from the previous experiments was not included here. Therefore, the main measure of familiarity was the participants' responses to the two questions: *How familiar are you with the accent spoken in Newcastle?* (familiarity) and *How many members of your social network (family and friends) are from Newcastle?* (social network). Table 5.6 provides the mean values and standard deviations of the responses. According to their self-reports, the participants were, on average, highly familiar with Newcastle/Tyneside English and had a strong social network in the region. However, for the social network measure, some participants actually indicated the lowest score on the scale. This was never the case for the familiarity scale, which always had ratings above 4 (out of 6). Since this experiment did not include the contrast between two accents, no linear regression models were run on the self-reports.

	Newcastle (English)
Familiarity with Accent	5.8 (0.4)
Social Network	4.6 (1.9)

Table 5.6: Self-reported familiarity and social network information: mean and standard deviation from seven-point Likert scales (0 to 6)

5.4 Stimuli

The stimuli for the attitude manipulation in the main experiment were identical to the ones from the pilot study (see subsection 5.2.2). Briefly, they consisted of some short sentences for the speaker information component, (in)correct responses to trivia questions for the judgement tasks and introductions as well as investment encouragements for the investment task. For the lexical decision task that followed, the TE stimulus set from Experiment 1 was used. It consisted of 112 short sentences (56 per speaker) taken from Stringer and Iverson (2020). Two versions of each sentence were created: one directly taken from Stringer and Iverson (2020) with a real word in sentence-final position and one with a nonword from Rastle et al. (2002) in this position. The stimuli for the recall task were triplets of stories that were each three sentences long and differed from each other in a single word only. The changed words between stories were either semantically related (*doctor* → *nurse*) or semantically unrelated (*doctor* → *teacher*). In total, 20 story triplets (12 experimental and 8 fillers) were used. Two versions of a story were played to the participants during one trial of the recall task. Detailed information on the development of the stimuli for the lexical decision and recall task can be found in subsection 3.2.2.

5.5 Procedure

The procedure of the attitude experiment is a combination of the tasks from the pilot study above and the tasks from Experiments 1 and 2. Only TE_1 and TE_2 were included as speakers in this study because including NZE and, thus, a manipulation of familiarity would have produced, first, a too complex design and, second, an experiment lasting more than one hour. In its final form, the online experiment consisted of the tasks below. Table 5.7 shows the Latin-square design that was used to mitigate order effects. The participants were allocated to one of the four groups iteratively via LabVanced. While not as equal as possible, the number

of participants across groups was very similar and there were at least nine participants per group.[4]

(1) **Headphone check**: To ensure that the participants were wearing headphones during the experiment, they completed a headphone check and had to identify the softest from a sequence of three sounds at least five out of six times correctly.

(2) **Speaker information**: The participants were informed that they would hear two speakers throughout the experiment, who were introduced to them as Speaker A and Speaker B. They heard a speech sample from each speaker and were provided with an introductory passage about each speaker, which was either positive or negative (see subsection 5.2.3.2).

(3) **Judgement task**: The participants heard the two speakers' responses to five trivia questions each. In total, there were ten trials. The participants had to decide if the responses were right or wrong. For one speaker, all answers were correct. The other speaker always provided incorrect responses (see subsection 5.2.3.3).

(4) **Investment task**: The participants completed the investment task with both speakers. The return behaviour of one speaker was generous while the other speaker behaved in a mean way during the task. There were eight trials per speaker for a total of 16 trials (see subsection 5.2.3.4).

(5) **Lexical decision task**: For this experiment, only TE_1 and TE_2 were included. The participants heard sentences from these speakers and had to decide via key press if the target (i.e. the final item of the sentence) was a real word or a nonword. After ten practice sentences, the participants completed 28 trials per speaker, 14 containing words and another 14 containing nonwords. A maximum of three (non)word trials was presented in a row and the trials were blocked by speaker. The recording of which speaker was used for a given sentence in the stimulus set was determined pseudorandomly in R.

(6) **Recall task**: The participants heard two stories in a row. The stories were three sentences long and either identical or different from each other in a single word. The participants had to, first, decide via key press if the stories were the same or different and, second, when they selected 'different', note down the change between the two

[4] The uneven number is due to some participants not completing the study or due to their data being excluded from analysis (see section 5.3).

stories. The changed words were either semantically related or unrelated. After four practice trials, the participants completed 10 trials per speaker for a total of 20 trials. The trials were blocked by speaker. Out of the 10 trials per speaker, four served as fillers and six as experimental trials (two unchanged, two related changes and two unrelated changes). The recording of which speaker was used for a given trial was decided in a pseudorandom manner in R.

(7) **Demographic and language background questionnaire**: The participants were asked to provide some demographic and residency information. Familiarity and social network measures were also elicited.

Participant Group		A	B	C	D
Number of Participants		10	9	12	10
Task	*Speaker Information: Speaker A*	TE_1 negative	TE_1 positive	TE_2 negative	TE_2 positive
	Speaker Information: Speaker B	TE_2 positive	TE_2 negative	TE_1 positive	TE_1 negative
	Judgement Task: Speaker A	TE_1 incorrect	TE_1 correct	TE_2 incorrect	TE_2 correct
	Judgement Task: Speaker B	TE_2 correct	TE_2 incorrect	TE_1 correct	TE_1 incorrect
	Investment Task: Speaker A	TE_1 mean	TE_1 generous	TE_2 mean	TE_2 generous
	Investment Task: Speaker B	TE_2 generous	TE_2 mean	TE_1 generous	TE_1 mean
	Lexical Decision Task	TE_1 first	TE_1 first	TE_2 first	TE_2 first
	Recall Task	TE_1 first	TE_1 first	TE_2 first	TE_2 first

Table 5.7: Task order, content and number of participants for the four groups in the main experiment on attitude

5.6 Data Analysis

5.6.1 Lexical Decision Task

Only word trials were included in the analysis of the lexical decision task. Both lexical decision accuracy and lexical decision latencies (LDLs) were analysed. Latencies were also used to detect outliers in the dataset. When the latencies were below 200 ms[5] or above 2.5 standard deviations from the participant-specific mean, they were excluded from the analysis (see Clopper et al., 2016, 92).

To analyse the data, (generalised) linear mixed effects models were applied to it in R with the 'lme4' package. The predictors included in each model are shown below and when the results are presented later on. Significance of predictors was determined by means of log-likelihood model comparisons, starting with the full model, with the 'afex' package. Since there was a theoretical motivation behind including each predictor in the model, none of them was dropped even if it did not reach significance. Pairwise comparisons were conducted with the 'emmeans' package. Bonferroni corrections were applied when multiple pairwise comparisons were necessary.

Two general remarks are relevant for the attitude data. First, there was no need for a fixed effect of **accent** in the models any more because all participants only heard the familiar TE during the experiment. Second, the new fixed effect **speaker manipulation** was included in the models to code if the speaker was presented positively or negatively during the first part of the experiment.

For the lexical decision data, the following model predictors were used:

- **Speaker manipulation** was a categorical predictor and had two levels: 'positive' and 'negative'. It referred to how the speaker was presented to the participants during the first part of the experiment. The reference level of speaker manipulation was set to 'positive'.

[5] In Experiments 1 and 2, the lower cut-off value for latencies was 250 ms rather than 200 ms. For the attitude dataset, a cut-off value of 250 ms would have resulted in an exclusion of 335 observations, which constituted almost 30% of the dataset. On average, the participants in Experiment 3 made their lexical decisions more quickly (mean latency of 527 ms before exclusion of outliers) than the ones in Experiment 1 (mean latency of 627 ms) and Experiment 2 (mean latency of 855 ms; Tyneside participants). Since the participants were generally faster in Experiment 3, the lower cut-off value was set to 200 ms instead. This reflected the increased speed and prevented an exclusion of a considerable portion of the dataset as outliers.

- **Block** coded if a specific trial was part of the 'first half' or 'second half' presented in a speaker's voice. Its purpose was to track adaptation effect, that is to say changes in accuracy and/or LDLs while listening to one speaker. Its reference level was 'first half'.
- **Word frequency** was a continuous predictor and added to check if more frequent targets were associated with better performance (see e.g. Grainger, 1990 and Whaley, 1978). It was of secondary interest. The word frequency measures were obtained from the SUBTLEX database (Brysbaert & New, 2009). Word frequency (in occurrences per million words) was first log10-transformed and then standardised. The logarithmic transformation is common for word frequency (see Van Heuven et al., 2014) and the standardisation allowed for more meaningful model coefficients (Winter, 2019, 89).
- **Speech rate** was a continuous predictor and operationalised in syllables per second. It was included as a control predictor to account for differences between the two speakers. Including 'speaker' as part of the random effects structure of the models would have resulted in singularity issues. Hence, speech rate was included as a fixed effect.

The models further included two random effects:

- Random intercepts for **participants** were added to account for the multiple measurements taken per participant. If possible, random slopes for speaker manipulation by participant were also added.
- Random intercepts for trial targets or **items** were added following the same rational. Random slopes were not included here.

5.6.2 Recall Task

The analysis of the recall data was based on two measures. 'Key press accuracy' coded if the participants' decision as to whether or not there was a change between the two stories was correct. The second measure was 'correction accuracy'. It was only considered for trials which did include a change between stories and for which the participants pressed the correct key initially. A correction was coded as accurate if the participants provided both the changed word in the second story and the original word from the first story. The participants' input was checked manually, for example to correct typos, and then analysed in R.

Significant effects were identified with the procedure laid out in subsection 5.6.1. The model components for the recall task included:

- **Speaker Manipulation**: categorical with the levels 'positive' (reference level) and negative.
- **Story type** coded if the story pair in a trial was 'unchanged' (reference level), 'related' or 'unrelated'. This predictor was used to test the effect of semantic proximity between the changes on recall. Differences here could indicate less-detailed processing or good-enough representations.
- An interaction between **speaker manipulation** and **story type** was included as a somewhat exploratory term. It was supposed to show if less-detailed processing was more prevalent in one of the manipulations.
- **Block**: categorical with the levels 'first half' (reference level) and 'second half'.
- **Speech rate**: continuous, measured in syllables per second and standardised.

Two random effects were added

- random intercepts for **participants** and, if possible, random slopes for speaker manipulation by participant
- random intercepts for **items**

5.7 Results

5.7.1 Lexical Decision Task

5.7.1.1 Accuracy

The results from two participants were excluded for the analysis of the accuracy data because their presence in the dataset resulted in model convergence and singularity issues.[6] From the remaining data, 176 observations were excluded as outliers, which resulted in a total number of 916 lexical decisions for the accuracy analysis. The average accuracy per speaker manipulation is shown in Figure 5.4. As can be seen, accuracy was overall very high, which did not come as a surprise as the participants heard a highly familiar accent in quiet during

[6] Quite many of the responses from these participants would have been excluded as outliers based on the analysis procedure described in subsection 5.6.1. For one participant, the outlier responses were concentrated in the second block. This might have been the reason why including them in the dataset resulted in the problems with the models.

the experiment. The mean accuracy was 1.6 percentage points lower in the negative speaker manipulation.

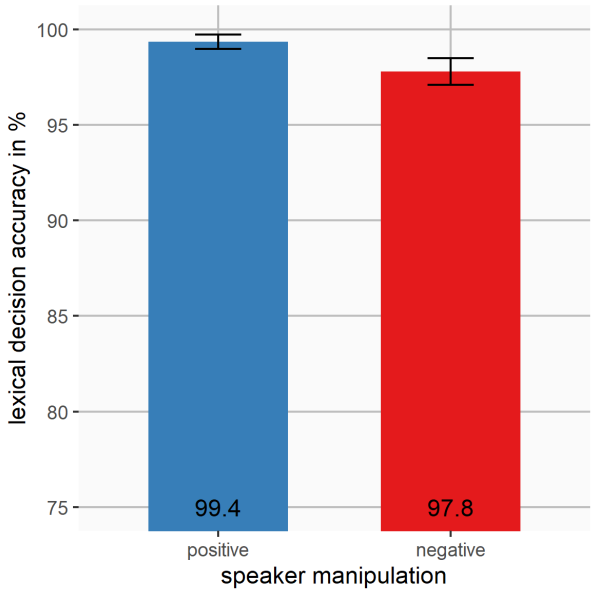


Figure 5.4: Lexical decision accuracy as mediated by speaker manipulation (bars: mean, error bars: ±one SE, colour: speaker manipulation)

The accuracy data were fed into the following model, the output of which is provided in Table 5.8.

glmer(accuracy_lex_dec ~ speaker_manipulation + block + speech_rate_z + freq_log10_z + (1|participant) + (1|item))

Predictor	Estimate	Standard Error
(Intercept)	5.24	0.93
Speaker Manipulation (negative)	-1.34	0.67
Block (second half)	0.93	0.62
Speech Rate (z-scored)	0.46	0.38
Word Frequency (log10, z-scored)	0.34	0.32

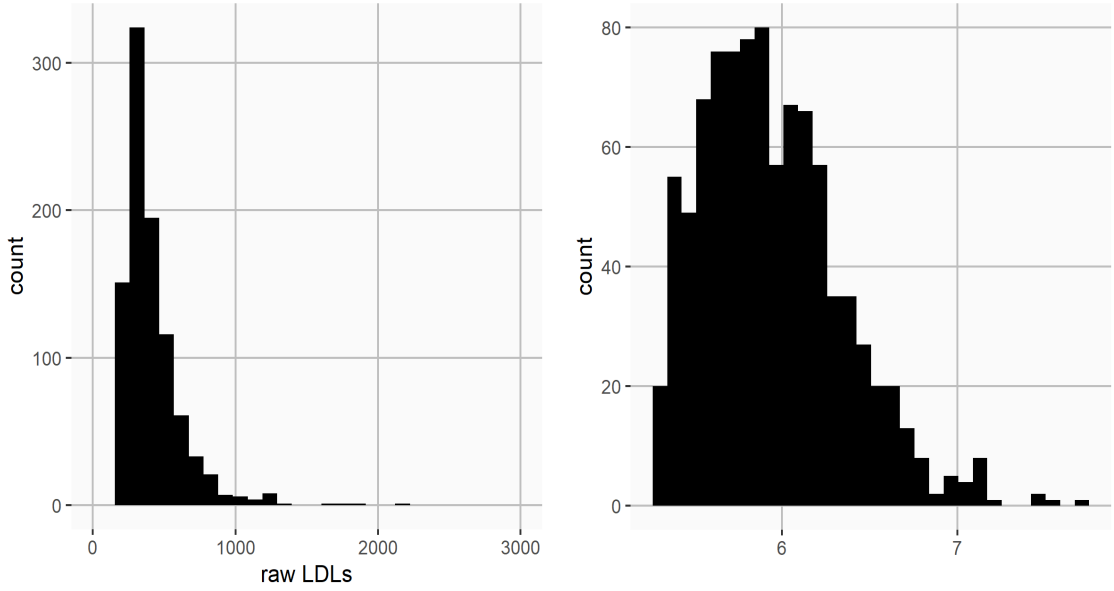
Table 5.8: Model output: lexical decision accuracy

The model comparisons showed a significant effect of speaker manipulation, $\chi^2(1) = 4.73$, $p = .030$. The other predictors in the model did not reach significance: block, $\chi^2(1) = 2.43$, $p = .119$; word frequency, $\chi^2(1) = 1.00$, $p = .317$; and speech rate, $\chi^2(1) = 1.61$, $p = .204$.

The main effect of speaker manipulation is visualised above in Figure 5.4. Lexical decisions were overall less accurate when the speaker was presented negatively, β = -1.34, SE = 0.67.

5.7.1.2 Lexical Decision Latencies

Latencies were considered for word trials with correct lexical decisions that had not previously been excluded as outliers. There was a total of 931 observations in the dataset.[7] In order to better fulfil the normality assumption for the residuals of the linear mixed effects models, the latencies were transformed logarithmically. The distribution of the latencies before and after this transformation can be found in Figure 5.5. The logarithmically condensed latencies were much closer in shape to a normal distribution than the distribution of the raw latencies.



(a) raw latencies *(b) log-transformed latencies*

Figure 5.5: Histograms for LDLs

Figure 5.6 shows how the LDLs varied between the two speaker manipulations in the experiment. Granted that the logarithmic transformation condensed the data, there appeared to be little variation between the speakers and the manipulations. The median value for the log-transformed latencies was slightly higher when the TE speakers were presented positively rather than negatively.

[7] The models ran without any problems with this dataset. Thus, the data from all 41 could be analysed (versus 39 participants for accuracy data; see footnote 6).

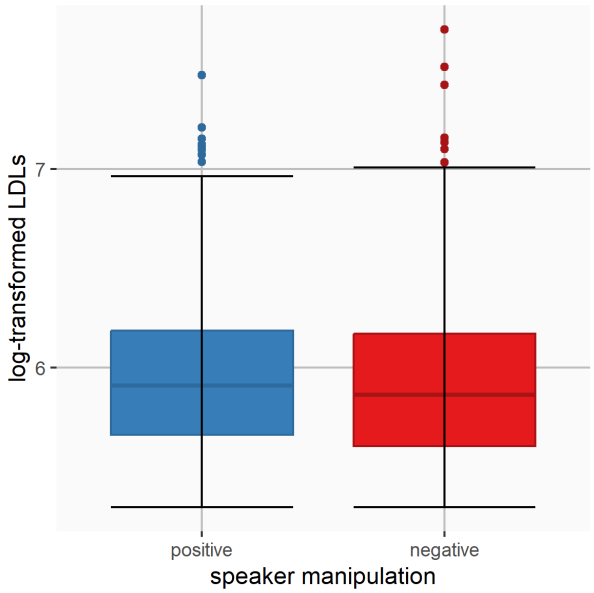


Figure 5.6: Log-transformed LDLs as mediated by speaker manipulation (colour: speaker manipulation)

The below model was used for the LDL data. Table 5.9 shows the model output.

$$\text{lmer(LDL\_log} \sim \text{speaker\_manipulation} + \text{block} + \text{speech\_rate\_z} + \text{freq\_log10\_z} + (1 + \text{speaker\_manipulation}|\text{participant}) + (1|\text{item}) \text{)}$$

Predictor	Estimate	Standard Error
(Intercept)	5.94	0.04
Speaker Manipulation (negative)	-0.03	0.03
Block (second half)	-0.01	0.02
Speech Rate (z-scored)	0.02	0.02
Word Frequency (log10, z-scored)	-0.05	0.02

Table 5.9: Model output: LDLs

The only model predictor to reach significance was word frequency, $\chi^2(1) = 3.86$, $p = .049$. The other predictors tested by means of model comparisons remained unsignificant: speaker manipulation, $\chi^2(1) = 1.36$, $p = .243$; block, $\chi^2(1) = 0.28$, $p = .596$; and speech rate, $\chi^2(1) = 1.17$, $p = .279$.

The effect of word frequency is visualised in Figure 5.7. As can be seen, LDLs were shorter for more frequent targets, $\beta = -0.05$, $SE = 0.02$. Thus, participants made their lexical decisions more quickly for more frequent targets. The difference between LDLs around the average word

frequency ($z = 0$) and around the highest word frequency ($z = 1.50$) was 405 ms - 381 ms = 24 ms.

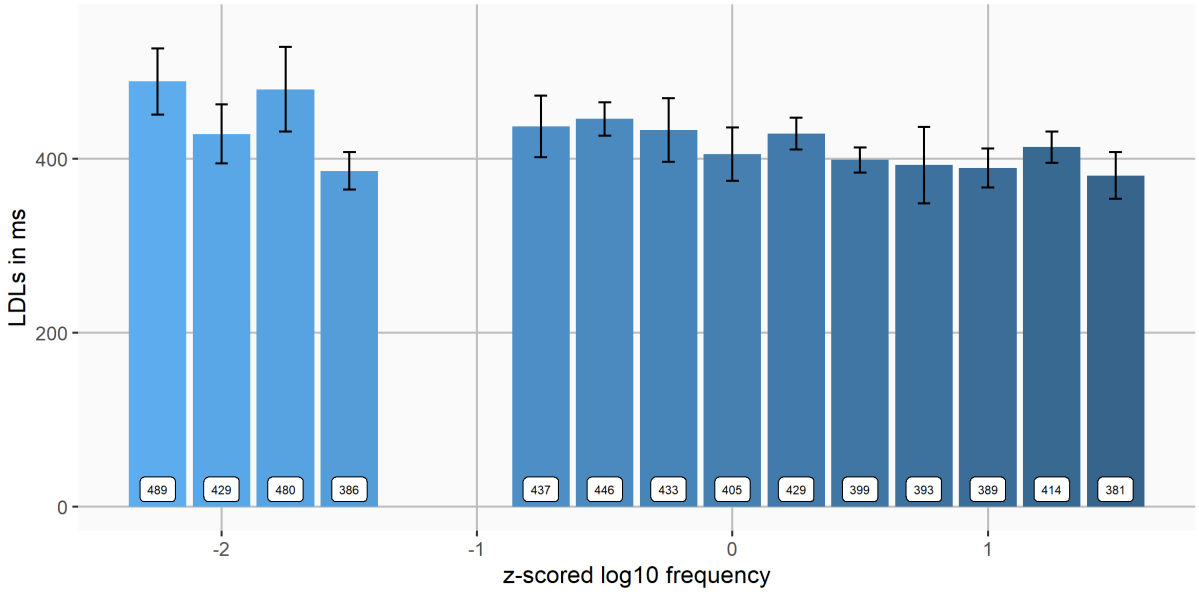


Figure 5.7: LDLs as mediated by word frequency (bars: mean, error bars: ±one SE)

5.7.2 Recall Task

5.7.2.1 Key Press Accuracy

492 responses were considered for the participants' accuracy of the key press in the attitude experiment. 433 of these responses were correct (88.0%) and 59 responses were incorrect (12.0%). Figure 5.8 demonstrates how average accuracy rates varied between combinations of story types and speaker manipulations. Overall, accuracy was very high. It was generally lower in the positive speaker manipulation. Additionally, in this manipulation, the accuracy rates were lowest for the related story type. This observation was somewhat surprising because semantic proximity was beneficial for recall in Experiments 1 and 2. In the negative condition, accuracy rates were lowest for the detection of semantically unrelated changes between the two stories. Across both speaker manipulations, stories without changes were associated with the highest accuracy rates.

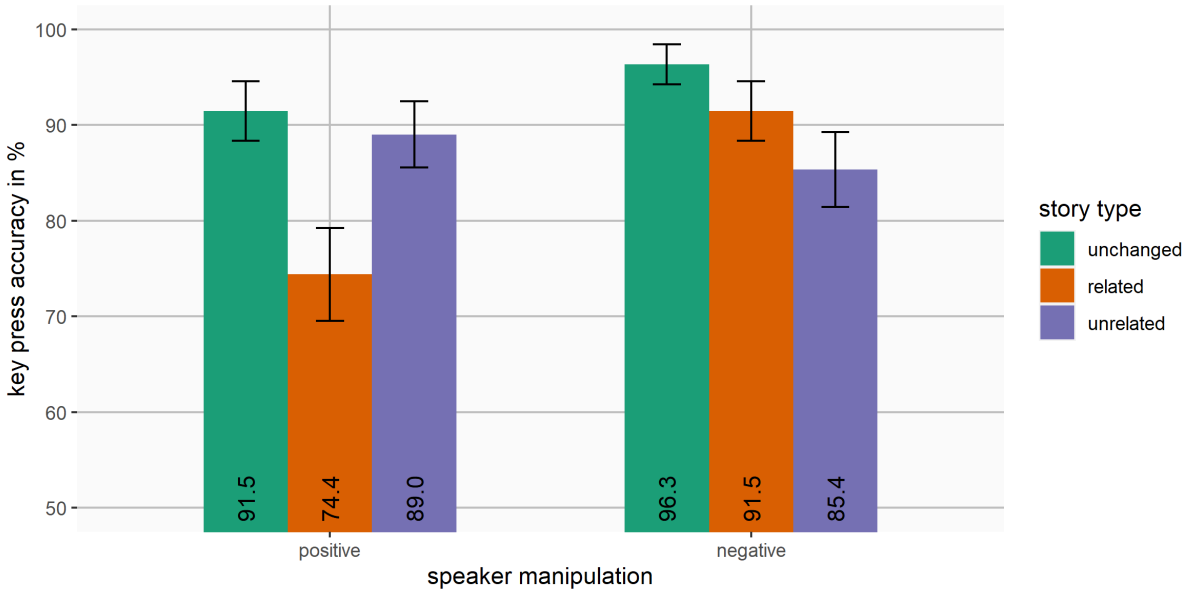


Figure 5.8: Key press accuracy as mediated by speaker manipulation and story type (bars: mean, error bars: ±one SE, colour: story type)

The following model was used to analyse the key press accuracy data. The output of the model is shown in Table 5.10.

glmer(accuracy_key_press ~ speaker_manipulation x story_type + block + speech_rate_z + (1|participant))[8]

Predictor	Estimate	Standard Error
(Intercept)	3.36	0.55
Speaker Manipulation (negative)	0.98	0.74
Story Type (related)	-1.50	0.52
Story Type (unrelated)	-0.29	0.57
Speaker Manipulation (negative) : Trial Type (related)	0.70	0.90
Speaker Manipulation (negative) : Trial Type (unrelated)	-1.42	0.90
Block (second half)	-0.96	0.33
Speech Rate (z-scored)	-0.11	0.16

Table 5.10: Model output: key press accuracy

[8] As opposed to the model specifications in subsection 5.6.2, the model for key press accuracy does not include random intercepts for items. Including these random intercepts resulted in issues of singular fit for some of the reduced models, which made model comparisons challenging. After careful consideration, the random intercepts for item were removed because model outputs showed that they accounted for very little variation in the data. For example, the variance of these intercepts in the full model would have been smaller than 10^{-6}. The variance for the random intercepts for participants, on the other hand, was 1.15. Hence, intercepts for participants were kept and those for items were dropped.

Three main effects emerged as significant from the model comparisons: speaker manipulation, $\chi^2(1) = 4.79$, $p = .029$; story type, $\chi^2(2) = 8.13$, $p = .017$; and block, $\chi^2(1) = 8.91$, $p = .003$. Additionally, there was a significant interaction between speaker manipulation and story type, $\chi^2(2) = 8.81$, $p = .012$. Speech rate did not reach significance, $\chi^2(1) = 0.53$, $p = .467$.

The main effect of speaker manipulation is shown in Figure 5.9. Key press accuracy was higher when the speaker was presented negatively, $\beta = 0.74$, $SE = 0.61$.[9] Key press accuracy was 6.1 percentage points higher for this speaker manipulation.

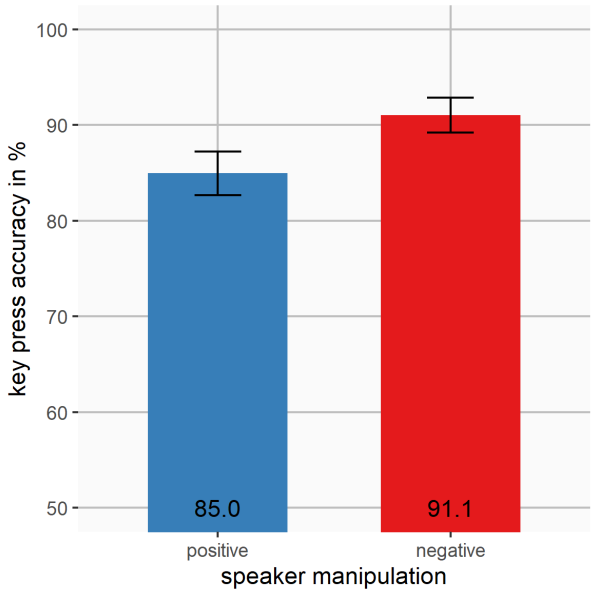


Figure 5.9: Key press accuracy as mediated by speaker manipulation (bars: mean, error bars: ±one SE, colour: speaker manipulation)

Figure 5.10 shows the effect of story type on key press accuracy. Pairwise comparisons showed significant differences between unchanged and related stories (higher key press accuracy in unchanged stories), $z = 2.54$, $p = .033$; but not between unchanged and unrelated stories, $z = 2.21$, $p = .082$; or related and unrelated stories, $z = -0.42$, $p > .999$. As the graph shows, accuracy was lowest for story pairs with semantically related changes. For this story type, the accuracy was at 82.9 %, compared to 87.2% for unrelated stories and 93.9% for unchanged stories.

[9] The model output in Table 5.10 shows simple effects of speaker manipulation and story type. This is how R generally prints the model coefficients. Here, there is a main effect of speaker manipulation and the corresponding coefficient of this main effect is provided. This coefficient was calculated manually based on the model output.

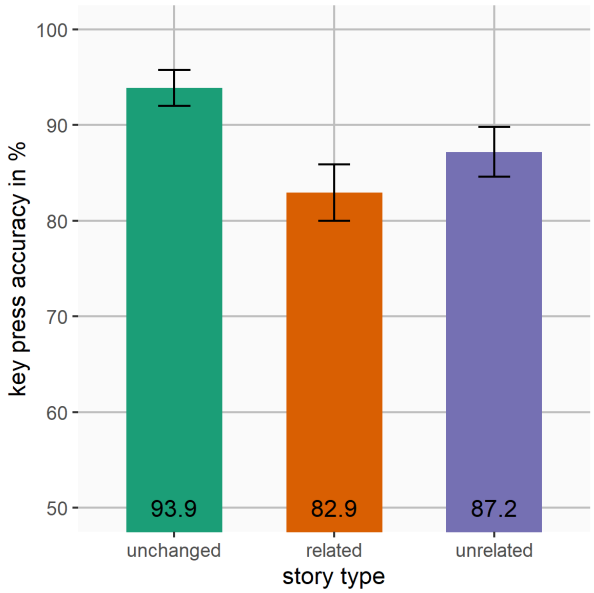


Figure 5.10: Key press accuracy as mediated by story type (bars: mean, error bars: ±one SE, colour: story type)

With regards to block, key press accuracy was lower during the second half of trials in a speaker's voice, β = -0.96, SE = 0.33. The effect of block is shown in Figure 5.11. There was an average difference of 7.3 percentage points between the two blocks, resulting in a mean accuracy of 84.4% during the second half.

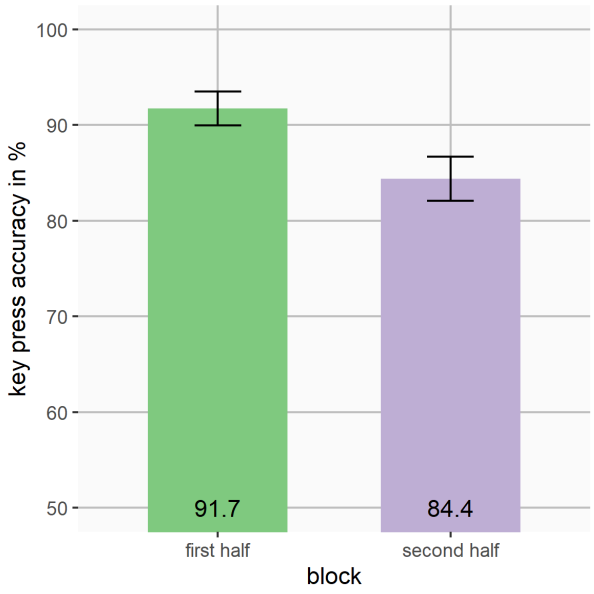


Figure 5.11: Key press accuracy as mediated by block (bars: mean, error bars: ±one SE, colour: block)

The interaction between story type and speaker manipulation is visualised in Figure 5.8 above. The main effects of story type and speaker manipulation emerge here in that, overall,

key press accuracy was lower for the negative speaker manipulation and for related stories. Pairwise comparisons were conducted between the speaker manipulations by story type and demonstrated significant differences between the positive and the negative speaker manipulation for related stories (higher key press accuracy in the negative manipulation), z = -3.18, p = .002; but not for unchanged stories, z = -1.32, p = .186; or for unrelated stories, z = 0.87, p = .387. When the speaker was presented positively, participants' key press accuracy was 17.1 percentage points lower for the related story type. Thus, the interaction between story type and speaker manipulation was carried by stories with semantically related changes.

5.7.2.2 Correction Accuracy

When the participants correctly indicated during a trial that there was a difference between the two stories, that trial was considered for the correction accuracy measure. As a consequence, the dataset consisted of 279 observations. There were 254 accurate corrections (91.0%) and 25 inaccurate corrections (9.0%). The correction accuracy per speaker manipulation and story type is provided in Figure 5.12. Overall, the accuracy rates were high and the participants were better at correcting semantically related as opposed to semantically unrelated changes, especially when the speakers were presented negatively.

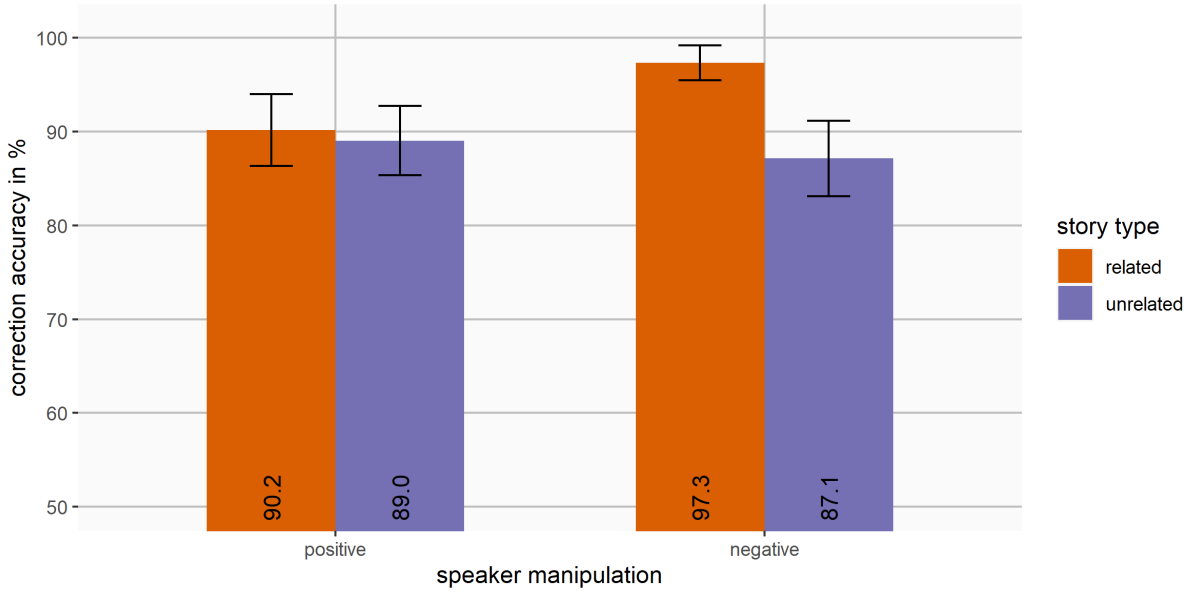


Figure 5.12: Correction accuracy as mediated by speaker manipulation and story type (bars: mean, error bars: ±one SE, colour: story type)

Correction accuracy was analysed by means of the below model. Its output is provided in Table 5.11.

glmer(accuracy_correction ~ speaker_manipulation x story_type + block + speech_rate_z + (1|participant) + (1|item))

Predictor	Estimate	Standard Error
(Intercept)	2.19	0.63
Speaker Manipulation (negative)	1.40	0.86
Story Type (unrelated)	-0.05	0.67
Speaker Manipulation (negative) : Trial Type (unrelated)	-1.43	1.02
Block (second half)	0.59	0.48
Speech Rate (z-scored)	-0.33	0.28

Table 5.11: Model output: correction accuracy

None of the model components reached significance: speaker manipulation, $\chi^2(1) = 1.92$, $p = .166$; story type, $\chi^2(1) = 1.70$, $p = .192$; block, $\chi^2(1) = 1.56$, $p = .212$; speech rate, $\chi^2(1) = 1.47$, $p = .225$; nor the interaction between speaker manipulation and story type, $\chi^2(1) = 2.12$, $p = .146$.

5.8 Summary of Findings

The aim of Experiment 3 was to find out how attitudes towards a speaker influence lexical processing and recall. Rather than choosing speakers of different varieties of English that had previously been found to be evaluated as more or less positive, the experiment included a three-fold manipulation of the participants' attitudes. The manipulation aimed at a positive/negative portrayal of the two speakers with regards to status and solidarity. These dimensions are very common in research into language attitudes. Using this methodology made it possible to capitalise on the existing recordings of TE_1 and TE_2 for the lexical decision and recall task from Experiment 1. It further avoided potential confound variables that the introduction of more speakers might have caused. Finally, the design of the study could be kept comparatively simple and easy to implement as an online experiment via LabVanced.

The results from the lexical decision task included an effect of speaker manipulation on accuracy as well as an effect of word frequency on LDLs. The participants' lexical decisions were

more accurate in the positive speaker manipulation. A potential link between this finding and cognitive theory is working memory. Previous research found a correlation between working memory capacity and emotional state (Figueira et al., 2017). Granted that the negative speaker manipulation resulted in the corresponding emotional state, it might have reduced working memory and, as a result, accuracy in lexical access. This explanation would, however, raise the question why there was no effect of speaker manipulation on LDLs. For LDLs, the results showed faster lexical decisions for more frequent words, which is a well attested pattern in lexical processing (see e.g. Grainger, 1990).

The findings from the recall task showed that the key press accuracy was significantly higher when the speaker was presented negatively rather than positively. Accuracy rates were overall lowest for related stories and decreased during the second half of trials in a specific speaker's voice. Importantly, the interaction between speaker manipulation and story type showed that, in the positive condition, the effect of semantic proximity on recall from Experiments 1 and 2 was reversed. Semantically related changes were detected less reliably. This could be due to differential processing modes induced by the two types of speaker manipulation. Specifically, a positive portrayal of the speakers might result in less item-specific and more relational encoding (see Kensinger, 2009; Levine & Bluck, 2004; Storbeck, 2013; Storbeck & Clore, 2005 in subsection 5.1.1). Relational encoding generates less granular memory representations, which might make it more difficult to catch semantically related changes between the stories.

With regards to the research questions raised in subsection 5.1.2, the key findings from Experiment 3 are summarised here:

(1) How do language attitudes affect lexical processing?
Lexical decision accuracy was generally higher when the speakers were presented positively rather than negatively.

(2) Is there evidence for adaptation in terms of lexical processing?
The results from Experiment 3 do not suggest adaptation, neither in terms of accuracy nor in terms of LDLs.

(3) What is the effect of word frequency on lexical processing?
Participants made their lexical decisions more quickly when the word frequency of the sentence-final target was higher.

(4) How do language attitudes and the semantic proximity of the change affect recall?
Key press accuracy was generally lower in the positive speaker manipulation. Additionally, an interaction was found for key press accuracy. Here, accuracy rates were especially low for stories with semantically related changes in the positive speaker manipulation. This might be due to less item-specific processing induced by positive affect. No effects were found for correction accuracy.

(5) Overall, is there evidence for less-detailed processing when the speaker is evaluated positively versus negatively?
Less-detailed processing appears to be more likely to occur in the positive speaker manipulation. Here, the reduced key press accuracy rates for related stories suggest that memory representations are not granular enough to catch the changes between stories. Thus, less-detailed processing might constitute one end of a processing mechanism scale, the other end of which is occupied by item-specific processing.

The theoretical implications of the results from this experiment are reviewed further in the following discussion chapter, which will also links the results of the three experiments back to existing literature and evaluates limitations as well as future directions of the research presented here.

Chapter 6

Discussion

6.1 Introduction

The online experiments from the previous three chapters investigated the effects of familiarity, intelligibility and attitude on speech processing, specifically lexical access and short-term recall. All three factors of interest were found to affect speech processing. In the following, they will first be reviewed individually, in response to the research questions from section 1.1 (recapped below). The findings of the current research will be discussed in the context of previous studies and how they fit or do not fit into existing theoretical frameworks. Next, potential interactions between the three factors will be taken into account, before the relevance of individual voice characteristics on research into speech processing will be addressed. Finally, limitations of the current research design will be focused on and potential mitigations thereof as well as new avenues for future research will be presented.

6.2 Recap of Research Questions

The research questions from each experimental chapter are recapped below. The following discussion of results will centre around these questions.

6.2.1 Familiarity

(1) How does accent familiarity affect lexical processing?

(2) Is there evidence for adaptation in terms of lexical processing?

(3) What is the effect of word frequency on lexical processing?

(4) How do accent familiarity and semantic proximity of the change affect recall?

(5) Overall, is there evidence for less-detailed processing of the unfamiliar versus familiar accent?

6.2.2 Intelligibility

(1) How do adverse listening conditions and accent familiarity affect lexical processing?

(2) Is there evidence for adaptation in terms of lexical processing?

(3) What is the effect of word frequency on lexical processing?

(4) How do adverse listening conditions, accent familiarity and semantic proximity of the change affect recall?

(5) Overall, is there evidence for less-detailed processing in quiet versus in noise and/or for the unfamiliar versus familiar accent?

6.2.3 Attitude

(1) How do language attitudes affect lexical processing?

(2) Is there evidence for adaptation in terms of lexical processing?

(3) What is the effect of word frequency on lexical processing?

(4) How do language attitudes and the semantic proximity of the change affect recall?

(5) Overall, is there evidence for less-detailed processing when the speaker is evaluated positively versus negatively?

6.3 Familiarity

The findings from the experiment on familiarity (see Chapter 3) were largely in line with previous research on lexical processing, which suggests a beneficial effect of familiarity with an accent. In general, the Tyneside participants' lexical decisions were more accurate for their own accent versus the unfamiliar New Zealand English (NZE). There was also an effect of word frequency, with more frequent targets being associated with higher lexical decision

accuracy. The interaction between accent and block showed that the participants' lexical decision accuracy increased for the Tyneside English (TE) speakers but decreased for the NZE speakers. With regards to lexical decision latencies (LDLs), there was a main effect of block such that participants made quicker decisions during the second half of trials for a speaker. Taken together, these results strongly suggest a familiarity benefit for lexical processing. For the familiar accent, the participants become more accurate and faster during the experiment. For the unfamiliar accent, there also is an increase in speed but this increase in speed is associated with a decrease in accuracy.

These findings resonate with past studies on the lexical processing of (un)familiar accents. For example, Floccia et al. (2006) found that their participants' LDLs were longer for unfamiliar accents if these accents were presented with the participants' home accents in a randomised manner. Similarly, in Floccia et al. (2009), LDLs were longer for Irish English in comparison with the participants' Plymothian home accent. This was the case for a randomised as well as a blocked presentation of accents. Impe et al.'s (2009) results for different accents of Dutch also followed this pattern and demonstrated an interesting processing imbalance with regards to standard accents. Their participants from Belgium were faster at making lexical decisions for Netherlandic accents than vice versa. That standard accents can function as centres of gravity, which are processed with relative ease, is sensible due to their widespread presence in broadcasting media and overt prestige.

The results from the recall task, however, were somewhat surprising in that the participants performed better for the unfamiliar accent. The general pattern for story type was that key press and correction accuracy was highest for unchanged stories, followed by related and unrelated stories. Importantly, these results do not suggest that unfamiliar L1 accents are processed in less detail. If this were the case, recall accuracy should be lower for the unfamiliar accent, especially when the change between the stories is semantically related. On the contrary, the unfamiliar accent seems to have been recalled better, which indicates more effortful encoding and retrieval by the listeners. It could be that listening to stories in an unfamiliar accent resulted in a more item-specific rather than relational processing mode. While this shift in processing mode has previously been found mostly when participants were exposed to challenging listening conditions or in a more negative emotional state (see sections 6.4 and 6.5), it is plausible that listeners employ more item-specific rather than relational processing for unfamiliar accents. This improves performance in the change-detection paradigm, which revolves around the identification of a single changed word between two stories. The finding

that related rather than unrelated changes were recalled better could be due to a semantic priming effect in that the original word primes a common semantic field (McNamara, 2005; Perea & Rosa, 2002), which eventually aids recall.

With regards to previous research on the recall of familiar versus unfamiliar L1 accents, one important study was conducted by Frances et al. (2018), who semi-replicated Lev-Ari and Keysar's (2012) change detection paradigm with different accents of Spanish. While the participants' ratings showed that they usually found the speakers of non-local accents less intelligible and more accented, this did not influence recall. Grohe and Weber (2018) showed a short-term familiarity effect for the recall of individual words in different accents of German. Reading the words aloud also improved recall performance. Clopper et al.'s (2016) findings suggest that lexical encoding, which ultimately supports recall, is stronger for accents that overtly and covertly carry more prestige and relevance. Thus, the current findings provide novel insights in that it was the unfamiliar accent that was associated with improved recall performance. However, this pattern was reversed under adverse listening conditions, the effect of which will discussed in the following section.

6.4 Intelligibility

For the purposes of the current research, intelligibility was defined as a categorical variable that coded if the stimuli in the lexical decision and recall task were presented in quiet or in noise at a signal-to-noise ratio (SNR) of 0dB (see Chapter 4). The reasoning behind including noise-masked stimuli was threefold. First, natural conversations rarely occur under the conditions found in a speech lab or recording booth. Thus, added noise increases the ecological validity of the study. Second, effects of accent familiarity on speech processing in some previous studies only surfaced when noise was added to the stimuli (e.g. Adank & McQueen, 2007; see below). Third, when the participants were asked how easy or difficult they found it to understand the two accents used in Experiment 1, they provided mean ratings of 5.8 (out of 6) and 4.6 for TE and NZE speakers, respectively. This indicates that they found all four speakers quite easy to understand, regardless of their accent. However, when noise was added to the signal, the participants from Tyneside rated the comprehensibility of NZE at an average of 2.4. Likewise, the mean comprehensibility ratings provided by the participants from New Zealand for TE in noise were 1.8. The decrease in comprehensibility ratings for the respective familiar accent of each participant group was much lower, which was indicative of the findings from Experiment 2.

Two sets of results were considered for intelligibility and its interaction with familiarity: first, data from Tyneside versus New Zealand participants for stimuli presented in noise and; second, data from two groups of Tyneside participants only for stimuli presented in quiet versus in noise. The latter was achieved through a comparison of the dataset from Experiment 1 and the Tynesider data from Experiment 2. The noise-only data showed that participants from both Tyneside and New Zealand were generally more accurate and faster in their lexical decisions for the TE speakers. This could be due to differential effects of the noise masking procedure on the four speakers used in the experiments (see section 6.7). Importantly, in addition to this main effect of accent, the interaction between accent and participant origin emerged as significant for the accuracy as well as the LDL data. With regards to accuracy, participants from New Zealand did better for their own accent than participants from Tyneside on NZE, although not to a significant extent. The Tynesiders' accuracy was higher than the New Zealanders' accuracy when they performed the lexical decision task for TE in noise. For LDLs, a similar pattern emerged in that Tynesiders made accurate lexical decisions more quickly for their own accent than New Zealanders.

These results further support the finding from Experiment 1 that accent familiarity is beneficial for lexical processing. TE in noise was processed both more quickly and more accurately by listeners who were familiar with it as compared to those unfamiliar with it. This conclusion was substantiated by the comparison of the Tynesider data from Experiments 1 and 2. For these participants, lexical decision accuracy dropped significantly more for the unfamiliar NZE when noise was added to the signal. LDLs were also much longer for NZE than for TE in noise. In terms of adaptation, the results suggested lower and upper bounds of this process in that there was evidence for adaptation in noise for TE and in quiet for NZE. The added noise might be necessary for adaptation effects to emerge for familiar accents because lexical processing might be at ceiling in quiet. For unfamiliar accents, lexical processing is more challenging even in ideal conditions, which leaves room for adaptation. Presenting unfamiliar accents in noise, however, might result in a floor effect for adaptation because the task becomes too challenging.

The above finding that effects of familiarity emerge more strongly in noise concurs with some previous research. For example, Adank and McQueen's (2007) Standard Southern British English (SSBE) participants' performance in a sensibility judgment task was only worse for Glaswegian English (GE) as compared to their own accent when the signal was noise-masked. Similarly, Adank and Janse (2010) showed that listeners were more tolerant of noise when it was

added to a standard accent of Dutch rather than an artificial accent created by the researchers. In Stringer and Iverson (2019), SSBE participants could repeat back disproportionately fewer key words for noise-masked sentences read by GE rather than SSBE speakers. This difference was less marked in quiet. In short, reduced intelligibility appears to interact with familiarity such that unfamiliar accents are especially difficult to process under adverse listening conditions.

As for the recall task, there were two key insights. First, in terms of key press accuracy in noise, there was a three-way interaction between accent, participant origin and story type such that Tyneside participants performed better when they heard their own accent and when the story did include a change. Furthermore, a comparison of the Tynesider data from Experiments 1 and 2 showed that their key press accuracy was lower in noise, especially for the unfamiliar NZE and for stories with a change. Second, in terms of correction accuracy in noise, participants were generally better at recalling semantically related as opposed to unrelated changes, regardless of the accent the stories were presented in. This finding was replicated from Experiment 1. Since there was no significant interaction with accent, the current pattern of results does not suggest a less-detailed processing of unfamiliar accents in noise. However, it does resonate with findings from previous research that lexical processing, and the subsequent encoding and recall thereof, is more item-specific under adverse listening conditions (Hällgren et al., 2001; Kjellberg et al., 2008; Ljung et al., 2009; Marsh et al., 2015). If noise-masking induced relational rather than item-specific processing, the recall of semantically related changes should be worse because the differences between related items are not encoded in enough detail to enable successful recall. Item-specific processing, on the other hand, focuses precisely on these details, which improves recall. Semantic priming effects from the original to the changed word might have further increased the recall performance for semantically related changes (McNamara, 2005; Perea & Rosa, 2002).

Following this discussion, the most valid generalisation from these results is to state a familiarity benefit for recall in noise, which echoes the findings for the lexical decision task and those of previous research into narrative recall, granted that the processing of unfamiliar accents requires more cognitive resources (e.g. Adank & Janse, 2010; Adank & McQueen, 2007; Piquado et al., 2012; Stringer & Iverson, 2019; Ward et al., 2016). In terms of speech processing, this benefit need not be due to less-detailed processing but could be caused by unfamiliar accents interfering in the lexical matching mechanism. This generalisation matches the findings from a comparison of the Tynesider data in Experiments 1 and 2. In quiet, their recall performance was actually slightly better for the unfamiliar accent. However, the

opposite pattern became apparent in noise. Again, this was due to stories that did include changes rather than unchanged ones. This comparison further demonstrates the importance of including a manipulation of intelligibility. It was only under adverse listening conditions that the recall disadvantage for the unfamiliar accent was found. This strongly suggests an interaction between familiarity and listening conditions in that negative effects on an unfamiliar accent might only surface in noise but not in quiet.

6.5 Attitude

The participants' attitudes towards the two TE speakers were manipulated in Experiment 3 (see Chapter 5) such that they were intended to hold positive attitudes towards one speaker and negative attitudes towards the other. This was achieved by means of providing information about the speakers, a judgement task and an investment task. The attitude manipulation focused mostly on the dimensions of status and solidarity. For the lexical decision task, there was a significant effect of speaker manipulation. When a speaker was presented negatively, the participants' lexical decisions were overall less accurate. The size of this effect was relatively small because performance in this task was close to ceiling, which is congruent with the findings on familiarity since, in Experiment 3, participants from the North East completed the experimental tasks in quiet and with an accent they were highly familiar with.

This finding is somewhat similar to Lev-Ari et al.'s (2019) results, which included poorer lexical processing of a Southern American accent if the participants held negative implicit attitudes towards its speakers and thought the speaker was from there. Generally, the decreased accuracy in the negative condition could be due to decreased working memory capacity. Figueira et al. (2017) found that negative affect results in reduced working memory capacity. Granted that a negative attitude towards a speaker and accent induces negative affect, this might result in poorer working memory and, thus, lexical decision accuracy. Interestingly, no according pattern was found for the LDLs in the attitude experiment. Increased task difficulty might have elicited a similar effect of attitude (and affect) on LDLs. There was, however, an effect of frequency on LDLs in that reactions were faster for more frequent targets. This finding is well attested in the speech processing literature (e.g. Grainger, 1990).

The results from the recall task were intriguing. Contrary to the results from the familiarity and the intelligibility experiments, semantic proximity did not generally aid recall at a closer inspection of the data. For the key press accuracy measure, there was an interaction between story type and speaker manipulation, most of the variance of which was associated with

stories with semantically related changes. For this story type, participants' accuracy was much better if the speaker was presented negatively. In the positive speaker manipulation, however, the lowest accuracy rate was found for stories with semantically related changes. These findings match with the wider literature on differential processing modes induced by affect. Although such research does not necessarily focus on the processing of speech, several studies have found that negative affect is associated with more item-specific processing, which encourages more deliberate and detailed bottom-up processing rather than the application of schematic information (Kensinger, 2009; Levine & Bluck, 2004; Storbeck, 2013; Storbeck & Clore, 2005). Here, the negative speaker manipulation might induce item-specific processing while the positive manipulation triggers relational processing. Recall, which was measured within a change-detection paradigm, benefits from item-specific processing since the targets are encoded in more detail rather than in their relation to other concepts stored in memory. This benefit becomes apparent for semantically related changes especially. If these changes were not encoded with sufficient detail, the participants' representations were not granular enough to catch the use of a semantically related word in the second story. For example, the use of *beer* in the first story and *cider* in the second story could be detected and corrected less easily following relational processing.

Two conclusions can be drawn from the results of the attitude experiment. First, the verbal change-detection paradigm used here corresponds to other recall tasks used in past research in that performance is enhanced by more item-specific processing, which, in turn, is induced by negative affect. This suggests that the way recall was operationalised here is valid and congruent with previous studies. Second, Lev-Ari and Keysar's (2012) notion of "less-detailed processing" and the notion of relational processing from the memory literature might refer to the same or at least very closely related processes. In fact, Lev-Ari and Keysar (2012, 535) comment that "less[-]detailed processing is similar to the notion of gist processing in the sense that both assume that the final representation focuses on important concepts. But our account could be different from gist processing as it assumes that peripheral details could also be specified if strong expectations direct attention to them". If the same process is at hand in Lev-Ari and Keysar's (2012) study and the current experiment, its activation appears to be modulated by competing factors. In Lev-Ari and Keysar (2012), relational processing would have been induced by expectations about L2 speech. Here, it was positive attitude and affect that promoted relational processing. This is interesting because L2 speech has been shown to generate less positive attitudes (e.g. Dragojevic, 2020). It might be that the influence of speaker identity, in particular their L1/L2 speaker status, outweighs attitude and affect

when processing mechanisms are selected. As will be commented on below, more research is needed here to find out if less-detailed and relational processing can be consolidated and how potentially competing parameters influence its activation.

6.6 Interfaces: Familiarity, Intelligibility and Attitude

The previous sections mostly looked at familiarity, intelligibility and attitude in isolation. The main experimental findings were discussed with regards to findings from previous studies and the underlying mechanisms of speech processing. The aim of this section is to broaden the scope and look at the interfaces between the three factors of interest. Relevant findings from the three experiments above will be reviewed alongside research that did not necessarily look at lexical processing or recall. Figure 6.1 serves to structure this discussion and shows that three pathways can be explored. In the following, the interface between familiarity and intelligibility is addressed first as Experiment 2 effectively encompassed both factors by incorporating a familiar as well as an unfamiliar accent under adverse listening conditions. Next, familiarity and attitude are considered, before the interface between intelligibility and attitude is focused on.

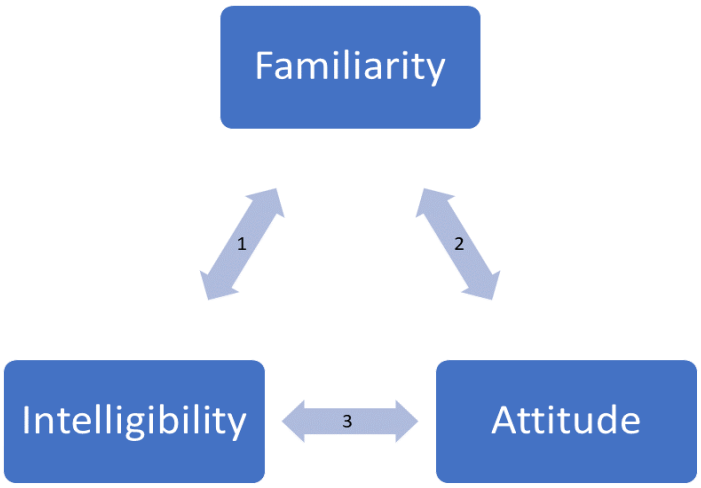


Figure 6.1: Interfaces between familiarity, intelligibility and attitude

6.6.1 Familiarity and Intelligibility

For the interface between familiarity and intelligibility, the results from Experiments 1 and 2 demonstrate that a beneficial effect of familiarity for lexical processing might only emerge

under adverse listening conditions. As shown above, this is in line with previous research (Adank & Janse, 2010; Adank & McQueen, 2007; Stringer & Iverson, 2019). Another relevant finding was the window of opportunity for the adaptation of LDLs. Adaptation effects were found in Tynesiders for the familiar accent in noise and the unfamiliar accent in quiet. These combinations of familiarity and intelligibility made lexical processing more challenging and opened a window of opportunity for adaptation to occur, without increasing the task difficulty to such an extent that adaptation was not possible. In terms of recall, performance by Tyneside participants was better for their own accent than NZE in noise. Additionally, both in quiet and in noise, there was a stable effect of story type. Semantic proximity between the changed words aided recall. This suggests a combination of semantic priming and more item-specific processing under adverse listening conditions.

Apart from these experimental findings, there is also research on the effects of comprehensibility that is relevant here. Comprehensibility is a metacognitive cue and refers to how easy or difficult a listener finds it to understand a speaker (Derwing & Munro, 1997; Oppenheimer, 2008). If listeners perceive speakers of an accent as hard to understand, these perceptions might influence their processing. An indication of this is Hanulíková et al.'s (2012) finding that grammatical errors in L2 speech elicit smaller P600 effects in L1 listeners of Dutch. Another example is Babel and Russell (2015), who found that participants with a social network consisting predominantly of Asian Canadians transcribed the speech of these individuals worse than participants with a predominantly White Canadian social network.[1] This is surprising as it would be expected that such a social network would increase exposure to and familiarity with the speech of Asian Canadians. Babel and Russell (2015, 2831) argue that this effect "is due to stereotyped associations learned through interactions with Asian Canadians". Few instances of interactions with Asian Canadians whose linguistic proficiency of English is limited warps listeners' expectations towards a generally lower proficiency of Asian Canadians. Most importantly to the current discussion is that this finding suggests that increased familiarity might, in fact, be detrimental for speech processing if it is not accompanied by increased comprehensibility. The evaluation of someone's accent based on their ethnicity also alludes to the role of attitudes, which will be considered next.

[1] It must be noted that the social network was operationalised by Babel and Russell (2015, 2827) rather crudely by asking participants whether their social network was predominantly White or Asian. This self-report was not backed up by an ethnographic analysis of the composition of their network.

6.6.2 Familiarity and Attitude

Interactions between familiarity and attitude were not directly explored in the current research since Experiment 3 only included TE, which the participants were highly familiar with.[2] In terms of lexical processing for the two attitude manipulations, it was found that a negative presentation of the speaker was associated with poorer lexical decision accuracy. However, negative attitudes could be beneficial for recall in that they trigger more item-specific processing through negative affect. In Experiment 1, a recall advantage was found for the unfamiliar accent. Thus, it might be the case that a negative presentation of a speaker with an unfamiliar accent results in even better recall if the effects of attitude/affect and (un)familiarity are additive. Importantly, this could only be reasonably assumed in quiet since Experiment 2 demonstrated that adverse listening conditions elicit a different pattern of results.

More broadly, increased familiarity with an accent appears to be a double-edged sword in terms of attitudes. Derwing et al. (2002), for example, found that improvements in attitudes towards L2 speech were highest if their participants learned about cross-cultural communication and the linguistic characteristics of L2 speech. Babel and Russell (2015), on the other hand, suggest that increased exposure to an accent and, thus, increased familiarity with it might actually have a detrimental effect if stereotypes are learned and reinforced (see Dragojevic's (2020) fluency principle in the following subsection). Clopper (2017) further suggests that salience or level of enregisterment plays a role for the relationship between familiarity and attitude and the effect of these two factors on speech processing. Specifically, Clopper (2017) found higher processing costs for a non-local and non-enregistered accent (Northern American) than for a non-local but enregistered accent (Southern American). If an accent is socially salient, that is it is easily identifiable and strongly associated with certain stereotypes, this might boost familiarity and, ultimately, aid speech processing. In this context, Lev-Ari et al.'s (2019) finding that lexical access is partly mediated by implicit attitudes becomes relevant. They found that the influence of indexical information ('speaker from the American South') on the selection of lexical competitors increased for participants that held a negative implicit bias against a Southern accent. More research is needed to evaluate if this is only the case for negative attitudes or if positive attitudes will have a similar effect.

[2] As will be shown in subsection 6.7.1, however, individual voice features may still influence language attitudes within the same accent.

6.6.3 Intelligibility and Attitude

Finally, the interface between intelligibility and attitude needs to be considered. Again, this interface was not directly investigated as Experiment 3 did not include a manipulation of listening conditions. However, the current findings and previous research suggest that adverse listening conditions and negative attitudes/affect promote more item-specific rather than relational processing. If these effects are additive, it could be that listeners who hold negative attitudes towards speakers perform especially well in noise, at least for the recall task. In terms of lexical processing, the pattern might be the opposite because, first, task difficulty increased in noise and; second, negative attitudes/affect were shown to be associated with decreased working memory capacity. Generally, increasing task difficulty by masking the signal with noise might overshadow any effects induced by attitude.

The general link between intelligibility, comprehensibility and language attitudes has been researched by Dragojevic et al. (2017) and Dragojevic (2020), who demonstrated how the development of attitudes is likely driven by comprehensibility. In their 2017 study, they aimed to examine if heavily accented L2 speech is evaluated more negatively in attitudinal studies because it induces processing disfluencies and/or because speakers with a stronger accent are more prototypical members of a group against which listeners hold biases. The effect of processing (dis)fluency on the development of language attitudes can be direct (see Oppenheimer, 2008), or indirect through affect: "the more difficult a speaker's speech is to process, the more negative affect listeners are likely to experience [...] and, in turn, the more negatively they are likely to evaluate the speaker [...]" (Dragojevic et al., 2017, 391). Therefore, language attitudes might be affected by the prototypicality of the speaker (pathway ab; see Figure 6.2), directly by processing fluency (pathway cd) or indirectly by the mediation of processing fluency through affect (pathway cef).

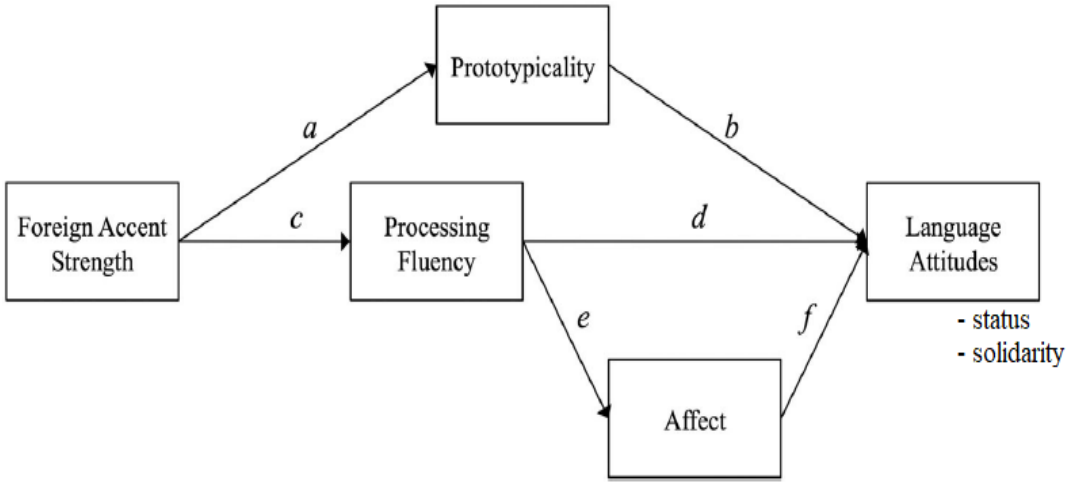


Figure 6.2: Pathways for the influence of L2 accents on language attitudes (Dragojevic et al., 2017, 388; slightly altered)

To test these three pathways, Dragojevic et al. (2017, 391-393) designed a matched-guised paradigm with a speaker of English with either a mild or a strong L2 accent. Participants listened to a story narrated by one or the other speaker and then rated the speaker on several scales, including attitudes, affect experienced while listening and comprehensibility.[3] They found that the strength of the accent did not influence solidarity ratings. The status ratings, however, were mediated by processing fluency/comprehensibility, both directly and indirectly through affect, while prototypicality did not have a significant effect (Dragojevic et al., 2017, 393-394, 396). The effect of processing fluency on status ratings is evident along both the direct pathway (cd) and the one mediated by affect (cef). Thus, negative status ratings are elicited because listeners experience heavy L2 accents as more difficult to process and because these processing difficulties result in negative emotions.

In the follow-up experiments by Dragojevic (2020), it was found that increases in processing fluency/comprehensibility, for example through the presence of subtitles, had the reverse effect: Language attitudes are more positive if cues that aid fluency are available. These effects were stronger for status than solidarity ratings. Based on these findings, Dragojevic (2020, 173) formulates the fluency principle: "Listeners' processing fluency (i.e., the ease with which listeners process a person's speech) [see footnote 3] is a general metacognitive cue to their language attitudes. Increases (decreases) in listeners' processing fluency, regardless of source, can positively (negatively) bias their evaluations of speakers (especially ratings of

[3] Dragojevic et al. (2017) and Dragojevic (2020) use the term *processing fluency* to refer to a metacognitive cue of the perceived ease of processing. This is congruent with the definition of *comprehensibility* presented in subsection 2.5.2.1.

speakers' status), independent of stereotyping. Processing fluency can exert its effects on speaker evaluations directly through the application of naïve theories, and indirectly through affect" (see above and Figure 6.2 for an explanation of these pathways). Importantly, Dragojevic's (2020) conclusions are all based on explicit attitudes, which need not converge with implicit attitudes (Pantos & Perkins, 2012). Further research is needed to see how implicit attitudes are affected by changes in processing fluency/comprehensibility.

In conclusion, the interfaces between familiarity, intelligibility and attitude demonstrate the complexity of speech processing when only three factors are considered. Further complexity is added by the individual voice characteristics of the speakers used. These are the topic of the following section.

6.7 Influence of Individual Voice Characteristics

To increase ecological validity, linguistic experiments often include more than one speaker. In fact, if there is only one speaker or one speaker per condition, it might not be clear if a significant pattern in the results is due to the experimental manipulation or idiosyncratic characteristics of the chosen speaker. For the current research, two speakers were included for each accent to mitigate the effect of these characteristics and to make the results more generalisable. Although the speakers within each accent were matched as best as possible for gender, age, location and residence history, it became apparent that individual voice characteristics still played a role as to how the speakers were perceived and how their voices responded to experimental manipulations. This section serves to reflect on two aspects in this regard. The differential attitudinal evaluations of TE_1 and TE_2 illustrate how intra-accent variation might have influenced the results. The effects of the noise masking procedure on the speakers from Tyneside versus those from New Zealand is an example of speaker-specific variation and its potential impact on the experimental findings.

6.7.1 Attitudes towards Tyneside Speakers

Both TE_1 and TE_2 were deemed representative speakers of TE by phonetically trained listeners. Using only TE speakers in the experiment on attitude was useful because the participants from the North East were highly familiar with this accent and had a strong social network in the region, as demonstrated by their responses in the questionnaire at the end of the experiment. Since Experiments 1 and 2 showed significant effects of familiarity on lexical

access and recall, it was essential to control this factor during Experiment 3. Previous research into language attitudes found that TE generally receives higher ratings in terms of social attractiveness than prestige. In Sharma et al. (2022), TE was ranked 32nd and 26th for prestige and social attractiveness, respectively, out of 38 global accents of English. While these ranks do not seem very high, it is likely that the participants from the current research would have given comparatively higher ratings for TE because of a positive in-group effect on the ratings (Dragojevic et al., 2017). In fact, the participants in the pilot study for Experiment 3 gave quite high ratings for the TE speakers, even when they were portrayed negatively.

However, the ratings varied between TE_1 and TE_2 in that, across almost all rating categories and speaker manipulations, TE_1 was ranked higher than TE_2 in the pilot study (see Table 6.1). This consistency in ratings suggests that voice characteristics beyond the regional accent, which the two speakers share, influenced the results. In fact, there have been several studies that investigated the correlation between acoustic measures and attributions of solidarity, attractiveness, charisma and trust, to name but a few examples (e.g. Rosenberg & Hirschberg, 2021; Strangert & Gustafson, 2008; Torre, 2017; Torre, Goslin, & White, 2020; Torre & White, 2021; Weiss & Burkhardt, 2010, 2012). These studies found that positive evaluations of speakers were associated with a larger F0 range, fewer disfluencies, lower speech rate and a smiling voice.

Speaker Manipulation	Rating Scale	TE_1	TE_2
positive	*status*	5.8 (0.7)	4.6 (1.7)
	solidarity	6.0 (0.6)	5.0 (1.6)
	dynamism	5.2 (0.8)	5.2 (1.4)
negative	*status*	4.4 (1.6)	3.2 (1.4)
	solidarity	4.7 (1.9)	3.7 (1.0)
	dynamism	4.3 (1.4)	4.4 (1.2)

Table 6.1: Ratings as mediated by speaker manipulation, rating scale and speaker in the pilot study from Experiment 3 (mean, standard deviation)

Since the speakers recorded the stimuli in a very controlled environment and in citation rather than conversational speech style, disfluencies were not present in the data. In fact, if recordings contained hesitations or disfluencies, they were re-recorded with the speakers. As no video recordings were made, it was difficult to assess if speakers smiled while producing the stimuli. F0 range and speech rate, on the other hand, could be measured with the recordings available and might offer a first insight as to why listeners rated TE_1 higher than TE_2. For

both measures, all recordings for experimental trials in the lexical decision task and in the recall task were used. Thus, the sentences with words in the lexical decision task and the stories with(out) a change in the second sentences in the recall task were the dataset for this analysis. The files were concatenated in Praat. Next, F0 and speech rate measures were taken.

An overview of F0 measures is provided in Table 6.2.[4] The last row of Table 6.2 shows an F0 interval that ranges two standard deviations above and below the mean F0. The numbers show that this interval is larger for TE_2 (228 Hz) than for TE_1 (168 Hz). As F0 was found to correlate positively with attitude ratings, this result is not in line with what previous research would predict.[5]

F0 Measure	**TE_1**	**TE_2**	**NZE_1**	**NZE_2**
Mean	167 Hz	227 Hz	203 Hz	161 Hz
Standard Deviation	42 Hz	57 Hz	42 Hz	43 Hz
Range with 2 Standard Deviations	[83 Hz; 251 Hz]	[113 Hz; 341 Hz]	[119 Hz; 287 Hz]	[75 Hz; 247 Hz]

Table 6.2: F0 measurements for the four speakers in the experiments

Table 6.3 shows speech rate measures for the four speakers. TE_2's speech rate was, on average, 0.2 syllables per second higher than TE_1's. In terms of the spread of the data, there is no difference between the two speakers as their standard deviation is equal. Hence, the range given in the final row has the same extension around the mean for TE_1 and TE_2. The results here are congruent with the finding from previous research that speech rate correlates negatively with attitude ratings. TE_1's speech rate is lower and she received higher ratings from the participants (see footnote 5).

Speech Rate (syllables/second)	**TE_1**	**TE_2**	**NZE_1**	**NZE_2**
Minimum	3.6	3.6	3.8	4.3
Maximum	5.7	6.3	5.6	6.8
Mean	4.4	4.6	4.8	5.2
Standard Deviation	0.5	0.5	0.4	0.4
Range with 2 Standard Deviations	[3.4; 5.4]	[3.6; 5.6]	[4.0; 5.6]	[4.4; 6.0]

Table 6.3: Speech rate measurements for the four speakers in the experiments

[4] The measures are given for all four speakers because F0 range will become relevant again in subsection 6.7.2. The same applies to Table 6.3.

[5] No statistical analysis was carried out here because, effectively, there were only two data points that could be used: one of the F0 measures per speaker and their average ratings on one of the rating scales. This would have resulted in an extremely underpowered analysis.

Taken together, only speech rate provides some explanatory power here in terms of the differential evaluations of TE_1 and TE_2. A thorough investigation of further acoustic features is beyond the scope of this thesis but would provide an intriguing angle for future research (see section 6.8). Importantly, the speaker-based approach here does not take into account the listener's perspective. As much as there is variation between TE_1 and TE_2 with regards to their voice features, different listeners could vary in their evaluation of these features.

Including more than one speaker will inevitably lead to the introduction of voice characteristics that are very hard to control and that might affect the results. However, as long as these characteristics do not become confound variables that overshadow the effect of the main experimental manipulation, the inclusion of several speakers should be encouraged in order to increase the generalisability of the results.

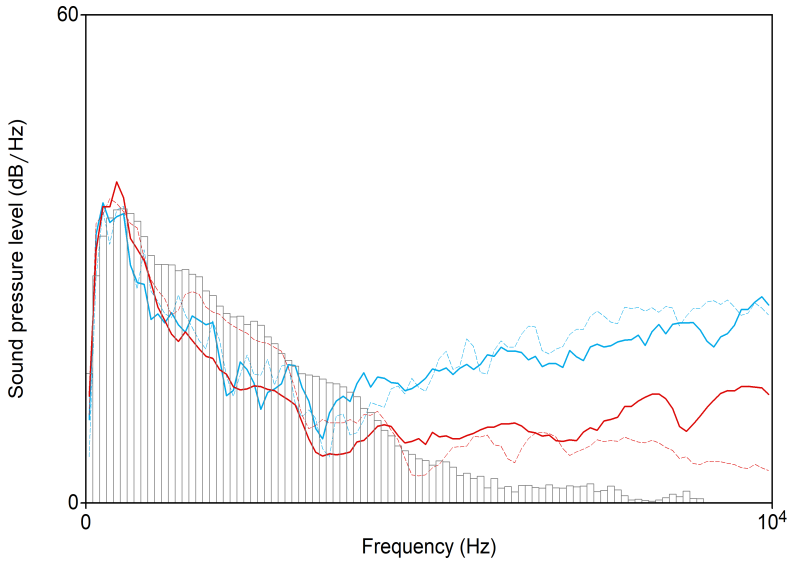
6.7.2 Noise-Masking

Although the same noise masking procedure was applied to the four speakers in Experiment 2, it seemed as though the two NZE speakers' voices were more effectively masked. To mitigate the potential impact of this on the intrerpretation of accent effects, participants from both Tyneside and New Zealand were recruited for the experiment (see section 4.2.2). The question remains, however, why some voices appeared to be disproportionately more affected. In general, speakers might exhibit certain characteristics that make their speech more intelligible and, for example, result in better transcription performance for these speakers (Knight, 2022). Bradlow, Torretta, and Pisoni's (1996) results showed that (English) speakers were more intelligible if they were female, had a wider F0 range and produced more dispersed vowels, especially on the close/open dimension of their vowel space.

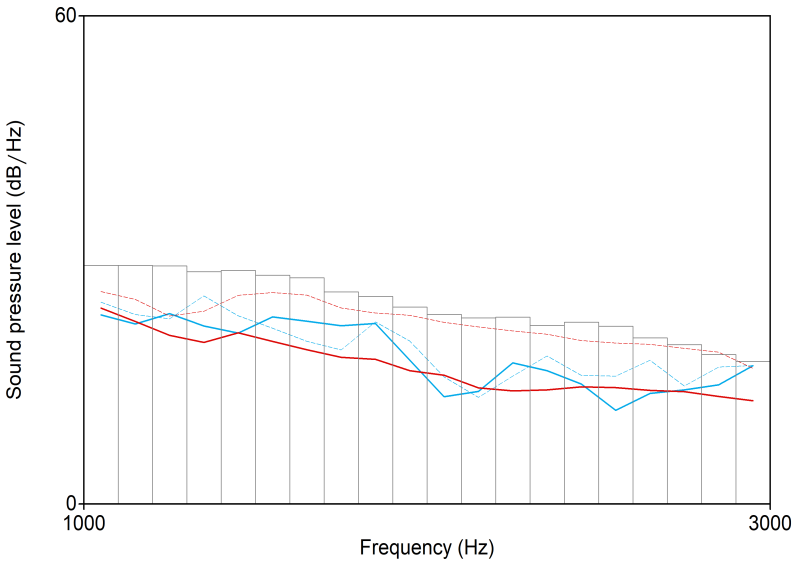
Another relevant line of research here are studies on clear speech. Clear speech is less inherent to specific speakers' voices. Instead, it is a speech style that any speaker can adopt in circumstances that might impede on communication, such as settings with a lot of background noise (Uchansky, 2008). While the speakers used for the current experiments did not receive instructions that would necessarily induce such a speaking style, its characteristics are still helpful to consider as clear speech results in increased intelligibility. In their review of clear speech, Smiljanić and Bradlow (2009) differentiated global and segmental characteristics of clear speech. In terms of global characteristics, they identified, among others, slower speech rate, greater F0 range, higher intensity and increased amplitude in the 1 to 3 kHz interval of long-term average spectra. At the segmental level, clear speech exhibits, for example, more

extreme positions of vowels in the vowel space and clearer releases of plosives. Importantly, the features of clear speech are language-dependent and vary according to where disambiguation is needed. For example, Bradlow et al. (1996) above found that the F1 dimension of vowels was especially relevant for higher intelligibility in English, which makes sense given that English vowels are more tightly clustered along the close/open rather than the front/back dimension.

F0 and speech rate measures for the four speakers are shown in Tables 6.2 and 6.3, respectively. In terms of F0, the two NZE speakers had a very a similar range around the mean to TE_1. Overall, TE_2 had the highest range (228 Hz). However, given that TE_1 seemed to be easier to hear in noise than NZE_1 and NZE_2 but shared a similar F0 range, fundamental frequency alone cannot be the only relevant predictor here. The pattern for speech rate is congruent with previous research in that NZE_1 and especially NZE_2 have higher speech rates than the TE speakers (see footnote 5). Importantly, both F0 range and speech rate do not consider how the speakers' voices interacted with the multi-speaker babble noise that was used in the masking procedure. This angle is taken in the following via an investigation of long-term average spectra.



(a) frequency range: 0 - 10 kHz



(b) frequency range: 1 - 3 kHz

Figure 6.3: Long-term average spectra at different frequency ranges for speakers from Experiment 2 (blue solid: TE_1, blue dashed: TE_2, red solid: NZE_1, red dashed: NZE_2, bars: noise signal)

The long-term average spectra of the experimental stimuli produced by the four speakers for Experiment 2 were computed in Praat. As can be seen in Figure 6.3, the spectra were computed in two frequency ranges: one from 0 to 10 kHz and another one from 1 to 3 kHz as the latter has been found to be especially important for intelligibility (Smiljanić & Bradlow, 2009). For both figures, the TE and NZE speakers are shown in blue and red, respectively. TE_1 and NZE_1 are represented by solid lines while the dashed lines show the spectra for TE_2 and NZE_2. The superimposed grey bars indicate the spectrum of the noise file that was

used to mask the stimuli. The spectrum from 1 to 3 kHz shows overall less energy for NZE_1 across the frequency range. However, NZE_2 actually has the most energy across almost the entire range. The top spectrum demonstrates that the TE and NZE speakers mainly differ for frequencies above 4 kHz. Here, the spectral energy is consistently higher for the TE speakers. While this area of the graph is not part of the critical range from 1 to 3 kHz, the consistently lower long-term average spectral energy for the NZE speakers might explain why they are more difficult to hear in noise. This observation, in conjunction with what was discussed above for F0 range and speech rate, offers a first insight into the differential results of the noise masking procedure. The following section addresses how future research could expand on this and other insights of the current research.

6.8 Directions for Future Research

As discussed above, the results of the three experiments presented here have provided new insights into how familiarity, intelligibility and attitude affect the lexical processing and recall of speech. Given the limited time scale and scope of the project, not all avenues of research could be explored in full detail. The purpose of this section is to identify directions for future research.

6.8.1 Attitude Manipulation

In Experiment 3, the participants' attitude was manipulated in three different ways. The results of the corresponding pilot study supported the success of this manipulation. The results were interpreted within an account that assumes that negative attitudes result in negative affect, which, in turn, triggers more item-specific processing. Importantly, only two speakers of the same accent were used in order to control the participants' familiarity with this accent. While this was essential for the current set of experiments, there are two aspects that require further consideration.

First, the link between attitude and affect needs to be further explored. While the current assumption is reasonable that the two are positively correlated, the pilot study measured affect only indirectly through the responses in the language attitude questionnaire. Future research could substantiate the assumption that negative affect is induced if subjects interact with an attitude object which they hold negative attitudes towards. Such research should incorporate speech stimuli to maintain construct validity.

Second, there is likely a difference between experimentally induced attitudes towards accents and attitudes that have developed over decades by means of socialisation, metalinguistic commentary and individual preferences. The advantage of attitude manipulation by experimental design is that individual voice characteristics can be controlled for through a limited number of speakers included. However, since the participants for Experiment 3 were recruited from the North East, they might have generally held more positive attitudes towards TE and an in-group bias. It would be very interesting to compare the results found for the negative attitude manipulation with results obtained for similar tasks with an accent that is usually negatively evaluated by a listener population. Based on language attitudes across the UK, suitable accents for this negative condition could be Birmingham, Liverpool or Essex English (Sharma et al., 2022). Accent selection could also occur on a more local level, which might help to control familiarity more easily. For example, there is a long-established rivalry between Newcastle and Sunderland (e.g. Beal et al., 2012; Burbano-Elizondo, 2015) and the relative geographical proximity of these two cities makes it more likely for listeners from Newcastle to be familiar with Sunderland English and vice versa. Again, the results for the corresponding negatively evaluated accent could be compared to the ones obtained here for the negative condition. Another possibility would be to conduct an experiment with traditionally positively and negatively evaluated accents as well as a deliberate attitude manipulation. While the design of this experiment would be more complex, it might result in even more negative/positive evaluations of the speakers, which could then affect speech processing.

6.8.2 Speaker Selection

The discussion of individual voice characteristics in section 6.7 has shown that including more than one speaker per condition is essential to increase the generalisability of experimental results. However, it inevitably also results in the introduction of idiosyncratic characteristics, which influence results beyond the desired experimental manipulation. For example, as the results from Experiments 2 and 3 have indicated here, voices might be harder or easier to hear under noise and listeners might like the voices of certain speakers more less. In terms of speaker choice and subsequent recording, the current research was limited because of the COVID-19 restrictions in place at the time of recording. However, if more time and resources are available for future studies, a two-step speaker selection might be helpful. Provided that two speakers are to be selected per accent used, several speakers of each accent could be recorded first. These first recordings would not need to include all stimuli to be incorporated into the main experiment. Instead, they could be the basis of a pilot investigation, which could look at the

long-term average spectra of each speaker, for example. Next, two speakers per accent with similar spectra are selected, which should result in more even noise-masking results across speakers. Along a similar vein, for language attitudes, speakers could be selected that were evaluated similarly by listeners during piloting. With regards to statistical modelling, speaker could be included in the random effects structure of the models. However, this might result in the necessity of high numbers of participants to yield sufficient statistical power. Therefore, a thorough speaker selection might be more sensible, before data is collected for future studies.

6.8.3 Exploration of Further Interfaces and Replication of Findings

The discussion of the three interfaces from Figure 6.1 has provided valuable insights as to how familiarity, intelligibility and attitude might interact. It has further demonstrated that there are more avenues to be explored here. For instance, intelligibility and attitude could be investigated in a single study, within which positively and negatively evaluated accents are presented under ideal and adverse listening conditions. A similar approach could be taken to research the interface between familiarity and attitude in more detail. Even a combination of all three factors within one study would be possible, although this could result in an overly complicated design, which might lead to exhausting experiments for the participants and insufficient statistical power. Another factor that has not been considered in this thesis but in many previous studies is working memory, which could be integrated into Figure 6.1 and future experiments. As can be seen, the current experiments are only a starting point for more research in this area of speech processing, which might also consolidate potentially congruent cognitive processing mechanisms that have been treated separately in the literature so far (e.g. less-detailed, gist, relational versus item-specific processing, etc).

Finally, several aspects of the current research are novel and original and, therefore, require replication. For example, Experiment 2 is the first study of its kind to research the immediate recall of short stories under adverse listening conditions as mediated by the participants' familiarity with the two accents used. In addition, the attitude manipulation in Experiment 3 is an innovative approach to the research of the effects of positive versus negative attitudes (and affect). The implementation of novel methodologies is necessary to generate new findings and to advance theories. However, these findings also need to be consolidated by means of replication in order to strengthen their reliability and validity. Otherwise, there is a risk of a 'replication crisis', which, according to Winter (2019, 47), is "looming around the corner" for linguistics (see also Ioannidis, 2005; Nieuwland et al., 2018). Therefore, the findings

from the experiments presented in this thesis should be replicated and expanded upon, ideally under controlled circumstances. The headphone check implemented here was an attempt to focus the participants' attention as much as possible in the online experiment. However, now that experiments under laboratory conditions are possible again, a replication of the experiments in the lab would be an ideal way of minimising distractions for the participants and, thus, of consolidating the current findings.

Chapter 7

Conclusion

The aim of this thesis was to investigate how lexical processing and short-term recall are mediated by familiarity, intelligibility and attitude. For each factor of interest, a separate online experiment was conducted. One of the aims of this research was to investigate if less-detailed processing, which has been proposed for L2 speech (Lev-Ari & Keysar, 2012), extends to (unfamiliar) L1 accents. Lexical processing was measured via a lexical decision task, during which participants had to decide if the sentence-final target was a real English word or a nonword. The recall task included the subsequent presentation of two short stories. Here, the participants first indicated if the two stories were identical or different. If they chose the latter, they were asked to recall the changed words between the two stories. The changed words were either semantically related or unrelated.

In terms of lexical processing, there were four key findings. First, there was a clear familiarity benefit, which emerged both in quiet and in noise. In quiet, Tynesiders' lexical decisions became faster and more accurate for Tyneside English (TE) while a speed accuracy tradeoff was found for New Zealand English (NZE). Under adverse listening conditions, the lexical processing of TE and NZE was more successful in participants from Tyneside and New Zealand, respectively. Second, adaptation to accents appears to take place within a window of opportunity in that it occurs when lexical processing is challenging (familiar accent in noise or unfamiliar accent in quiet) but not too easy (familiar accent in quiet) or too challenging (unfamiliar accent in noise). Third, negative attitudes were associated with poorer lexical access, potentially due to reduced working memory capacity. Finally, performance in the lexical decision task was better for more frequent targets.

The findings from the recall task can be summarised in three points. First, a familiarity benefit emerged for this task only under adverse listening conditions. In quiet, Tynesiders' recall was actually better for short stories presented in NZE rather than TE. When noise was added to the signal, on the other hand, participants from Tyneside performed better for their own accent. Second, semantic proximity between the changed words generally seemed to aid recall, unless the speaker was presented positively in Experiment 3. Finally, a negative presentation of the speaker in Experiment 3 was associated with improved recall. These findings were interpreted within a framework of semantic priming and the competing cognitive operations of item-specific versus relational processing. Recall benefits from semantic priming and item-specific processing. Presenting an unfamiliar accent might elicit item-specific processing. The effect of this processing mode might be beneficial in quiet but not strong enough to emerge when task difficulty is increased under adverse listening conditions. Here, the familiarity benefit seems to outweigh the effects of processing mode induced by the listening conditions. Negative attitudes towards a speaker also appear to induce item-specific processing and, thus, improved recall by means of affect. Positive attitudes, on the other hand, promote relational processing, which results in the poorer recall of semantically related changes. Relational processing seems to be similar in nature to Lev-Ari and Keysar's (2012) less-detailed processing. Importantly, the current results do not suggest such a processing mode for unfamiliar as opposed to familiar accents.

In conclusion, the current research produced novel insights into the lexical processing and recall of regional L1 speech. Familiarity, intelligibility and attitude were investigated by means of separate experiments. The interactions between these factors remain underexplored and should be addressed in future studies, which should also look into the specific effects of individual voice characteristics.

References

Ackerman, L. (2019). *Ordinal data.* Retrieved June 30, 2023 from https://verbingnouns.github.io/notebooks/ordinal_data.html.

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(2), 520–529. doi: 10.1037/a0013552

Adank, P., & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychology and Aging*, *25*(3), 736–740. doi: 10.1037/a0020054

Adank, P., & McQueen, J. M. (2007). The effect of an unfamiliar regional accent on spoken word comprehension. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS), August 06-10, 2007* (pp. 1925–1928). Saarbrücken, Germany: Universität des Saarlandes.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*(3), 163–187. doi: 10.1016/0010-0277(94)90042-6

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics*, *32*(2), 233–250. doi: 10.1016/S0095-4470(03)00036-6

Araya-Salas, M., & Smith-Vidaurre, G. (2017). WarbleR: An r package to streamline analysis of animal acoustic signals. *Methods in Ecology and Evolution*, *8*(2), 184–191. doi: 10.1111/2041-210X.12624

Araya-Salas, M., & Smith-Vidaurre, G. (2023). *WarbleR.* Retrieved June 30, 2023 from https://marce10.github.io/warbleR/index.html.

Arndt, J., & Reder, L. M. (2003). The effect of distinctive visual information on false recognition. *Journal of Memory and Language*, *48*(1), 1–15. doi: 10.1016/S0749-596X(02)00518-1

Audacity Team. (2023). *Audacity: Free, open source, cross-platform audio software.* Retrieved June 30, 2023 from https://www.audacityteam.org/.

Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America*, *137*(5), 2823–2833. doi: 10.1121/1.4919317

Baranowski, M. (2017). Class matters: The sociolinguistics of GOOSE and GOAT in Manchester English. *Language Variation and Change*, *29*(3), 301–339. doi: 10.1017/S0954394517000217

Barlaz, M. (2023). *Ordinal logistic regression in R.* Retrieved June 30, 2023 from https://marissabarlaz.github.io/portfolio/ols/.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... Krivitsky, P. N. (2023). *Lme4: Linear mixed-effects models using 'Eigen' and S4.* Retrieved June 30, 2023 from https://cran.r-project.org/web/packages/lme4/index.html.

Bauer, L., & Warren, P. (2008). New Zealand English: Phonology. In K. Burridge & B. Kortmann (Eds.), *Varieties of English 3: The Pacific and Australasia* (pp. 39–63). Berlin, Germany: De Gruyter Mouton.

Bauer, L., Warren, P., Bardsley, D., Kennedy, M., & Major, G. (2007). New Zealand English. *Journal of the International Phonetic Association*, *37*(1), 97–102. doi: 10.1017/S0025100306002830

Baxter, F., Khattab, G., Krug, A., & Du, F. (2022). Recall of own speech following interaction with L2 speakers: Is there evidence for fuzzier representations? *Frontiers in Communication*, *7*, 1–11. doi: 10.3389/fcomm.2022.840041

Beal, J. C., Elizondo, L. B., & Llamas, C. (2012). *Urban North-eastern English: Tyneside to Teesside.* Edinburgh, United Kingdom: Edinburgh University Press.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. doi: 10.1006/game.1995.1027

Bishop, H., Coupland, N., & Garrett, P. (2005). Conceptual accent evaluation: Thirty years of accent prejudice in the UK. *Acta Linguistica Hafniensia*, *37*(1), 131–154. doi: 10.1080/03740463.2005.10416087

Bladon, R. A. W., Henton, C. G., & Pickering, J. B. (1984). Towards an auditory theory of speaker normalization. *Language & Communication*, *4*(1), 59–69. doi: 10.1016/0271-5309(84)90019-3

Boberg, C. (2008). English in Canada: Phonology. In E. Schneider (Ed.), *Varieties of English 2: the Americas and the Caribbean* (pp. 144–160). Berlin, Germany: De Gruyter Mouton.

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer.* Retrieved June 30, 2023 from https://www.fon.hum.uva.nl/praat/.

Bohn, A., & Berntsen, D. (2007). Pleasantness bias in flashbulb memories: Positive and negative flashbulb memories of the fall of the Berlin Wall among East and West Germans. *Memory & Cognition*, *35*(3), 565–577. doi: 10.3758/bf03193295

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3), 255–272. doi: 10.1016/S0167-6393(96)00063-5

Brunellière, A., Dufour, S., Nguyen, N., & Frauenfelder, U. H. (2009). Behavioral and electrophysiological evidence for the impact of regional variation on phoneme perception. *Cognition*, *111*(3), 390–396. doi: 10.1016/j.cognition.2009.02.013

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. doi: 10.3758/BRM.41.4.977

Burbano-Elizondo, L. (2015). Sunderland. In R. Hickey (Ed.), *Researching Northern English* (pp. 183–204). Amsterdam, Netherlands: John Benjamins.

Bybee, J. (2017). Grammatical and lexical factors in sound change: A usage-based approach. *Language Variation and Change*, *29*(3), 273–300. doi: 10.1017/S0954394517000199

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology*, *98*, 73–101. doi: 10.1016/j.cogpsych.2017.08.003

Centers for Disease Control and Prevention. (2022). *What noises cause hearing loss?* Retrieved June 30, 2023 from https://www.cdc.gov/nceh/hearing_loss/what_noises_cause_hearing_loss.html.

Cherry, E. C. (1953). Some experiment on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. doi: 10.1121/1.1907229

Christensen, R. H. B. (2022). *Ordinal: Regression models for ordinal data.* Retrieved June 30, 2023 from https://cran.r-project.org/web/packages/ordinal/index.html.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. doi: 10.1121/1.1815131

Clopper, C. G. (2017). Dialect interference in lexical processing: Effects of familiarity and social stereotypes. *Phonetica*, *74*(1), 25–59. doi: 10.1159/000446809

Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: Intelligibility and classification. *Language and Speech*, *51*(3), 175–198. doi: 10.1177/0023830908098539

Clopper, C. G., Tamati, T. N., & Pierrehumbert, J. B. (2016). Variation in the strength of lexical encoding across dialects. *Journal of Phonetics*, *58*, 87–103. doi: 10.1016/j.wocn.2016.06.002

Cohen, A. O., Dellarco, D. V., Breiner, K., Helion, C., Heller, A. S., Rahdar, A., ... Casey, B. (2016). The impact of emotional states on cognitive control circuitry and function. *Journal of Cognitive Neuroscience*, *28*(3), 446–459. doi: 10.1162/jocn_a_00906

Coupland, N., & Bishop, H. (2007). Ideologised values for British accents. *Journal of Sociolinguistics*, *11*(1), 74–93. doi: 10.1111/j.1467-9841.2007.00311.x

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? In W. S. Sossin, J.-C. Lacaille, V. F. Castellucci, & S. Belleville (Eds.), *Progress in brain research* (Vol. 169, pp. 323–338). Amsterdam, Netherlands: Elsevier.

Daidone, D. (2017). *Praat script: Concatenation of groups of three sound files with pauses.* Retrieved June 30, 2023 from http://www.ddaidone.com/uploads/1/0/5/2/105292729/concatenate_sound_triads_from_table.txt.

Derwing, T. M., & Munro, M. J. (1997). ACCENT, INTELLIGIBILITY, AND COMPREHENSIBILITY: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1–16. doi: 10.1017/S0272263197001010

Derwing, T. M., Rossiter, M. J., & Munro, M. J. (2002). Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingual and Multicultural Development*, *23*(4), 245–259. doi: 10.1080/01434630208666468

Dewaele, J.-M., & McCloskey, J. (2015). Attitudes towards foreign accents among adult multilingual language users. *Journal of Multilingual and Multicultural Development*, *36*(3), 221–238. doi: 0.1080/01434632.2014.909445

DiCanio, C. (2011). *Praat script: Renaming of intervals.* Retrieved June 30, 2023 from http://www.acsu.buffalo.edu/~cdicanio/scripts/Replace_labels.praat.

Docherty, G. J., & Foulkes, P. (2014). An evaluation of usage-based approaches to the modelling of sociophonetic variability. *Lingua*, *142*, 42–56. doi: 10.1016/j.lingua.2013.01.011

Dolcos, F., & McCarthy, G. (2006). Brain systems mediating cognitive interference by emotional distraction. *Journal of Neuroscience*, *26*(7), 2072–2079. doi: 10.1523/JNEUROSCI.5042-05.2006

Drager, K., & Kirtley, M. J. (2016). Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. In A. M. Babel (Ed.), *Awareness and control in sociolinguistic research* (pp. 1–24). Cambridge, UK: Cambridge University Press.

Dragojevic, M. (2020). Extending the fluency principle: Factors that increase listeners' processing fluency positively bias their language attitudes. *Communication Monographs*, *87*(2), 158–178. doi: 10.1080/03637751.2019.1663543

Dragojevic, M., Giles, H., Beck, A.-C., & Tatum, N. T. (2017). The fluency principle: Why foreign accent strength negatively biases language attitudes. *Communication Monographs*, *84*(3), 385–405. doi: 10.1080/03637751.2017.1322213

Edwards, J. G. H., Zampini, M. L., & Cunningham, C. (2018). The accentedness, comprehensibility, and intelligibility of Asian Englishes. *World Englishes*, *37*(4), 538–557. doi: 10.1111/weng.12344

Evans, B. G., & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *The Journal of the Acoustical Society of America*, *121*(6), 3814–3826. doi: 10.1121/1.2722209

Fagyal, Z., Hassa, S., & Ngom, F. (2002). L'opposition [e]–[ɛ] en syllabes ouvertes de fin de mot en français parisien: Étude acoustique préliminaire. In *XXIVèmes journées d'Étude sur la parole, 24-27 juin 2002* (pp. 165–168). Nancy, France: Nantes Université.

Fagyal, Z., Kibbee, D., & Jenkins, F. (2006). *French: A linguistic introduction*. Cambridge, UK: Cambridge University Press.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. doi: 10.1111/1467-8721.0015

Figueira, J. S. B., Oliveira, L., Pereira, M. G., Pacheco, L. B., Lobo, I., Motta-Ribeiro, G. C., & David, I. A. (2017). An unpleasant emotional state reduces working memory capacity: electrophysiological evidence. *Social Cognitive and Affective Neuroscience*, *12*(6), 984–992. doi: 10.1093/scan/nsx030

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: A unified javascript framework for online studies. In *2017 International Conference on Computational Social Science IC^{2}S^2, July 10-13, 2016* (pp. 1–3). Cologne, Germany.

Floccia, C., Butler, J., Goslin, J., & Ellis, L. (2009). Regional and foreign accent processing in english: Can listeners adapt? *Journal of Psycholinguistic Research*, *38*(4), 379–412. doi: 10.1007/s10936-008-9097-8

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology. Human Perception and Performance*, *32*(5), 1276–1293. doi: 10.1037/0096-1523.32.5.1276

Foulkes, P., & Docherty, G. (1999). Derby and Newcastle: Instrumental phonetics and variationist studies. In P. Foulkes & G. Docherty (Eds.), *Urban voices: Accent studies in the British Isles.* Milton Park, UK: Routledge.

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, *34*(4), 409–438. doi: 10.1016/j.wocn.2005.08.002

Frances, C., Costa, A., & Baus, C. (2018). On the effects of regional accents on memory and credibility. *Acta Psychologica*, *186*, 63–70. doi: 10.1016/j.actpsy.2018.04.003

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125. doi: 10.1037//0096-1523.6.1.110

Gasper, K., & Clore, G. L. (2002). Attending to the big picture: Mood and global versus local processing of visual information. *Psychological Science*, *13*(1), 34–40. doi: 10.1111/1467-9280.00406

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. doi: 10.1177/1745691614551642

Giles, H. (1970). Evaluative reactions to accents. *Educational Review*, *22*(3), 211–227. doi: 10.1080/0013191700220301

Giles, H., & Billings, A. C. (2004). Assessing language attitudes: Speaker evaluation studies. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 187–209). Hoboken, NJ: John Wiley & Sons.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183. doi: 10.1037//0278-7393.22.5.1166

Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language*, *122*(2), 92–102. doi: 10.1016/j.bandl.2012.04.017

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, *29*(2), 228–244. doi: 10.1016/0749-596X(90)90074-A

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. doi: 10.1037//0022-3514.74.6.1464

Grohe, A.-K., & Weber, A. (2018). Memory advantage for produced words and familiar native accents. *Journal of Cognitive Psychology*, *30*(5-6), 570–587. doi: 10.1080/20445911.2018.1499659

Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, *24*(4), 878–887. doi: 10.1162/jocn_a_00103

Hanulíková, A., & Weber, A. (2010). *Production of English interdental fricatives by Dutch, German, and English speakers.* Retrieved June 30, 2023 from https://www.researchgate.net/publication/50809673_Production_of_English_interdental_fricatives_by_Dutch_German_and_English_speakers.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*(4), 865–892. doi: 10.1515/LING.2010.027

Hay, J., & Maclagan, M. A. (2008). *New Zealand English*. Edinburgh, UK: Edinburgh University Press.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484. doi: 10.1016/j.wocn.2005.10.001

Hogg, M. A., & Vaughan, G. M. (2018). *Social psychology* (8th ed.). London, UK: Pearson Education.

Hällgren, M., Larsby, B., Lyxell, B., & Arlinger, S. (2001). Evaluation of a cognitive test battery in young and elderly normal-hearing and hearing-impaired persons. *Journal of the American Academy of Audiology*, *12*(7), 357–370. doi: 10.1055/s-0042-1745620

Impe, L., Geeraerts, D., & Speelman, D. (2009). Mutual intelligibility of standard and regional dutch language varieties. In J. Nerbonne, C. Gooskens, S. Kürschner, & R. Van Bezooijen (Eds.), *International journal of humanities and arts computing Volume 2* (pp. 101–118). Edinburgh, UK: Edinburgh University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. doi: 10.1371/journal.pmed.0020124

Isen, A. M., & Daubman, K. A. (1984). The influence of affect on categorization. *Journal of Personality and Social Psychology*, *47*(6), 1206–1217. doi: 10.1037/0022-3514.47.6.1206

ITV News. (2013). *28% feel accent discrimination.* Retrieved June 30, 2023 https://www.itv.com/news/story/2013-09-25/regional-accents-discrimination-friendliest/.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.

Johnson, K. (2008). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (1st ed., pp. 363–389). Hoboken, NJ: John Wiley & Sons.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384. doi: 10.1006/jpho.1999.0100

Kazanina, N., Bowers, J. S., & Idsardi, W. (2018). Phonemes: Lexical access and beyond. *Psychonomic Bulletin & Review*, *25*(2), 560–585. doi: 10.3758/s13423-017-1362-0

Kensinger, E. A. (2009). Remembering the details: Effects of emotion. *Emotion Review*, *1*(2), 99–113. doi: 10.1177/1754073908100432

Kensinger, E. A., & Schacter, D. L. (2006). Amygdala activity is associated with the successful encoding of item, but not source, information for positive and negative stimuli. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *26*(9), 2564–2570. doi: 10.1523/JNEUROSCI.5241-05.2006

Kizach, J. (2014). Analyzing likert-scale data with mixed-effects linear models - a simulation study. Tübingen, Germany: Universität Tübingen. (Linguistic Evidence 2014: Empirical, theoretical and computational perspectives)

Kjellberg, A., Ljung, R., & Hallman, D. (2008). Recall of words heard in noise. *Applied Cognitive Psychology*, *22*(8), 1088–1098. doi: 10.1002/acp.1422

Knight, S. (2022). *Acoustic measures of speech perception in noise.* Personal communication.

Koops, C., Gentry, E., & Pantos, A. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics*, *14*(2), 93–101.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2007). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81. doi: 10.1016/j.cognition .2007.07.013

Kraus, J. (2017). *A sociophonetic study of the Urban Bahamian Creole vowel system* (Unpublished doctoral dissertation). Ludwig-Maximilians-Universität München.

Kraut, R., & Wulff, S. (2013). Foreign-accented speech perception ratings: A multifactorial case study. *Journal of Multilingual and Multicultural Development*, *34*(3), 249–263. doi: 10.1080/01434632.2013.767340

Kristiansen, T., Garrett, P., & Coupland, N. (2005). Introducing subjectivities in language variation and change. *Acta Linguistica Hafniensia*, *37*(1), 9–35. doi: 10.1080/03740463 .2005.10416081

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Instruction manual and affective ratings, technical report A-8.* The Centre for Research in Psychophysiology, University of Florida.

Lehiste, I., & Peterson, G. E. (1961). Transitions, glides, and diphthongs. *The Journal of the Acoustical Society of America*, *33*(3), 268–277. doi: 10.1121/1.1908638

Lennes, M. (2002). *Praat script: Saving segments as separate files.* Retrieved June 30, 2023 from http://phonetics.linguistics.ucla.edu/facilities/acoustic/save_labeled_intervals_to_wav_sound_files.txt.

Lenth, R. V., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., ... Singmann, H. (2023). *Emmeans: Estimated marginal means, aka least-squares means.* Retrieved June 30, 2023 from https://cran.r-project.org/web/packages/emmeans/index.html.

Lev-Ari, S. (2015a). Adjusting the manner of language processing to the social context: Attention allocation during interactions with non-native speakers. In R. K. Mishra, N. Srinivasan, & F. Huettig (Eds.), *Attention and vision in language processing* (pp. 185–195). New Delhi, India: Springer India.

Lev-Ari, S. (2015b). Comprehending non-native speakers: Theory and evidence for adjustment in manner of processing. *Frontiers in Psychology*, *5*, 1–12. doi: doi.org/10.3389/fpsyg.2014.01546

Lev-Ari, S., Dodsworth, R., Mielke, J., & Peperkamp, S. (2019). The different roles of expectations in phonetic and lexical processing. In G. Kubin & Z. Kačič (Eds.), *Proceedings of Interspeech 2019* (pp. 2305–2309). Graz, Austria.

Lev-Ari, S., Ho, E., & Keysar, B. (2018). The unforeseen consequences of interacting with non-native speakers. *Topics in Cognitive Science*, *10*(4), 835–849. doi: doi.org/10.1111/tops.12325

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*, 1093–1096. doi: 10.1016/j.jesp.2010.05.025

Lev-Ari, S., & Keysar, B. (2012). Less-detailed representation of non-native language: Why non-native speakers' stories seem more vague. *Discourse Processes*, *49*(7), 523–538. doi: 10.1080/0163853X.2012.698493

Levine, L., & Bluck, S. (2004). Painting with broad strokes: Happiness and the malleability of event memory. *Cognition and Emotion*, *18*(4), 559–574. doi: 10.1080/02699930341000446

Levon, E., Sharma, D., Watt, D., Perry, C., Cardoso, A., Ye, Y., & Ilbury, C. (2023). *Accent Bias Britain: Results.* Retrieved June 30, 2023 from https://accentbiasbritain.org/results-labels/.

Lindemann, S. (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics*, *7*(3), 348–364. doi: 10.1111/1467-9481.00228

Lippi-Green, R. (2011). *English with an accent: Language, ideology and discrimination in the united states* (2nd ed.). Milton Park, UK: Routledge.

Ljung, R., & Kjellberg, A. (2008). Recall of spoken words presented with a prolonged reverberation time. Foxwood, CT. (9th International Congress on Noise as a Public Health Problem (ICBEN))

Ljung, R., Sörqvist, P., Kjellberg, A., & Green, A.-M. (2009). Poor listening conditions impair memory for intelligible lectures: Implications for acoustic classroom standards. *Building Acoustics*, *16*(3), 257–265. doi: 10.1260/135101009789877031

Lobanov, B. M. (1971). Classification of russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606–608. doi: 10.1121/1.1912396

Lyberg Åhlander, V., Haake, M., Brännström, J., Schötz, S., & Sahlén, B. (2015). Does the speaker's voice quality influence children's performance on a language comprehension test? *International Journal of Speech-Language Pathology*, *17*(1), 63–73. doi: 10.3109/17549507.2014.898098

Marsh, J. E., Ljung, R., Nöstl, A., Threadgold, E., & Campbell, T. A. (2015). Failing to get the gist of what's being said: Background noise impairs higher-order cognitive processing. *Frontiers in Psychology*, *6*, 1–10. doi: 10.3389/fpsyg.2015.00548

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71–102. doi: 10.1016/0010-0277(87)90005-9

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 576–585. doi: 10.1037//0096-1523.15.3.576

Martin, C. D., Garcia, X., Potter, D., Melinger, A., & Costa, A. (2016). Holiday or vacation? The processing of variation in vocabulary across dialects. *Language, Cognition and Neuroscience*, *31*(3), 375–390. doi: 10.1080/23273798.2015.1100750

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562. doi: 10.1080/03640210802035357

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. doi: 10.1016/0010-0285(86)90015-0

McCloy, D. (2013). *Praat script: Noise masking of stimuli.* Retrieved June 30, 2023 from https://github.com/drammock/praat-semiauto/blob/master/MixSpeechNoise.praat.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. London, UK: Taylor & Francis.

Mearns, A. (2015). Tyneside. In R. Hickey (Ed.), *Researching Northern English.* Amsterdam, Netherlands: John Benjamins.

Montgomery, C. (2018). The perceptual dialectology of England. In N. Braber & S. Jansen (Eds.), *Sociolinguistics in England* (pp. 127–164). London, UK: Palgrave Macmillan UK.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. doi: 10.1111/j.1467-1770.1995.tb00963.x

Newcastle University IT Service. (2023). *Form builder.* Retrieved June 30, 2023 from https://services.ncl.ac.uk/itservice/services/collaboration/formbuilder/.

Niedenthal, P. M., & Setterlund, M. B. (1994). Emotion congruence in perception. *Personality and Social Psychology Bulletin*, *20*(4), 401–411. doi: 10.1177/0146167294204007

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*(1), 62–85. doi: doi.org/10.1177/0261927X99018001005

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, 1–24. doi: 10.7554/eLife.33468

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. doi: 10.1007/s10459-010-9222-y

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. doi: 10.1016/S0010-0285(03)00006-9

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. Pothos & A. Wills (Eds.), *Formal approaches in categorization* (pp. 18–39). Cambridge, UK: Cambridge University Press.

Nycz, J. (2015). Second dialect acquisition: A sociophonetic perspective. *Language and Linguistics Compass*, *9*(11), 469–482. doi: 10.1111/lnc3.12163

Nycz, J. (2018). Stylistic variation among mobile speakers: Using old and new regional variables to construct complex place identity. *Language Variation and Change*, *30*(2), 175–202. doi: 10.1017/S0954394518000108

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. doi: 10.3758/bf03206860

Office for National Statistics. (2021). Population projections for local authorities: Table 2 [Computer software manual]. Retrieved June 30, 2023 from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/datasets/localauthoritiesinenglandtable2.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237–241. doi: 10.1016/j.tics.2008.02.014

O'Connor, B. P. (2002). A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories. *Assessment*, *9*(2), 188–203. doi: 10.1177/1073191102092010

Pantos, A. J., & Perkins, A. W. (2012). Measuring implicit and explicit attitudes toward foreign accented speech. *Journal of Language and Social Psychology*, *32*(1), 3–20. doi: 10.1177/0261927X12463005

Pearce, M. (2009). A perceptual dialect map of North East England. *Journal of English Linguistics*, *37*(2), 162–192. doi: 10.1177/0075424209334026

Perea, M., & Rosa, E. (2002). The effects of associative and semantic priming in the lexical decision task. *Psychological Research*, *66*(3), 180–194. doi: 10.1007/s00426-002-0086-5

Pereira, M. G., Volchan, E., de Souza, G. G. L., Oliveira, L., Campagnoli, R. R., Pinheiro, W. M., & Pessoa, L. (2006). Sustained and transient modulation of performance induced by emotional picture viewing. *Emotion*, *6*(4), 622–634. doi: 10.1037/1528-3542.6.4.622

Perry, T. L., Ohde, R. N., & Ashmead, D. H. (2001). The acoustic bases for gender identification from children's voices. *The Journal of the Acoustical Society of America*, *109*(6), 2988–2998. doi: 10.1121/1.1370525

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, *2*(1), 33–52. doi: 10.1146/annurev-linguistics-030514-125050

Pinet, M., Iverson, P., & Huckvale, M. (2011). Second-language experience and speech-in-noise recognition: Effects of talker-listener accent similarity. *The Journal of the Acoustical Society of America*, *130*(3), 1653–1662. doi: 10.1121/1.3613698

Piquado, T., Benichov, J. I., Brownell, H., & Wingfield, A. (2012). The hidden effect of hearing acuity on speech recall, and compensatory effects of self-paced listening. *International Journal of Audiology*, *51*(8), 576–583. doi: 10.3109/14992027.2012.684403

Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *The Journal of the Acoustical Society of America*, *22*(6), 807–820. doi: 10.1121/1.1906694

Preston, D. R. (2013). The influence of regard on language variation and change. *Journal of Pragmatics*, *52*, 93–104. doi: 10.1016/j.pragma.2012.12.015

Prolific Team. (2023). *Prolific.* Retrieved June 30, 2023 from https://www.prolific.co/.

R Core Team. (2023). *R: A language and environment for statistical computing.* Retrieved June 30, 2023 from https://www.R-project.org/.

Rabbitt, P. (1966). Recognition: Memory for words correctly heard in noise. *Psychonomic Science*, *6*(8), 383–384. doi: 10.3758/BF03330948

Rabbitt, P. (1990). Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. *Acta Oto-Laryngologica. Supplementum*, *476*, 167–176. doi: 10.3109/00016489109127274

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339–1362. doi: doi.org/10.1080/02724980244000099

Richtsmeier, P. (2011). Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology*, *2*(1). doi: 10.1515/labphon.2011.005

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. doi: 10.1037/0278-7393.21.4.803

Rosenberg, A., & Hirschberg, J. (2021). Prosodic aspects of the attractive voice. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice attractiveness: Studies on sexy, likable, and charismatic speakers* (pp. 17–40). Berlin, Germany: Springer.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative english-speaking teaching assistants. *Research in Higher Education*, *33*(4), 511–531.

Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology*, *62*(1), 49–72. doi: 10.1146/annurev.psych.121208.131643

Schneider, B. A., Daneman, M., Murphy, D. R., & See, S. K. (2000). Listening to discourse in distracting settings: The effects of aging. *Psychology and Aging*, *15*(1), 110–125. doi: 10.1037//0882-7974.15.1.110

Scott, D. R., & Cutler, A. (1984). Segmental phonology and the perception of syntactic structure. *Journal of Verbal Learning and Verbal Behavior*, *23*(4), 450–466. doi: 10.1016/S0022-5371(84)90291-3

Sharma, D., Levon, E., & Ye, Y. (2022). 50 years of British accent bias: Stability and lifespan change in attitudes to accents. *English World-Wide*, *43*(2), 135–166. doi: 10.1075/eww.20010.sha

Simion, D. (2023). *Soundbible: Party crowd.* Retrieved June 30, 2023 from https://soundbible.com/2163-Party-Crowd.html.

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., ... Christensen, R. H. B. (2023). *Afex: Analysis of factorial experiments.* Retrieved June 30, 2023 from https://cran.r-project.org/web/packages/tidyverse/index.html.

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and linguistics compass*, *3*(1), 236–264. doi: 10.1111/j.1749-818X.2008.00112.x

Smith, E. R., Mackie, D. M., & Claypool, H. M. (2014). *Social psychology* (4th ed.). New York, NY: Psychology Press.

Stats New Zealand. (2021). *2018 census totals by topic – national highlights (updated).* Retrieved June 30, 2023 from https://www.stats.govt.nz/information-releases/2018-census-totals-by-topic-national-highlights-updated.

Stevens, K. N. (2008). Features in speech perception and lexical access. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (1st ed., pp. 125–155). Hoboken, NJ: John Wiley & Sons.

Storbeck, J. (2013). Negative affect promotes encoding of and memory for details at the expense of the gist: Affect, encoding, and false memories. *Cognition & Emotion*, *27*(5), 800–819. doi: 10.1080/02699931.2012.741060

Storbeck, J., & Clore, G. L. (2005). With sadness comes accuracy; with happiness, false memory: Mood and the false memory effect. *Psychological Science*, *16*(10), 785–791. doi: 10.1111/j.1467-9280.2005.01615.x

Stout, D. M., Shackman, A. J., & Larson, C. L. (2013). Failure to filter: Anxious individuals show inefficient gating of threat from working memory. *Frontiers in Human Neuroscience*, *7*, 1-10. doi: 10.3389/fnhum.2013.00058

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, *18*(1), 86–100. doi: 10.1177/0261927X99018001006

Strand, E. A. (2000). *Gender stereotype effects in speech processing* (Unpublished doctoral dissertation). The Ohio State University.

Strangert, E., & Gustafson, J. (2008). What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. In D. Burnham (Ed.), *Proceedings of Interspeech 2008* (pp. 1688–1698). Brisbane, Australia.

Stringer, L., & Iverson, P. (2019). Accent intelligibility differences in noise across native and nonnative accents: Effects of talker–listener pairing at acoustic–phonetic and lexical levels. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2213–2226. doi: 10.1044/2019_JSLHR-S-17-0414

Stringer, L., & Iverson, P. (2020). Non-native speech recognition sentences: A new materials set for non-native speech perception research. *Behavior Research Methods*, *52*(2), 561–571. doi: 10.3758/s13428-019-01251-z

Strongman, K. T., & Woosley, J. (1967). Stereotyped reactions to regional accents. *British Journal of Social and Clinical Psychology*, *6*(3), 164–167. doi: 10.1111/j.2044-8260.1967.tb00516.x

Sturt, P., Sanford, A. J., Stewart, A., & Dawydiak, E. (2004). Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin & Review*, *11*(5), 882–888. doi: 10.3758/bf03196716

Sumner, M. (2013). A phonetic explanation of pronunciation variant effects. *The Journal of the Acoustical Society of America*, *134*(1), EL26–EL32. doi: 10.1121/1.4807432

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2013). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology*, *4*, 1–13. doi: 10.3389/fpsyg.2013.01015

Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, *60*(4), 487–501. doi: 10.1016/j.jml.2009.01.001

Surprenant, A. M. (1999). The effect of noise on memory for spoken syllables. *International Journal of Psychology*, *34*(5), 328–333. doi: 10.1080/002075999399648

Torre, I. (2017). *The impact of voice on trust attributions* (Unpublished doctoral dissertation). University of Plymouth.

Torre, I., Goslin, J., & White, L. (2020). If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior*, *105*, 106215. doi: 10.1016/j.chb.2019.106215

Torre, I., & White, L. (2021). Trust in vocal human–robot interaction: Implications for robot voice design. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice attractiveness: Studies on sexy, likable, and charismatic speakers* (pp. 299–316). Berlin, Germany: Springer.

Traxler, M. J. (2012). *Introduction to psycholinguistics: Understanding language science*. Hoboken, NJ: Wiley-Blackwell.

Trudgill, P. (1982). *On dialect: Social and geographical perspectives*. Hoboken, NJ: Wiley-Blackwell.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., Hale, S., & Rose, N. S. (2008). Auditory-visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *International Journal of Audiology*, *47*(Suppl 2), S31–S37. doi: 10.1080/14992020802301662

Uchansky, R. M. (2008). Clear speech. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (1st ed., pp. 207–235). Hoboken, NJ: John Wiley & Sons.

Van Berkum, J. J. A. V., Brink, D. V. D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580–591. doi: 10.1162/jocn.2008.20054

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190. doi: 10.1080/17470218.2013.8505

Vaughn, C. R. (2019). Expectations about the source of a speaker's accent affect accent adaptation. *The Journal of the Acoustical Society of America*, *145*(5), 3218–3232. doi: 10.1121/1.5108831

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *The Journal of the Acoustical Society of America*, *60*(1), 198–212. doi: 10.1121/1.381065

Verde, M. F. (2021). Retrieval-induced forgetting and inhibition: A critical review. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 47–80). Cambridge, MA: Academic Press.

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751. doi: 10.1038/nature02447

Walker, A., & Hay, J. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology*, *2*(1), 219–237. doi: 10.1515/labphon.2011.007

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, *57*(8), 1124–1133. doi: 10.3758/bf03208369

Wang, X., & Xu, L. (2021). Speech perception in noise: Masking and unmasking. *Journal of Otology*, *16*(2), 109–119. doi: 10.1016/j.joto.2020.12.001

Warburton, J. (2020). *The merging of the GOAT and THOUGHT vowels in Tyneside English: Evidence from production and perception* (Unpublished doctoral dissertation). Newcastle University.

Ward, C. M., Rogers, C. S., Van Engen, K. J., & Peelle, J. E. (2016). Effects of age, acoustic challenge, and verbal working memory on recall of narrative speech. *Experimental aging research*, *42*(1), 97–111. doi: 10.1080/0361073X.2016.1108785

Warren, A., & Gibson, J. (2023). *Vocoder: Introduction to MIDI and computer music.* Retrieved June 30, 2023 from https://cecm.indiana.edu/361/rsn-vocoder.html.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393. doi: 10.1126/science.167.3917.392

Wasiuk, P. A., Radvansky, G. A., Greene, R. L., & Calandruccio, L. (2021). Spoken narrative comprehension for young adult listeners: effects of competing voices and noise. *International Journal of Audiology*, *60*(9), 711–722. doi: 10.1080/14992027.2021.1878397

Watt, D. J. L. (1998). *Variation and change in the vowel system of Tyneside English* (Unpublished doctoral dissertation). Newcastle University.

Watt, D. J. L., & Allen, W. (2003). Tyneside English. *Journal of the International Phonetic Association*, *33*(2), 267–271. doi: 10.1017/S0025100303001397

Watt, D. J. L., & Fabricius, A. (2003). Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1 - F2 plane. *Leeds Working Papers in Linguistics and Phonetics*, *9*(9), 159–173.

Watt, D. J. L., & Milroy, L. (1999). Patterns of variation and change in three Newcastle vowels: Is this dialect levelling? In P. Foulkes & G. J. Docherty (Eds.), *Urban voices: Accent studies in the British Isles* (pp. 25–46). London, UK: Arnold.

Weiss, B., & Burkhardt, F. (2010). Voice attributes affecting likability perception. In K. Hirose (Ed.), *Proceedings of Interspeech 2010* (pp. 2014–2017). Makuhari, Japan.

Weiss, B., & Burkhardt, F. (2012). Is 'not bad' good enough? Aspects of unknown voices' likability. In R. Sproat (Ed.), *Proceedings of Interspeech 2012* (pp. 510–513). Portland, OR.

Wells, J. C. (1982). *Accents of English: Volume 1*. Cambridge, United Kingdom: Cambridge University Press.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*(2), 143–154. doi: 10.1016/S0022-5371(78)90110-X

Wickham, H., & RStudio. (2023). *Tidyverse: Easily install and load the 'tidyverse'.* Retrieved June 30, 2023 from https://tidyverse.tidyverse.org/.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Milton Park, UK: Routledge.

Woods, K. J., Siegel, M., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception & Psychophysics*, *79*(7), 2064–2072. doi: 10.3758/s13414-017-1361-2

Xie, W., Ye, C., & Zhang, W. (2022). Negative emotion reduces visual working memory recall variability: A meta-analytical review. *Emotion*, *23*(3), 859-871. doi: 10.1037/emo0001139

Zehr, J., & Schwarz, F. (2018). *PennController for internet based experiments (IBEX)*. Retrieved June 30, 2023 from https://www.pcibex.net/. doi: 10.17605/OSF.IO/MD832

Zsiga, E. C. (2022). *Phonetics/phonology interface*. Edinburgh, UK: Edinburgh University Press.

Appendices

Appendix A

Participant Information and Consent

The information sheets, declarations of informed consent and post-experiment debriefs were presented to the speakers and participants in different formats and across different digital platforms. Therefore, the formatting of these documents in the following might be slightly different from the one that the speakers and participants saw.

A.1 Participant Information Sheets

A.1.1 Recordings 2020

(1) Your recordings will be used for linguistic research on speech processing.

(2) In particular, we are interested in how fast and accurately accents are processed based on how familiar listeners are with them, how intelligible they find them and what attitude they have towards them.

(3) The current research is conducted by Andreas Krug as part of his PhD studies at Newcastle University.

(4) This research is supervised by Prof. Ghada Khattab and Dr Laurence White from the School of Education, Communication and Language Sciences at Newcastle University.

(5) The recording session will take about 60/90 minutes with intermittent breaks.

(6) Your recordings will be used for research purposes only and your information will be kept secure at all times.

(7) If you have any questions, regards or concerns regarding this research, please contact us via email at A.Krug2@newcastle.ac.uk or by telephone at +44 (0)789 568 2137.

This study has been approved by the Research Ethics Committee of the Faculty of Humanities and Social Sciences at Newcastle University.

A.1.2 Recordings 2022

(1) Your recordings will be used for linguistic research on speech processing.

(2) In particular, we are interested in how fast and accurately accents are processed based on how familiar listeners are with them, how intelligible they find them and what attitude they have towards them.

(3) The current research is conducted by Andreas Krug as part of his PhD studies at Newcastle University.

(4) This research is supervised by Prof. Ghada Khattab and Dr Laurence White from the School of Education, Communication and Language Sciences at Newcastle University.

(5) The recording session will take about 45 minutes with intermittent breaks.

(6) Your recordings will be used for research purposes only and your information will be kept secure at all times.

(7) If you have any questions, regards or concerns regarding this research, please contact us via email at A.Krug2@newcastle.ac.uk or by telephone at +44 (0)789 568 2137.

(8) You can reach campus security by telephone at +44 191 208 6817.

This study has been reviewed and approved by the Research Ethics Committee of the Faculty of Humanities and Social Sciences at Newcastle University.

A.1.3 Experiment 1: Pilot Study

(1) You are invited to take part in a research study entitled 'Speech Processing'.

(2) Please read the following text carefully before agreeing to be in the study.

(3) This study is conducted by Andreas Krug as part of his PhD studies at Newcastle University.

(4) This research is supervised by Prof. Ghada Khattab and Dr Laurence White from the School of Education, Communication and Language Sciences at Newcastle University.

(5) The purpose of this study is to research speech processing. We aim to investigate how quickly and accurately you can process the information in sentences and short stories.

(6) You must be wearing headphones for this experiment. Doing the experiment without headphones will lead to exclusion from the experiment.

(7) If you agree to be in this study, you will be asked to do several tasks: a decision task, a memory task, a matching task and a transcription task. You will also be asked to do a headphone check and to provide some demographic information.

(8) Your participation in this study should take no longer than 45 minutes.

(9) There will be some tasks to check that you pay attention throughout the experiment.

(10) You are free to decide whether or not to participate. If you decide to participate, you are free to withdraw at any time during the experiment without any negative consequences for you.

(11) All responses you give or other data collected will be kept confidential. The records of this study will be kept secure and private. All files containing any information you give are password protected. In any research report that may be published, no information will be included that will make it possible to identify you individually. There will be no way to connect your name to your responses at any time during or after the study.

(12) You will receive an Amazon voucher worth £7.50/two SONA credits for your participation/£7.50 for your participation via Prolific.

(13) An anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.

(14) After you finish the experiment, you can decide if you would like to withdraw your data from the experiment. You will not be able to do so afterwards and a fully anonymous version of your data will be included in the dataset.

(15) If you have any questions, regards or concerns regarding this research, please contact me via email at A.Krug2@newcastle.ac.uk or by telephone at 0789 568 2137.

This study has been reviewed and approved by the Research Ethics Committee of the Faculty of Humanities and Social Sciences at Newcastle University.

A.1.4 Experiment 1: Main Study

(1) You are invited to take part in a research study entitled 'Speech Processing'.

(2) Please read the following text carefully before agreeing to be in the study.

(3) This study is conducted by Andreas Krug as part of his PhD studies at Newcastle University.

(4) This research is supervised by Prof. Ghada Khattab and Dr Laurence White from the School of Education, Communication and Language Sciences at Newcastle University.

(5) The purpose of this study is to research speech processing. We aim to investigate how quickly and accurately you can process the information in sentences and short stories.

(6) You must be wearing headphones for this experiment. Doing the experiment without headphones will lead to exclusion from the experiment.

(7) If you agree to be in this study, you will be asked to do several tasks: a decision task, a memory task, a matching task and a transcription task. You will also be asked to do a headphone check and to provide some demographic information.

(8) Your participation in this study should take no longer than 45 minutes.

(9) There will be some tasks to check that you pay attention throughout the experiment.

(10) You are free to decide whether or not to participate. If you decide to participate, you are free to withdraw at any time during the experiment without any negative consequences for you.

(11) All responses you give or other data collected will be kept confidential. The records of this study will be kept secure and private. All files containing any information you give are password protected. In any research report that may be published, no information will be included that will make it possible to identify you individually. There will be no way to connect your name to your responses at any time during or after the study.

(12) You will receive an Amazon voucher worth £7.50/£7.50 for your participation via Prolific.

(13) An anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.

(14) After you finish the experiment, you can decide if you would like to withdraw your data from the experiment. You will not be able to do so afterwards and a fully anonymous version of your data will be included in the dataset.

(15) If you have any questions, regards or concerns regarding this research, please contact me via email at A.Krug2@newcastle.ac.uk or by telephone at 0789 568 2137.

This study has been reviewed and approved by the Research Ethics Committee of the Faculty of Humanities and Social Sciences at Newcastle University.

A.1.5 Experiment 2: Pilot Studies

(1) You are invited to take part in a research study entitled 'Transcription Study'.

(2) Please read the following text carefully before agreeing to be in the study.

(3) This study is conducted by Andreas Krug as part of his PhD studies at Newcastle University.

(4) This research is supervised by Prof. Ghada Khattab and Dr Laurence White from the School of Education, Communication and Language Sciences at Newcastle University.

(5) The purpose of this study is to research transcription performance. We aim to investigate how accurately you can transcribe short sentences.

(6) You must be wearing headphones for this experiment. Doing the experiment without headphones will lead to exclusion from the experiment.

(7) If you agree to be in this study, you will be asked to do two tasks: a headphone check and a transcription task. You will also be asked to provide some demographic information.

(8) Your participation in this study should take no longer than 18 minutes.

(9) There will be a task to check that you pay attention throughout the experiment.

(10) You are free to decide whether or not to participate. If you decide to participate, you are free to withdraw at any time during the experiment without any negative consequences for you.

(11) All responses you give or other data collected will be kept confidential. The records of this study will be kept secure and private. All files containing any information you give are password protected. In any research report that may be published, no information will be included that will make it possible to identify you individually. There will be no way to connect your name to your responses at any time during or after the study.

(12) An anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.

(13) After you finish the experiment, you can decide if you would like to withdraw your data from the experiment. You will not be able to do so afterwards and a fully anonymous version of your data will be included in the dataset.

(14) You have to complete this experiment in full-screen mode. If you exit full-screen mode (via the F11 key), the experiment will be paused and you cannot continue until you re-enter full-screen mode.

(15) You will receive £3.00 for your participation via Prolific/a voucher worth NZ$10.00 for your participation.

This study has been reviewed and approved by the Research Ethics Committee of the Faculty of Humanities and Social Sciences at Newcastle University.

A.1.6 Experiment 2: Main Study

(1) You are invited to take part in a research study entitled 'Speech Processing'.

(2) Please read the following text carefully before agreeing to be in the study.

(3) This study is conducted by Andreas Krug as part of his PhD studies at Newcastle University.

(4) This research is supervised by Prof. Ghada Khattab and Dr Laurence White from the School of Education, Communication and Language Sciences at Newcastle University.

(5) The purpose of this study is to research speech processing. We aim to investigate how quickly and accurately you can process the information in sentences and short stories when there is background noise.

(6) You must be wearing headphones for this experiment. Doing the experiment without headphones will lead to exclusion from the experiment.

(7) If you agree to be in this study, you will be asked to do several tasks: a decision task, a memory task and a matching task. You will also be asked to do a headphone check and to provide some demographic information.

(8) Your participation in this study should take no longer than 45 minutes.

(9) You will receive an Amazon or a Waterstones voucher worth £7.50/two SONA credits/Prezzee voucher worth NZ$15.00 for your participation.

(10) There will be some tasks to check that you pay attention throughout the experiment.

(11) You are free to decide whether or not to participate. If you decide to participate, you are free to withdraw at any time during the experiment without any negative consequences for you.

(12) All responses you give or other data collected will be kept confidential. The records of this study will be kept secure and private. All files containing any information you give are password protected. In any research report that may be published, no information will be included that will make it possible to identify you individually. There will be no way to connect your name to your responses at any time during or after the study.

(13) An anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.

(14) After you finish the experiment, you can decide if you would like to withdraw your data from the experiment. You will not be able to do so afterwards and a fully anonymous version of your data will be included in the dataset.

(15) You have to complete this experiment in full-screen mode. If you exit full-screen mode (via the F11 key), the experiment will be paused and you cannot continue until you re-enter full-screen mode.

This study has been reviewed and approved by the Research Ethics Committee of the Faculty of Humanities and Social Sciences at Newcastle University.

A.1.7 Experiment 3: Pilot Study

(1) You are invited to take part in a research study entitled 'Voice Perception'.

(2) Please read the following text carefully before agreeing to be in the study.

(3) This study is conducted by Andreas Krug as part of his PhD studies at Newcastle University.

(4) This research is supervised by Prof. Ghada Khattab and Dr Laurence White from the School of Education, Communication and Language Sciences at Newcastle University.

(5) The purpose of this study is to research voice perception. We aim to investigate what you think about people following a number of tasks.

(6) You must be wearing headphones for this experiment. Doing the experiment without headphones will lead to exclusion from the experiment.

(7) If you agree to be in this study, you will be asked to do three tasks: a judgement task, an investment task and a rating task. You will also be asked to do a headphone check and to provide some demographic information.

(8) Your participation in this study should take no longer than 20 minutes.

(9) You will receive £3.50 in Prolific credits for your participation.

(10) There will be some tasks to check that you pay attention throughout the experiment.

(11) You are free to decide whether or not to participate. If you decide to participate, you are free to withdraw at any time during the experiment without any negative consequences for you. However, you will only receive the full Prolific credits if you finish the study.

(12) All responses you give or other data collected will be kept confidential. The records of this study will be kept secure and private. All files containing any information you give are password protected. In any research report that may be published, no information will be included that will make it possible to identify you individually. There will be no way to connect your name to your responses at any time during or after the study.

(13) An anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.

(14) After you finish the experiment, you can decide if you would like to withdraw your data from the experiment. You will not be able to do so afterwards and a fully anonymous version of your data will be included in the dataset.

(15) You have to complete this experiment in full-screen mode. If you exit full-screen mode (via the F11 key), the experiment will be paused and you cannot continue until you re-enter full-screen mode.

This study has been reviewed and approved by the Research Ethics Committee of the Faculty of Humanities and Social Sciences at Newcastle University.

A.1.8 Experiment 3: Main Study

(1) You are invited to take part in a research study entitled 'Speech Processing'.

(2) Please read the following text carefully before agreeing to be in the study.

(3) This study is conducted by Andreas Krug as part of his PhD studies at Newcastle University.

(4) This research is supervised by Prof. Ghada Khattab and Dr Laurence White from the School of Education, Communication and Language Sciences at Newcastle University.

(5) The purpose of this study is to research research speech processing. We aim to investigate how quickly and accurately you can process the information in sentences and short stories.

(6) You must be wearing headphones for this experiment. Doing the experiment without headphones will lead to exclusion from the experiment.

(7) If you agree to be in this study, you will be asked to do several tasks: a judgement task, an investment task, a decision task and a memory task. You will also be asked to do a headphone check and to provide some demographic information.

(8) Your participation in this study should take no longer than 45 minutes.

(9) You will receive two SONA credits/£7.50 in Prolific credits for your participation.

(10) There will be some tasks to check that you pay attention throughout the experiment.

(11) You are free to decide whether or not to participate. If you decide to participate, you are free to withdraw at any time during the experiment without any negative consequences for you. However, you will only receive the full SONA credits if you finish the study.

(12) All responses you give or other data collected will be kept confidential. The records of this study will be kept secure and private. All files containing any information you give are password protected. In any research report that may be published, no information will be included that will make it possible to identify you individually. There will be no way to connect your name to your responses at any time during or after the study.

(13) An anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.

(14) After you finish the experiment, you can decide if you would like to withdraw your data from the experiment. You will not be able to do so afterwards and a fully anonymous version of your data will be included in the dataset.

(15) You have to complete this experiment in full-screen mode. If you exit full-screen mode (via the F11 key), the experiment will be paused and you cannot continue until you re-enter full-screen mode.

This study has been reviewed and approved by the Research Ethics Committee of the Faculty of Humanities and Social Sciences at Newcastle University.

A.2 Declarations of Informed Consent

A.2.1 Recordings 2020

- I agree to make recordings for linguistic research on speech processing.
- I have read the participant information sheet and understand the information provided.
- I have been informed that I can contact the researcher via email at A.Krug2@newcastle.ac.uk or via telephone at +44 789 2137 if I have any questions.
- I will be provided with a copy of this form for my records.

A.2.2 Recordings 2022

- I agree to make recordings for linguistic research on speech processing.
- I have read the participant information sheet and understand the information provided.
- I have been informed that I can contact the researcher via email at A.Krug2@newcastle.ac.uk or via telephone at +44 789 568 2137 if I have any questions.
- I have been informed that I can reach campus security by telephone at +44 191 208 6817.
- I will be provided with a copy of this form for my records.

A.2.3 Experiment 1: Pilot Study

- I agree to take part in this study, the purpose of which is to investigate speech processing.
- I have read and understood the information provided on the previous pages.
- I have been informed that I may withdraw from the study without penalty of any kind.
- I have been informed that all my responses will be kept confidential and secure, and that I will not be identified in any report or other publication resulting from this research.
- I have been informed that an anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.
- I have been informed that I must be wearing headphones and that I will be excluded from the experiment if I do not.
- I have been informed that, once I provide my final consent at the end of the experiment, my anonymised data cannot be excluded from the dataset.
- I have been informed that the investigator will answer any questions regarding the study and its procedures. The investigator's email is A.Krug2@newcastle.ac.uk. He can be contacted via email or by telephone at 0789 568 2137.
- Please feel free to take a screenshot of this page for your own records.

A.2.4 Experiment 1: Main Study

- I agree to take part in this study, the purpose of which is to investigate speech processing.
- I have read and understood the information provided on the previous pages.
- I have been informed that I may withdraw from the study without penalty of any kind.
- I have been informed that all my responses will be kept confidential and secure, and that I will not be identified in any report or other publication resulting from this research.
- I have been informed that an anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.
- I have been informed that I must be wearing headphones and that I will be excluded from the experiment if I do not.
- I have been informed that, once I provide my final consent at the end of the experiment, my anonymised data cannot be excluded from the dataset.
- I have been informed that the investigator will answer any questions regarding the study and its procedures. The investigator's email is A.Krug2@newcastle.ac.uk. He can be contacted via email or by telephone at 0789 568 2137.
- Please feel free to take a screenshot of this page for your own records.

A.2.5 Experiment 2: Pilot Studies

- I agree to take part in this study, the purpose of which is to investigate transcription performance.
- I have read and understood the information provided on the previous pages.
- I have been informed that I may withdraw from the study without penalty of any kind.
- I have been informed that all my responses will be kept confidential and secure, and that I will not be identified in any report or other publication resulting from this research.
- I have been informed that an anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.
- I have been informed that I must be wearing headphones and that I will be excluded from the experiment if I do not.

- I have been informed that, once I provide my final consent at the end of the experiment, my anonymised data cannot be excluded from the dataset.
- I have been informed that the investigator will answer any questions regarding the study and its procedures. The investigator's email is A.Krug2@newcastle.ac.uk. He can be contacted via email or by telephone on 0789 568 2137.
- Please feel free to take a screenshot of this page for your own records.

A.2.6 Experiment 2: Main Study

- I agree to take part in this study, the purpose of which is to investigate speech processing.
- I have read and understood the information provided on the previous pages.
- I have been informed that I may withdraw from the study without penalty of any kind.
- I have been informed that all my responses will be kept confidential and secure, and that I will not be identified in any report or other publication resulting from this research.
- I have been informed that an anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.
- I have been informed that I must be wearing headphones and that I will be excluded from the experiment if I do not.
- I have been informed that, once I provide my final consent at the end of the experiment, my anonymised data cannot be excluded from the dataset.
- I have been informed that the investigator will answer any questions regarding the study and its procedures. The investigator's email is A.Krug2@newcastle.ac.uk. He can be contacted via email or by telephone on 0789 568 2137.
- Please feel free to take a screenshot of this page for your own records.

A.2.7 Experiment 3: Pilot Study

- I agree to take part in this study, the purpose of which is to investigate voice perception.
- I have read and understood the information provided on the previous pages.
- I have been informed that I may withdraw from the study without penalty of any kind.

- I have been informed that all my responses will be kept confidential and secure, and that I will not be identified in any report or other publication resulting from this research.
- I have been informed that an anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.
- I have been informed that I must be wearing headphones and that I will be excluded from the experiment if I do not.
- I have been informed that, once I provide my final consent at the end of the experiment, my anonymised data cannot be excluded from the dataset.
- I have been informed that the investigator will answer any questions regarding the study and its procedures. The investigator's email is A.Krug2@newcastle.ac.uk. He can be contacted via email, via Prolific or by telephone on 0789 568 2137.
- Please feel free to take a screenshot of this page for your own records.

A.2.8 Experiment 3: Main Study

- I agree to take part in this study, the purpose of which is to investigate speech processing.
- I have read and understood the information provided on the previous pages.
- I have been informed that I may withdraw from the study without penalty of any kind.
- I have been informed that all my responses will be kept confidential and secure, and that I will not be identified in any report or other publication resulting from this research.
- I have been informed that an anonymised version of the data will be uploaded to a secure server so that it can be accessed by other researchers.
- I have been informed that I must be wearing headphones and that I will be excluded from the experiment if I do not.
- I have been informed that, once I provide my final consent at the end of the experiment, my anonymised data cannot be excluded from the dataset.
- I have been informed that the investigator will answer any questions regarding the study and its procedures. The investigator's email is A.Krug2@newcastle.ac.uk. He can be contacted via email, via Prolific or by telephone on 0789 568 2137.
- Please feel free to take a screenshot of this page for your own records.

A.3 Participants Debriefs

A.3.1 Experiment 1: Pilot Study

Thank you very much for taking part in this study. Your time and effort are greatly appreciated.

At the beginning of the study, you were informed that this study investigates speech processing. In particular, we were interested in how fast you can identify a real or nonsense word and how many changes you detected in the short stories.

This gives us an idea of how deeply you processed the audio. We are interested in how this level of processing is related to how familiar you are with the accent in the audio and how easily you can understand it.

Since you now know the details of the research, you may decide that you do not want us to use your answers.

A.3.2 Experiment 1: Main Study

Thank you very much for taking part in this study. Your time and effort are greatly appreciated.

At the beginning of the study, you were informed that this study investigates speech processing. In particular, we were interested in how fast you can identify a real or nonsense word and how many changes you detected in the short stories.

This gives us an idea of how deeply you processed the audio. We are interested in how this level of processing is related to how familiar you are with the accent in the audio and how easily you can understand it.

Since you now know the details of the research, you may decide that you do not want us to use your answers.

A.3.3 Experiment 2: Pilot Studies

Thank you very much for taking part in this study. Your time and effort are greatly appreciated.

At the beginning of the study, you were informed that this study investigates transcription performance. In particular, we were interested in how well you can transcribe different accents in noise.

This gives us an idea of the effects of noise and different accents on speech processing.

Since you now know the details of the research, you may decide that you do not want us to use your answers.

A.3.4 Experiment 2: Main Study

Thank you very much for taking part in this study. Your time and effort are greatly appreciated.

At the beginning of the study, you were informed that this study investigates speech processing. In particular, we were interested in how fast you can identify a real or nonsense word and how many changes you detected in the short stories.

This gives us an idea of how deeply you processed the audio. We are interested in how this level of processing is related to how familiar you are with the accent in the audio and how easily you can understand it.

Since you now know the details of the research, you may decide that you do not want us to use your answers.

A.3.5 Experiment 3: Pilot Study

Thank you very much for taking part in this study. Your time and effort are greatly appreciated.

At the beginning of the study, you were informed that this study investigates voice perception. In particular, we were interested in how Speaker A's and Speaker B's responses and behaviour influenced your final ratings.

This gives us an idea of your perception of Speaker A and Speaker B.

Since you now know the details of the research, you may decide that you do not want us to use your answers.

A.3.6 Experiment 3: Main Study

Thank you very much for taking part in this study. Your time and effort are greatly appreciated.

At the beginning of the study, you were informed that this study investigates speech processing. In particular, we were interested in how fast you can identify a real or nonsense word and how many changes you detected in the short stories, depending on whether the speakers are presented in a positive or a negative light.

This gives us an idea of how deeply you processed the audio. We are interested in how this level of processing is related to whether the speakers are presented in a positive or a negative light.

Since you now know the details of the research, you may decide that you do not want us to use your answers.

Appendix B

Stimuli from Experiments

B.1 Lexical Decision Task

For each sentence frame, there is one version of the sentence ending in a real English word and one version ending in a nonword (see subsection 3.2.2.1.1). These critical words are underlined in the following list. The nonword sentences were used as fillers in the lexical decision task.

(1) An elephant is bigger than a horse.

(2) An elephant is bigger than a trape.

(3) At the weekend we often go to the pool.

(4) At the weekend we often go to the trem.

(5) Children know a lot of clever tricks.

(6) Children know a lot of clever stask.

(7) Dishes are sometimes made with grapes.

(8) Dishes are sometimes made with ven.

(9) Every day I drink chocolate milk with my lunch.

(10) Every day I drink chocolate milk with my jipe.

(11) Every day I have a piece of fruit.

(12) Every day I have a piece of sharn.

(13) He could not find the correct key.

(14) He could not find the correct pev.

(15) I called the place to ask about a room.

(16) I called the place to ask about a yeap.

(17) I prefer ones with nice blue ink.

(18) I prefer ones with nice blue hule.

(19) I really like sandwiches that have cheese.

(20) I really like sandwiches that have yight.

(21) I really like to travel by ship.

(22) I really like to travel by yarm.

(23) I want to visit them for a week.

(24) I want to visit them for a frews.

(25) In that country they eat a lot of rice.

(26) In that country they eat a lot of teesk.

(27) It can be hard to find out the truth.

(28) It can be hard to find out the blinch.

(29) Lots of people work in that school.

(30) Lots of people work in that greck.

(31) Make sure you always use the soap.

(32) Make sure you always use the murf.

(33) Most weekends I work on a farm.

(34) Most weekends I work on a louth.

(35) My dog was taken by a thief.

(36) My dog was taken by a kib.

(37) My favourite food is called beef.

(38) My favourite food is called dran.

(39) My kids are playing with my niece.

(40) My kids are playing with my jadge.

(41) People are friendlier in the north.

(42) People are friendlier in the tib.

(43) People enjoy receiving a lot of cards.

(44) People enjoy receiving a lot of rarp.

(45) People get annoyed by their boss.

(46) People get annoyed by their yurk.

(47) People usually like their own bed.

(48) People usually like their own meev.

(49) Please be careful with that old vase.

(50) Please be careful with that old glate.

(51) Please help me with this big jar.

(52) Please help me with this big cuke.

(53) Put the cream on your skin.

(54) Put the cream on your lafe.

(55) She loves eating chips in her car.

(56) She loves eating chips in her gube.

(57) She loves reading and relaxing at the beach.

(58) She loves reading and relaxing at the brotch.

(59) Some children are really scared of bees.

(60) Some children are really scared of tase.

(61) Some kids have very big ears.

(62) Some kids have very big bup.

(63) Some people do not have nice clothes.

(64) Some people do not have nice smow.

(65) That country has many types of bird.

(66) That country has many types of plass.

(67) The athlete needs a new coach.

(68) The athlete needs a new spink.

(69) The children looked at the little calf.

(70) The children looked at the little brare.

(71) The food is better in the south.

(72) The food is better in the plail.

(73) The girl is sitting on the floor.

(74) The girl is sitting on the bleff.

(75) The lawyer has a powerful voice.

(76) The lawyer has a powerful suv.

(77) The man draws pictures of cows.

(78) The man draws pictures of flane.

(79) The man is choosing some nice meat.

(80) The man is choosing some nice tarb.

(81) The man is related to the king.

(82) The man is related to the snace.

(83) The mother made her child some delicious soup.

(84) The mother made her child some delicious scarc.

(85) The old lady has a very cute cat.

(86) The old lady has a very cute filk.

(87) The president wants to have peace.

(88) The president wants to have spabe.

(89) The students enjoy hearing simple songs.

(90) The students enjoy hearing simple fusk.

(91) The stuff is on a white plate.

(92) The stuff is on a white crench.

(93) The teacher is interested in art.

(94) The teacher is interested in nount.

(95) The vegetables are under the wooden box.

(96) The vegetables are under the wooden shirth.

(97) The woman has a really interesting nose.

(98) The woman has a really interesting drine.

(99) The woman knows a very famous chef.

(100) The woman knows a very famous yav.

(101) They are very scared of the big bull.

(102) They are very scared of the big rurk.

(103) Those people like to eat lots of fish.

(104) Those people like to eat lots of gof.

(105) When we go walking we take a tent.

(106) When we go walking we take a mawk.

(107) You can't change the action of your heart.

(108) You can't change the action of your jift.

(109) You can have lots of fun at the zoo.

(110) You can have lots of fun at the poy.

(111) You should put your tickets in the bin.

(112) You should put your tickets in the wouse.

B.2 Recall Task

There are three versions of each story. The three versions differ in a single word, which is either semantically related or semantically unrelated to the original word for the experimental stories. For filler stories, the semantic proximity of the change was not controlled to the same extent. The following list first provides the experimental stories (1-24), followed by the filler stories (25-40). The target words are underlined.

(1) After a lot of begging, Simon's parents finally gave in. They went to the shelter and adopted a kitten/puppy/snake for him. Simon was very happy and promised to take good care of it.

(2) Clara is following a healthier diet now. For breakfast, she has a mango/banana/smoothie and scrambled eggs. She says it helps her to start her day with more energy.

(3) Elena's night out in London ended badly for her. She lost her purse/handbag/phone in the night club near King's Cross. She was quite upset because she had to save up to buy it.

(4) Harry tried to change the light bulb in the living room. He had to stand on a stool/chair/ box to reach the ceiling. I was really worried that he would fall down.

(5) Henry had a bike accident two days ago. He broke his hand/arm/phone when he fell on the pavement. It's a good thing (that)[1] he was wearing a helmet.

(6) I wonder which three items I'd bring to a remote island. It would probably be my phone, lighter/matches/water and some food. I just hope I'll never actually end up on a remote island.

[1] NZE_2 inserted *that* here when producing this story.

(7) I would like to go to the museum on Saturday. They have a new exhibition on magazines/newspapers/technology from the last century. It will be interesting to see how things have changed since then.

(8) Karen is an active person with many hobbies. On sunny days, she enjoys playing football/basketball/music in the park. When it's raining, she usually reads a book at home.

(9) Leo's flat in the city centre is quite small. That's why he keeps some of his stuff in a box in the toilet/bathroom/hall. At least he makes the most out of the available space.

(10) Mary and Lisa went on a hiking tour in Scotland. The tour guide had asked them to bring waterproof jackets/coats/boots. They were glad they did because the weather was rough.

(11) My brother has accepted his first permanent job. From a very young age, he always wanted to be a doctor/nurse/teacher. However, he ditched those plans and is now a chef instead.

(12) My friend gave me a great tip for making tomato soup. He said that if you add some basil/parsley/juice, it enhances the flavour. I tried it the other day and it really works.

(13) Sam and Kate made an interesting discovery yesterday. They found a small chest full of rubies/diamonds/drugs in their attic. They had no idea where it came from or who put it there.

(14) Sarah has come up with a new idea for a drink. It contains lemon, ginger/garlic/wine and some other ingredients. She thinks it will be ideal for a hot summer day.

(15) Sometimes, working in our office can be annoying. For the past days, the printer/scanner/toilet has not been working properly. We called someone to fix it but they said they were really busy.

(16) The company presented their new survival tool. It will be made of steel/iron/plastic and it will have more gadgets. They didn't reveal the final design but it sounds promising.

(17) There has been another scandal in the Premier League. The football team released a statement on Facebook/Twitter/TV last night. I don't think that their fans will believe it though.

(18) Tom and Jay had a big fight last week. Tom broke Jay's laptop/computer/window but did not tell him about it. When Jay found out, he demanded an explanation.

(19) We couldn't decide what to have for dinner last night. We ended up ordering pizza/pasta/sushi from the nearby takeaway. It was as good as always and they even gave us free drinks.

(20) We finally had time to go to IKEA last week. We bought some new pillows/blankets/drawers and carpets for the house. It was necessary because the old ones were falling apart.

(21) We had a good time at Janet's birthday party. She was happy about the dress/skirt/book that we bought for her. She told us that was just what she had wished for.

(22) We had no idea what to do on Saturday. I suggested going down to the new arena/stadium/pub. Jenny wasn't up for it though and we stayed at home.

(23) We took a family trip to the zoo yesterday. The big crocodiles/alligators/elephants scared our youngest daughter Linda. She had never seen these animals in real life before.

(24) Yesterday I went to the new restaurant in town. I have to say that I really enjoyed the beer/cider/decoration there. The food, on the other hand, was not very good.

(25) After a two-year long trial, the criminal was convicted. He never confessed to any of the crimes. Nonetheless, he was sentenced to nine/eight/twelve years in prison.

(26) Elena bought herself a new orchid two/three/four weeks ago. To be honest, she doesn't really have green fingers. However, the flower looks really good and healthy so far.

(27) I had the chance to meet my boyfriend's mother Lisa/Karen/Laura last week. She was very hospitable and we had a good chat together. I'm really glad that we got along so well.

(28) I think my favourite breakfast food is waffles/pancakes/cereal. I could have that for breakfast every day. It wouldn't be the most healthy option though.

(29) James went/drove/walked to the shopping centre yesterday. He wanted to buy the newly released video game. However, the shop was already out of stock when he arrived.

(30) Jennifer Aniston/Lawrence/Garner is one of my favourite actresses. I really enjoy all the films that she's in. Of course, some are better than others.

(31) Jess bought some candles/flowers/pictures to decorate our flat. I was a bit sceptical (irritated)[2] because we already have a lot of clutter. They really make the place much more homely though.

(32) Joe wanted to surprise Karen for her birthday. He bought flowers and took her to their favourite pub. They had a few drinks and met their friend Ian/Kyle/Linda there.

(33) Lately, I've been swamped with work at the office. We have three new projects so there are many things to do. At least I can count/rely/depend on my hard-working team there.

(34) Martha came to the UK as an immigrant from Sweden/Japan/Bolivia. The other day she finally received her UK citizenship. This makes it much easier for her to do things here.

(35) My brother moved to Australia three years ago. He was offered a job there and took the opportunity. It's strange/weird/funny that he lives on the other side of the world now.

(36) My friend Lynne loves shopping for stationary. If she could, she would go to Paperchase every day. She bought a new set of binders/pens/pencils just yesterday.

(37) My son Troy is not a big fan of healthy food. He refuses to eat any kind of vegetable. At least we can persuade him to eat one apple/kiwi/orange each day.

(38) Pierce showed me his new house in the countryside. It has four bedrooms and is generally quite spacious. The facade is painted in his favourite colour green/red/blue.

(39) Sarah has been suffering from a rare disease for many years. Last week they told her that they found a donor organ for her. With the new kidney/liver/lung, her quality of life will improve.

[2] TE_1 struggled with the initial consonant cluster in *sceptical*. Therefore, she produced the word *irritated* instead.

(40) This report is very/highly/extremely relevant for the company's new project. They are currently finalising their first prototype. With this report, they might be able to speed up the process.

B.3 Transcription Task

B.3.1 Experiment 1

The words that were used to assess the participants' transcription accuracy are underlined.

(1) For breakfast children eat toast or cereal.

(2) He keeps his stuff in the garage.

(3) In my city we have very good weather.

(4) Some students will become a doctor.

(5) The man in the corner is the captain.

(6) The nurse doesn't want to have any children.

(7) The visitors thanked the kind driver.

(8) You must press the button on the control.

B.3.2 Experiment 2: Pilot Studies

(1) An elephant is bigger than a horse.

(2) Every day I have a piece of fruit.

(3) I called the place to ask about a room.

(4) I prefer ones with nice blue ink.

(5) I really like to travel by ship.

(6) Most weekends I work on a farm.

(7) My dog was taken by a thief.

(8) My favourite food is called beef.

(9) My kids are playing with my niece.

(10) People are friendlier in the north.

(11) Please be careful with that old vase.

(12) Please help me with this big jar.

(13) Put the cream on your skin.

(14) She loves eating chips in her car.

(15) She loves reading and relaxing at the beach.

(16) Some children are really scared of bees.

(17) That country has many types of bird.

(18) The athlete needs a new coach.

(19) The children looked at the little calf.

(20) The food is better in the south.

(21) The lawyer has a powerful voice.

(22) The man draws pictures of cows.

(23) The man is choosing some nice meat.

(24) The mother made her child some delicious soup.

(25) The old lady has a very cute cat.

(26) The students enjoy hearing simple songs.

(27) The stuff is on a white plate.

(28) The teacher is interested in art.

(29) The vegetables are under the wooden box.

(30) The woman knows a very famous chef.

(31) They are very scared of the big bull.

(32) You can have lots of fun at the zoo.

B.4 Accent Matching Task

(1) He often goes to the market to buy a camel.

(2) That special thing is called a circle.

(3) He went into the house for a minute.

(4) They gave a gift to the very lucky winner.

B.5 Investment Task

For the investment task, two types of sentences were recorded. During the first round of the task, the participants heard an introductory sentence from the speaker (1-2). During the remaining trials, they heard sentences that were supposed to encourage them to invest into the speakers (3-16).

(1) Welcome to the investment game. I hope we will enjoy playing it.

(2) Hello, welcome to the game. Let's see how much money we can make.

(3) As long as cooperation continues, I'm happy to earn a bit less.

(4) I am expecting you to share because that is exactly what I am doing.

(5) I think we can definitely go home with much more money than this.

(6) I trust you and I'm sure that we can both benefit from each other.

(7) I trust you and I will show you that you can trust me as well.

(8) I will demonstrate that you can trust me, just as I trust you.

(9) If I keep all your investment, you will not invest any more.

(10) If we both invest in each other, the final reward will be bigger.

(11) If we both invest in each other, we will surely raise our earnings.

(12) If we could talk face to face at the moment, you'd know that I am being honest.

(13) In my opinion, we should keep cooperating until the end.

(14) We can both win the game but we have to keep sharing our money.

(15) We could finish the game better off than this if only we tried harder.

(16) We should have a clear strategy: always help each other out.