Production, perception, and comprehension of subphonemic detail

Word-final /s/ in English

Dominic Schmitz

Studies in Laboratory Phonology 11

language science press

Studies in Laboratory Phonology

Chief Editor: Martine Grice

Editors: Doris Mücke, Taehong Cho

In this series:

1. Cangemi, Francesco. Prosodic detail in Neapolitan Italian.
2. Drager, Katie. Linguistic variation, identity construction, and cognition.
3. Roettger, Timo B. Tonal placement in Tashlhiyt: How an intonation system accommodates to adverse phonological environments.
4. Mücke, Doris. Dynamische Modellierung von Artikulation und prosodischer Struktur: Eine Einführung in die Artikulatorische Phonologie.
5. Bergmann, Pia. Morphologisch komplexe Wörter im Deutschen: Prosodische Struktur und phonetische Realisierung.
6. Feldhausen, Ingo & Fliessbach, Jan & Maria del Mar Vanrell. Methods in prosody: A Romance language perspective.
7. Tilsen, Sam. Syntax with oscillators and energy levels.
8. Ben Hedia, Sonia. Gemination and degemination in English affixation: Investigating the interplay between morphology, phonology and phonetics.
9. Easterday, Shelece. Highly complex syllable structure: A typological and diachronic study.
10. Roessig, Simon. Categoriality and continuity in prosodic prominence.
11. Schmitz, Dominic. Production, perception, and comprehension of subphonemic detail: Word-Final /s/ in English.

ISSN: 2363-5576

Production, perception, and comprehension of subphonemic detail

Word-final /s/ in English

Dominic Schmitz

language
science
press

Dominic Schmitz. 2022. *Production, perception, and comprehension of subphonemic detail: Word-final /s/ in English* (Studies in Laboratory Phonology 11). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/365

This book is the revised version of the author’s PhD dissertation
Heinrich-Heine-Universität Düsseldorf, 2022 (D61)
ISBN: 978-3-96110-401-7 (Digital)
978-3-98554-059-4 (Hardcover)

ISSN: 2363-5576
DOI: 10.5281/zenodo.7267830
Source code available from www.github.com/langsci/365
Errata: paperhive.org/documents/remote?type=langsci&id=365

Cover and concept of design: Ulrike Harbort
Typesetting: Dominic Schmitz
Proofreading: Amir Ghorbanpour, Amy Amoakuh, Annika Schiefner, Brett Reynolds, Elliott Pearl, Jeroen van de Weijer, Ludger Paschen, Jean Nitzke, Sarah Warchhold
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XƎLATEX

Language Science Press
xHain
Grünberger Str. 16
10243 Berlin, Germany
http://langsci-press.org

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

Contents

Acknowledgments

Thank you to Ingo Plag and Dinah Baer-Henney. Without your ideas my dissertation, of which this book is a slightly altered version, would not have been possible. Thank you for your invaluable expertise and enthusiasm, for listening to my ideas and questions, and for providing a constant flow of new input. Thank you for guiding me through my PhD and for pushing me to sharpen my thinking.

Thank you to Arne Lohmann, who always had an open door and open ear for my questions.

Thank you to everyone in the DFG research unit Spoken Morphology for listening to talks on word-final /s/ over and over again, and for giving me insightful feedback.

Thank you to the Deutsche Forschungsgemeinschaft for funding the research unit, giving me the opportunity to fully concentrate on my research.

Thank you to James White and Andrew Nevins for giving me the opportunity to run my production experiment at Chandler House, University College London, and to Andrew Clark for his technical support.

Thank you to Yu-Ying Chuang for providing help, support, and valuable input on my linear discriminative learning implementation. Without your input my LDL study would not have been possible.

Thank you to Jennifer Hay for making it possible to conduct my perception and comprehension experiments at the University of Canterbury, Christchurch, and to Wakayo Mattingley and Gia Hurring who run them on site. Thank you to Jessie Nixon for providing recordings for the experiments. I am very grateful for your help in times of global turmoil.

Thank you to everyone in the English linguistics department at the Heinrich-Heine-Universität Düsseldorf. A great team of colleagues makes working on your PhD much easier. A special thanks to Viktoria, Simon, and Marie who listened to my complaints, and provided feedback and food for thought wherever and whenever necessary.

Thank you to Julika, Sally, Jennifer, Lisa-Marie, Ann-Sophie, Nina, and Lukas for the countless hours you spent in front of Praat, for pilot-testing my perception and comprehension experiments, and for proofreading my work.

Thank you to my best friend, Janina, who provided invaluable mathematical insight and stimulating discussions as well as happy distractions to rest my mind outside my research.

Thank you to my family for having faith in me, for your sympathetic ear, and for reminding me that there is more than word-final /s/ in the world.

Thank you to my partner, Dennis, for your love, patience, and support, and most of all, for always believing in me.

1 Introduction

The complexities of speech production, perception, and comprehension are enormous. This circumstance has led to the development of numerous models and theories of language production, perception, and comprehension since the dawn of the modern study of language structures in the early twentieth century. Especially since the rise of psycholinguistics in the 1960s, psycholinguistic methods and findings contributed to the development of pertinent theoretical approaches to language structure. As has been shown repeatedly in recent linguistic research, however, it remains a challenge for most if not all established approaches to account for findings on more and more intricate features of language such as differences in subphonemic detail.

In research on speech production, it was shown that homophonous lexemes differ in their acoustic duration due to differences in frequency (e.g. Jurafsky et al. 2002; Lavoie 2002; Gahl 2008; Drager 2011; Lohmann 2018). Such findings indicate that phonologically identical lemmas may differ in their phonetic realisation. Similarly, fine phonetic differences were also found, for example, for bound versus free bases (e.g. Kemps, Ernestus, et al. 2005; Kemps, Wurm, et al. 2005), for final segments of a mono-morphemic stem versus the final segments of the same stem if followed by a suffix (e.g. Sugahara & Turk 2004; 2009), and for prefixes in prefixed versus so-called pseudo-prefixed words (e.g. Smith et al. 2012). A popular case for the research of such fine-grained phonetic detail below the word level is word-final /s/ and /z/ in English. Previous corpus studies (Zimmermann 2016; Plag et al. 2017; Tomaschek et al. 2019) showed that the acoustic duration of word-final /s/ depends on its morphological make-up, with non-morphemic /s/ being longest and auxiliary clitic /s/ being shortest in duration. However, previous experimental studies (Walsh & Parker 1983; Hsieh et al. 1999; Seyfarth et al. 2017) found effects in the opposite direction. It is thus the first general aim of this book to investigate by means of a production task whether such durational differences between different types of word-final /s/ really exist, and to find a potential explanation for the contradictory nature of previous results. I will introduce relevant theoretical approaches of speech production such as feed-forward formal theories of morphology-phonology interaction (e.g. Chomsky & Halle 1968; Kiparsky 1982), the framework of Prosodic Phonology (e.g. Booij 1983; Nespor

& Vogel 2007), psycholinguistic theories of speech production (e.g. Levelt et al. 1999; Roelofs & Ferreira 2019), exemplar-based models (e.g. Goldinger 1998; Pierrehumbert 2001; Gahl & Yu 2006), and discriminative learning (e.g. Rescorla 1988; Ramscar & Yarlett 2007; Ramscar et al. 2010) to discuss their respective explanatory limits. As the approach of discriminative learning can only be meaningfully discussed in light of an implementation of such an approach, a linear discriminative learning network (e.g. Baayen, Chuang, Shafaei-Bajestan, et al. 2019) is implemented. This implementation not only allows for a discussion of the general approach itself, but potentially offers insight into the nature of the durational differences in word-final /s/.

Research on the perception of fine phonetic detail found that listeners make use of segment durations as a cue for word boundaries (Shatzman & McQueen 2006b) and to assist in differentiating phonologically similar lemmas (Warner et al. 2004). Findings on bare versus suffixed stems indicate that listeners make use of acoustic duration as a cue for distinguishing such stems (Kemps, Ernestus, et al. 2005; Kemps, Wurm, et al. 2005; Blazej & Cohen-Goldberg 2015). Yet, there is barely any research on the question of how small such subphonemic durational differences may be to remain perceptible. This question is answered for individual segments by a rather dated study by Klatt & Cooper (1975). That is, to be perceptible, a durational difference in fricatives should be of 25 ms or more. However, these authors found that perceptibility is worse in fricatives and word-final position. Hence, the second general aim of this book is to explore how small a durational difference in word-final /s/ is perceptible in a same-different perception task. I will discuss the findings taking into account abstractionist approaches (e.g. Klatt 1979; McClelland & Elman 1986; Norris 1994; Norris & McQueen 2008), approaches relying on fine phonetic detail (e.g. Goldinger 1996), approaches combining abstract representations and fine phonetic detail (e.g. Hawkins & Smith 2001; Pierrehumbert 2002), and computational models of speech perception (e.g. ten Bosch et al. 2015; Baayen, Chuang, Shafaei-Bajestan, et al. 2019).

For an account of the influence of subphonemic detail on comprehension, one can consider the same results which have been brought forward to describe the perception of such fine phonetic detail. As subphonemic durational differences are used as a cue for word boundaries (Shatzman & McQueen 2006b), they are not only perceptible but also used in the comprehension of words. Differentiating between unsuffixed and suffixed stems by means of acoustic durations (Kemps, Ernestus, et al. 2005; Kemps, Wurm, et al. 2005; Blazej & Cohen-Goldberg 2015) does not only indicate that such differences are perceived, but also that such differences are made use of in comprehension. In general, however, there is little research available which directly asks the question of whether subphonemic durational differences significantly influence comprehension (e.g. Blazej &

Cohen-Goldberg 2015). Thus, it is the third general aim of this book to investigate this question. This is done by means of two number-decision tasks in a mouse-tracking paradigm. Taking into account the findings from these experiments, I will discuss the same set of theoretical approaches as are taken into account for the results of the perception study: abstractionist approaches, approaches relying on fine phonetic detail, approaches combining abstract representations and fine phonetic detail, and computational models of speech perception.

The overarching goal of this book, then, is to draw a more detailed, intricate, and exhaustive picture of the production, perception, and comprehension of subphonemic detail. To achieve this goal, two important methodological decisions were taken. First, where applicable, pseudowords as well as real words are used as target items to account for potentially confounding effects of lexical properties (e.g. effects of frequency, e.g. Gahl 2008; Lohmann 2018; effects of storage, e.g. Caselli et al. 2016). Second, sound statistical analyses are performed, relying on novel statistical techniques where appropriate. Overall, the findings presented in this book are the results of a thorough methodological approach to item design and statistical analysis, offering a reliable account of the nature of subphonemic detail and a strong foundation for future research.

This book is structured as follows. In Chapter 2, I will give a detailed overview of previous findings on the production, perception, and comprehension of subphonemic durational differences and introduce pertinent theoretical approaches. Taking these approaches as a starting point, hypotheses to be investigated in the subsequent chapters are derived. Chapter 3 will introduce the general method used in this book. It will discuss pseudowords as a type of item (Section 3.1.1) and present the pseudoword and real word items used across all studies of this book (Section 3.1.2). Statistical methods and procedures are described in Section 3.2. Then, the approach of linear discriminative learning is introduced (Section 3.3). Chapter 4 presents the production study investigating the production of subphonemic durational differences in word-final /s/, while Chapter 5, relying on the introduction to linear discriminative learning in Section 3.3, presents the implementation of such a linear discriminative learning network to account for the nature of the reported subphonemic durational differences. In Chapter 6, I will present the perception study, which consists of a same-different task to investigate the perceptibility of durational differences in word-final /s/. Chapters 7 and 8 introduce and discuss the two number-decision tasks used to examine the influence of subphonemic durational differences on comprehension. In Chapter 9, I will bring together the results of the individual studies presented in Chapters 4 to 8 and discuss them in light of the general aims set in the present and the hypotheses given the following chapter. Chapter 10 concludes this book.

2 Subphonemic differences in phonologically identical elements

Research of the last decades has repeatedly shown that subphonemic differences are found in the production of phonologically identical elements (Walsh & Parker 1983; Hsieh et al. 1999; Cho 2001; Jurafsky et al. 2002; Lavoie 2002; Sugahara & Turk 2004; 2009; Kemps, Ernestus, et al. 2005; Kemps, Wurm, et al. 2005; Gahl 2008; Drager 2011; Smith et al. 2012; Zimmermann 2016; Ben Hedia & Plag 2017; Plag et al. 2017; Seyfarth et al. 2017; Lohmann 2018; Ben Hedia 2019; Tomaschek et al. 2019; Plag et al. 2020) and that such differences can be perceived as well as be used in comprehension (e.g. Klatt & Cooper 1975; Warner et al. 2004; Kemps, Ernestus, et al. 2005; Kemps, Wurm, et al. 2005; Shatzman & McQueen 2006b). It is on these findings that the research presented in this book is grounded. Instead of following a specific theory that is to be confirmed, the studies this book reports on are of an explorative nature. The research questions addressed in the individual studies are met with all relevant theories at hand to provide elaborate discussions of the respective findings. The exploratory nature of this research is typical of research spanning multiple areas of a discipline. In the present case, findings, concepts, and approaches of morphology, phonology, phonetics, computational linguistics, and psycholinguistics are combined. While this at times may prove difficult, it also enriches the knowledge gain of the field by combining theoretical accounts spanning different subdisciplines. The overall aim of the approach in this book follows its overarching aim of establishing substantiated knowledge on subphonemic detail and its role in production, perception, and comprehension.

Previous findings as well as the main theoretical accounts concerning the production, perception, and comprehension of subphonemic detail are introduced in the following sections, Section 2.1 and Section 2.2, respectively. In both sections, I will first review relevant previous empirical findings before I then introduce pertinent theoretical approaches and models. Taking the theoretical accounts as motivation, I will derive the hypotheses to be explored in the studies of this book. Finally, I will sum up the hypotheses for a concise overview in Section 2.3.

2.1 Production

The evidence for the presence of morphological information at the subphonemic level emerges mainly from the study of homophonous lexemes, stems, and affixes.[1] For homophonous lexemes, Gahl (2008) and Lohmann (2018) investigated acoustic realisations of seemingly homophonous word pairs such as *time* and *thyme* and found the more frequent member of each pair to be of shorter acoustic duration. Further evidence for differing acoustic realisations of supposedly homophonous lexemes was found by Drager (2011). Drager compared realisations of *like* as adverb, verb, discourse particle, and as part of the quotative *be like*. Differences surfaced in several phonetic parameters. Similar effects were found for function words such as *four* and *for* and different uses of words such as *to*, which were investigated by Lavoie (2002) and Jurafsky et al. (2002). Such fine realisational differences indicate that at the phonetic level, two or more phonologically homophonous lemmas may differ in their realisation.

Similarly, evidence shows that seemingly homophonous elements below the word level have different phonetic realisations. Kemps, Ernestus, et al. (2005) and Kemps, Wurm, et al. (2005) found that in Dutch and English segmentally identical free and bound variants of a base (e.g. *help* without a suffix versus *help* in *helper*) differ acoustically. Sugahara & Turk (2004; 2009) found phonetic differences between the final segments of a mono-morphemic stem as compared to the final segments of the same stem if followed by a suffix, e.g. in *mist rain* versus *missed rain*. The stem had slightly longer rhymes if followed by certain suffixes. Seyfarth et al. (2017) found that for words ending in fricatives, the durations of a word's morphological relatives influence the realisation of that word. In their study, stems of multi-morphemic words showed longer durations than similar strings of segments in homophonous mono-morphemic words (e.g. *free* in *frees* versus *freeze*). They concluded that the durational targets of the multi-morphemic word's relatives influence the word's duration to such an extent that a durational difference between the respective multi-morphemic word and its homophonous mono-morphemic counterpart arise. A similar effect of morphological relations influencing duration was found for plurals and their bare stems in a corpus-based study by Engemann & Plag (2021).

For prefixes, Smith et al. (2012) found systematic realisational differences for *dis-* and *mis-* between prefixed and so-called pseudo-prefixed words (e.g. *discolour* versus *discover*). Prefixed words showed longer durations and longer voice onset times, among other things. Ben Hedia & Plag (2017) and Ben Hedia (2019) showed that the more segmentable a prefix, the longer the duration of its nasal.

[1] An earlier version of this section has been published in Schmitz, Baer-Henney, et al. (2021).

On the articulatory level, Cho (2001) found evidence for the variability of intergestural timing between identical strings in mono- versus multi-morphemic contexts. In their electropalatographic study, Cho showed that the timing of the gestures for [ti] and [ni] in Korean show more variation when the sequence is mono-morphemic (/mati/ 'knot' and /pani/ 'name') as compared to the timing of the same gestures in multi-morphemic sequences (/mat-i/ 'the oldest' and /pan-i/ 'class-NOM'), thus indicating that morphological structure is reflected in articulatory gestures, which in turn may lead to correlates in the acoustic signal. Hence, morphology is reflected in the phonetic realisation of otherwise identical strings of segments.

In sum, it seems that there is vast evidence for seemingly homophonous elements, that is, lexemes, bases, and affixes, to differ on the level of speech production. Differences on the level of segments have been reported as well. Previous corpus studies on word-final /s/ in English found realisational differences between non-morphemic, suffix, and clitic variants. Zimmermann (2016) on New Zealand English (data from QuakeBox corpus; Walsh et al. 2013) and Plag et al. (2017) as well as Tomaschek et al. (2019) on North American English (data from Buckeye Corpus of Conversational Speech; Pitt et al. 2007) found that non-morphemic /s/ showed longer durations than suffix and clitic /s/. In turn, suffix /s/ also showed longer durations than clitic /s/. While these results draw a clear picture of /s/ duration across morphological categories (including the non-morphemic /s/), they are subject to unbalanced data sets due to the nature of corpora. That is, corpus data may contain a huge number of confounding and moderator variables that experimental data can control for (e.g. Gries 2015).

Previous experimental studies, however, have reported less consistent results and show some problematic methods and analyses. Walsh & Parker (1983) carried out a production experiment with three homophonous word pairs (e.g. *Rex* and *wrecks*). They measured the duration of the word-final /s/ in both the mono- and the multi-morphemic word of each pair in three different conditions. Each word was produced by eight to ten participants. Condition I consisted of an unambiguous context; condition II consisted of a semantically neutral context; condition III consisted of a semantically anomalous context. While in two of these conditions there was a small difference of 9 ms in the means of the different types of /s/, there was none in the third condition. Still, the authors concluded that "speakers of English systematically lengthen morphemic /s/" (Walsh & Parker 1983: 204). However, their analysed data set was small (110 observations), included a mixture of common and proper nouns, and no phonetic covariates were integrated in their analysis. Further, instead of applying appropriate inferential statistical methods (e.g. t-tests or more advanced methods), the mean durations of the types of /s/ under investigation were compared impressionistically. Therefore, there are several reasons to be sceptical of their results.

In another study, Hsieh et al. (1999) measured /s/ duration in child-directed speech in data originally elicited for another study (on vowel durations in function words, see Swanson & Leonard 1994). The authors found plural /s/ to be longer than third-person singular /s/. However, as the data originally was not designed for this endeavour, half of all plural items occurred sentence-finally, while almost all third-person singular items occurred sentence-medially. The durational difference found between the suffixes may hence have been due to effects of phrase-final lengthening (e.g. Klatt 1976; Wightman et al. 1992) rather than to inherent phonetic differences due to morphological categories.

In a more recent study, Seyfarth et al. (2017) conducted a production experiment to collect data on non-morphemic, plural, and third-person singular /s/ and /z/ durations. They found the non-morphemic variant to be shorter than the morphemic instances. However, they did not find differences between the voiced and the voiceless allomorphs during their analysis. This may be a worrisome result, especially considering the small number of items with voiceless allomorphs (n = 6) as compared to the high number of items with voiced allomorphs (n = 20) in their data.

Recently, Plag et al. (2020) found plural and genitive plural /s/ to be of different durations. In their study, the genitive plural suffix showed significantly longer durations as compared to the plural suffix. An overview of the durational differences found in the aforementioned experimental studies is given in 2.1.

Table 2.1: Overview of durational differences of word-final /s/ found in previous studies.

Study	Findings
Zimmermann 2016; Plag et al. 2017, Tomaschek et al. 2019	non-morphemic > plural > clitics
Walsh & Parker 1983	plural > non-morphemic
Hsieh et al. 1999	plural > third-person singular
Seyfarth et al. 2017	plural > non-morphemic
Plag et al. 2020	genitive plural > plural

In sum, there is evidence that there may be durational differences between different types of /s/. However, while results of corpus studies are in line with each other, they might be flawed due to imbalanced data sets. Previous experimental studies, on the other hand, have often relied on small data sets and lacked phonetic covariates, appropriate statistical methods, or a proper distinction of voiced and voiceless segments. Another crucial difference between corpus and experimental studies is the use of homophones. While all previous experimental studies restricted their data to homophone pairs, corpus studies take into consideration all words. The limitation to homophones and the resulting competition between their representations might be a problem in itself, as it appears to be unclear how members of homophone pairs are stored and connected to their respective frequencies. In all cases, previous results were subject to potentially confounding effects of the lexical properties (e.g. effects of frequency, e.g. Gahl 2008; Lohmann 2018; effects of storage, e.g. Caselli et al. 2016) and contextual effects (e.g. phrase final lengthening, e.g. Klatt 1976; Wightman et al. 1992) of the items under investigation. Also, so far, no experimental study included clitics in their analysis, whereas corpus studies have suggested that clitics show different durations than suffixes.

A study is therefore called for that investigates the durational nature of different types of word-final /s/ in English, preferably an experimental study with carefully controlled data avoiding potentially confounding effects. This book presents such a study investigating word-final /s/ in English by means of a pseudoword production task. In this task, three types of word-final /s/ were elicited: mono-morphemic, plural, and clitic /s/ (with the auxiliaries *is* and *has*). It will address some of the issues of previous studies. More precisely, the use of pseudowords prevents potential lexical effects to confound findings (see Section 3.1.1), while the highly controlled task evades the influence of contextual effects. Even though the data will also contain homophone pairs to a certain extent, the individual members do not have lexical representations. That is, one can rule out effects of competition between homophonous lexical entries due to their similar representations. In addition, the use of pseudowords eliminates potential differences in duration due to differences in frequency between the homophones.

Let us now turn to the question of how morpho-phonetic effects can be explained at the theoretical level. Existing theories make different predictions concerning the possible presence of durational differences between different types of /s/. I will discuss four approaches here: feed-forward models of phonology-morphology interaction, Prosodic Phonology, exemplar theory, and discriminative learning.

In standard feed-forward formal theories of morphology-phonology interaction, all types of /s/, be they morphemic or non-morphemic, are treated in a similar way (e.g. Chomsky & Halle 1968; Kiparsky 1982). In the case of morphological word-final /s/, a process called *bracket erasure* is said to remove all morphological information from a pertinent word form once retrieved from the lexicon during the stage of *lexical phonology* and leaves speech production without an insight into the morphological makeup at the stage of *post-lexical phonology*. After retrieval, there is no informational difference between word-final morphemic and non-morphemic types of /s/. Thus, there is nothing in such a system that could account for realisational differences, e.g. different durations, between phonologically identical suffixes and non-morphemic segments. The realisation of clitics is a post-lexical process to begin with and thus outside the scope of any prediction by this theory.

In the framework of Prosodic Phonology, there is a complex mapping of morphological structure onto prosodic structure (Booij 1983; Nespor & Vogel 2007). Since prosodic boundaries may correlate with particular phonetic properties, segments at such boundaries may show systematic differences in phonetic implementation (see, for example, Keating 2006). Phonetic differences between two phonologically homophonous affixes could therefore result from a difference in the prosodic structure that goes with the two affixes. In particular, different types of word-final /s/ can be analysed as having different positions in the hierarchical prosodic configuration. These configurations co-determine the degree of integration of an /s/ to the word it belongs to. These different degrees of integration might then emerge as durational differences between types of /s/ in speech production.

Applying the approach of Selkirk (1996), non-morphemic /s/, uncontroversially, is an integral part of the prosodic word, as shown in Panel A of Figure 2.1. Goad (1998) analyses plural /s/ as an *internal clitic*, which is adjoined to the highest prosodic constituent below the prosodic word, as shown in Panel B. In Goad (2002), however, plural /s/ is analysed as an *affixal clitic*, like third-person singular /s/ in Goad (2003) and Goad & White (2019), as shown in Panel C. The prosodic status of the cliticized auxiliary /s/ is not entirely clear, but presumably, it is best analysed as *free clitic*, as in Panel D.

The Prosodic Phonology approach thus posits a structural prosodic difference between non-morphemic /s/, plural /s/, and clitic /s/. This prosodic difference might be mirrored in durational differences. It is, however, not so clear what particular phonetic effects this approach would predict and by which processing mechanism the structural prosodic differences would be translated into different articulations. The most plausible prediction would be that closer integration into the prosodic word would correlate with shorter durations: Non-morphemic /s/ should be shortest, clitic /s/ longest, and plural /s/ in between. From the perspective of phrase-final lengthening (e.g. Klatt 1976; Wightman et al. 1992), one should also expect that clitic /s/ is longest, as it immediately precedes a phrase boundary.

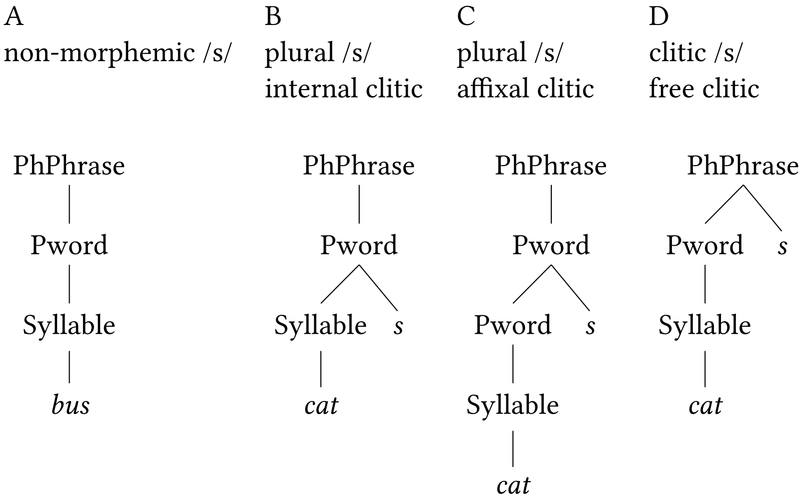


Figure 2.1: Prosodic structure of non-morphemic (A), plural (B, C), and clitic /s/ (D) as given in literature on Prosodic Phonology.

The distinction of lexical and post-lexical processing as introduced by the aforementioned standard feed-forward theories of morphology-phonology interaction is also an integral part of established theories in psycholinguistics. According to models of speech production such as the one proposed by Levelt et al. (Levelt et al. 1999; see Roelofs & Ferreira 2019 for an update), morphemic /s/ would not differ in its realisation from corresponding non-morphemic realisations of /s/. In such models, meanings are stored in the mental lexicon with their forms being represented phonologically. A module called *articulator* uses these phonological forms for speech production, hence, has no information on the lexical origin of particular segments. As a consequence, in this architecture, no systematic differences between different types of /s/ should emerge.

In contrast, exemplar-based models (e.g. Goldinger 1998; Bybee 2001; Pierrehumbert 2001; 2002; Gahl & Yu 2006) have an architecture that would in principle allow for morpho-phonetic effects. In such models, lexemes are linked to a frequency distribution over their phonetic outcomes as experienced by the individual speaker. These distributions are updated with each new experience: Experienced subtle subphonemic differences then may result in representations mirroring these properties. While such an account may allow for durational differences between different types of word-final /s/ to emerge from stored phonetic representations, it leaves open the question of how such systematic differences between clouds of exemplars would come about in the first place. The downside of this is that it is also unclear in which direction differences between different types of /s/ should play out.

Finally, there is the discriminative learning approach, which is based on simple but powerful principles of discriminative learning theory (Rescorla 1988; Ramscar & Yarlett 2007; Ramscar et al. 2010; see, for example, Baayen et al. 2011; Baayen, Chuang, Shafaei-Bajestan, et al. 2019 for its application to linguistic problems). According to this theory, learning results from exposure to informative relations among events in the individual's environment. Individuals use the associations between these events to create cognitive representations of their environment. Most importantly, associations and their resulting representations are updated constantly on the basis of new experiences. Associations are built between features (*cues*, e.g. biphones) and classes or categories (*outcomes*, e.g. different types of /s/) that co-occur in events in which the learner is predicting the outcomes from the cues (Tomaschek et al. 2019). The relation between cues and outcomes is modelled mathematically by the so-called Rescorla-Wagner equations (Rescorla & Wagner 1972; Wagner & Rescorla 1972; Rescorla 1988). Following these equations, an association strength or *weight* increases every time a cue and an outcome co-occur, while it decreases if a cue occurs without the outcome in a learning event. This results in a continuous recalibration of association strengths, which is a crucial part of discriminative learning.

In recent discriminative learning implementations, the association weights between semantic representations and phonetic representations have been shown to be predictive of phonetic durations (e.g. Stein & Plag 2021). With regard to final /s/, Tomaschek et al. (2019) show that the different durations of final /s/ can be understood as following from the extent to which words' phonological and collocational properties can discriminate between the inflectional functions expressed by the /s/. The input features (cues) for their discriminative network were the words (*lexomes* as pointers to the meaning of the forms) in a five-word window centred on the /s/-bearing word and the biphones in the phonological forms of these words. These cues are associated with the inflectional functions of the /s/. Two main measurements emerged as significant predictors of /s/ duration. The so-called *activation* (named *prior* in Tomaschek et al. 2019) is a measure of an outcome's baseline activation, i.e. of how well an outcome is entrenched in the lexicon. The other measure is *activation diversity*, which quantifies the extent to which the cues in the given context also support other targets. The general pattern now is the following: When the uncertainty about the targeted outcome increases, the acoustic duration of /s/ decreases. In other words, stronger support (both from long-term entrenchment and short-term from the context) for a morphological function leads to a longer, i.e. enhanced, acoustic signal. In sum, the discriminative approach predicts that differences between different types of /s/ may emerge from the associations of form and meaning that speakers develop as a result of their experience with the pertinent words. But what about pseudowords? It has recently been shown by Chuang et al. (2021) that these associations also play a role for pseudowords. Pseudowords have no representation in the lexicon, but, as these authors show, pseudowords nevertheless resonate with the lexicon due to their formal similarity with existing words. This resonance even influences subtle phonetic details such as duration. It is, however, yet unclear what kinds of durational differences can be expected between different types of /s/ in pseudowords.

Effects of informativity or predictability (which are also inherently present in discriminative learning approaches) are also to mention, as they may play a role as well (Seyfarth 2014; Cohen Priva 2015; Zee et al. 2021). Greater predictability of the word in its context has been found to lead to phonetic reduction, for example, to shortening in duration. On the other hand, higher paradigmatic predictability has been shown to correlate with longer duration (*paradigmatic enhancement*, e.g. Kuperman et al. 2007; Bell et al. 2021). As these informativity effects are necessarily bound to existing words, an experiment that uses pseudowords cannot straightforwardly test these approaches.

Based on the different theories laid out above, different hypotheses about durational differences between different types of /s/ in pseudowords can be set up. H prod$_1$, the *Feed-Forward Hypothesis*, arises from feed-forward approaches and is in accordance with the prediction that no systematic phonetic differences should be observed between different types of /s/. H prod$_2$, the *Prosodic Hypothesis*, is derived from prosodic approaches. According to these approaches, a higher degree of prosodic integration should correlate with shorter durations. Hence, non-morphemic /s/ should be shorter than plural /s/, and plural /s/ should be shorter than clitic /s/. Finally, exemplar-based approaches and discriminative learning approaches both predict the presence of morpho-phonetic effects, but it is unclear how these differences would play out for the three types of /s/ in the present production study. This is encapsulated in H prod$_3$, the *Emergence Hypothesis*.

In summary, the production study presented in Chapter 4 of this book intends to establish whether there are durational differences also with pseudowords, and if so, how these differences play out.

H prod$_1$: *Feed-Forward Hypothesis*

> There is no durational difference between word-final non-morphemic /s/, plural /s/, and auxiliary clitic /s/.

H PROD$_2$: *Prosodic Hypothesis*

There are durational differences between different types of word-final /s/: non-morphemic /s/ is shorter than plural /s/, plural /s/ is shorter than auxiliary clitic /s/.

H PROD$_3$: *Emergence Hypothesis*

There are durational differences between different types of word-final /s/ (non-morphemic, plural, and auxiliary clitic).

2.2 Perception and comprehension

Findings on subphonemic durational differences give rise to two further questions. First, are listeners able to perceive such subphonemic durational differences between different types of word-final /s/? That is, are listeners not only sensitive to differences between different phonemes (e.g. Goldstone & Hendrickson 2010) but can they pick up on differences between phonologically similar but phonetically different realisations? Second, if subphonemic durational differences are perceptible, are they used in comprehension? That is, does the perception of (un-)expected subphonemic features influence the comprehension process?

On the level of word perception and comprehension, Shatzman & McQueen (2006b) showed that listeners make use of segment durations as a cue for word boundaries. In their study, native speakers of Dutch listened to ambiguous sentences in which plosive-initial words, e.g. *pot* ‘jar’, were preceded by *eens* ‘once’. Additionally, the sentences could also refer to cluster-initial words instead, e.g. *een spot* ‘a spotlight’. The two readings were, among other acoustic features, different in regard to their /s/ durations: Word-initial /s/ was overall longer in duration than word-final /s/ (Δ = 51 ms). The authors found that listeners make use of such different durations for their lexical decision. That is, the durational difference between word-initial and word-final /s/ was perceptible and used for an informed lexical decision, i.e. in word comprehension.

Warner et al. (2004) investigated whether listeners perceive subphonemic differences in Dutch words of identical phonetic but different underlying phonological form, e.g. /met/ ‘measures (sg.)’ and /med/ ‘avoided (sg.)’, where both words phonetically are transcribed as [meit]. Productions of such word pairs showed differences in the subphonemic features between the members of a pair. One of these features was vowel duration, which listeners showed sensitivity to: Listeners were able to perceive subphonemic detail and to use this information in comprehension, even though differences were rather small, e.g. Δ = 3.5 ms for vowel duration.

Kemps, Ernestus, et al. (2005) and Kemps, Wurm, et al. (2005) found that listeners in Dutch and English are sensitive to the durational differences between stems in isolation and stems as parts of affixed word forms, e.g. *help* without a suffix versus *help* in *helper*. This finding is confirmed by similar results in Lee et al. (2020) and in Blazej & Cohen-Goldberg (2015). In their study, Blazej and Cohen-Goldberg showed that listeners make use of duration as a cue for distinguishing unsuffixed stems from suffixed stems, e.g. *clue* without a suffix versus *clue* in *clueless*. The authors found the influence of duration as a cue to be persistent in isolated and continuous speech, with full, reduced, and removed effects of articulation, and in implicit and explicit tasks.

Taking into account the aforementioned findings, the question arises what the just-noticeable difference to be perceived is. Klatt & Cooper (1975) found this difference for a change in duration to a single segment to be 25 ms. That is, below the durational difference of 51 ms found in Shatzman & McQueen (2006b) but well above the durational difference of 3.5 ms given in Warner et al. (2004). Further, according to Klatt and Cooper's findings, this just-noticeable difference threshold is influenced by several factors. Most importantly, differences in word-final position and differences in fricatives are less well perceptible.

In sum, evidence for the perception of subphonemic differences in phonologically similar segments and its effect on comprehension exists. However, such evidence is rather sparse and mainly concerned with lexical decisions or differentiation of unsuffixed and suffixed forms. To date, there is no study which looks into the perception and comprehension of phonologically identical but phonetically and morphologically different segments. Thus, two types of studies are called for. First, a study is needed that investigates whether durational differences found between such segments are perceptible. This is the aim of the same-different task I present in Chapter 6 of this book. Using real words as well as pseudowords, potential lexical effects are taken into account. Second, it needs to be investigated whether subphonemic detail is not only perceptible but also used in comprehension. This is the purpose of the two number-decision mouse-tracking tasks I present in Chapters 7 and 8. Using isolated real words with non-morphemic and plural /s/ in one of the tasks, and pseudowords with plural and clitic /s/ embedded in real word contexts in the other, a detailed image of whether comprehension is affected by subphonemic durational differences is drawn. That is, evidence for real words as well as for pseudowords and for several types of word-final /s/ will be illustrated.

Let us now turn to the question of how the perception and comprehension of subphonemic detail can be explained at the theoretical level. Existing theories of speech perception and comprehension make different predictions concerning the perception of subphonemic detail and its use in comprehension. I will discuss several groups of approaches here: Theories that make use of abstract representations, theories that rely on sets of features, theories that combine abstract representations and sets of features, and computational models of speech perception and comprehension.

In abstractionist models of speech perception, the phonetics of the incoming speech signal are translated into phonemic representations before the stage of lexical access. That is, the result of perception is of phonological nature and without information on phonetic detail. Well known examples of abstractionist approaches are the TRACE model (e.g. McClelland & Elman 1986), Shortlist (e.g. Norris 1994) and Shortlist B (Norris & McQueen 2008) as well as the speech perception and lexical access model introduced by Klatt (1979). All of these models have in common that perception of subphonemic detail is either considered to be a peripheral process at the margins of speech perception or that it is not considered at all. Additionally, some abstractionist models (e.g. Klatt 1979) perform time normalisation. Timing (and with that duration) is only conceived as important if it serves a discriminative role, e.g. in stress placement. As this group of abstractionist models does not integrate subphonemic detail in the process of perception, it cannot account for the perception of subphonemic detail and, consequently, its use in comprehension. If subphonemic detail is not considered for the outcome of the perception process, there is no need to perceive it in the first place. Thus, comprehension has no access to any subphonemic, pre-phonological-representation information.

Approaches that make use of features instead of abstract phonemic representations form another group of speech perception models. One such model is the Fuzzy Logical Model of Speech Perception (Massaro & Simpson 1987). It assumes that multiple sources of information influence speech perception, that listeners have continuous information about each source, and that the multiple sources are used together in the most meaningful manner. Sources contribute features of sounds as information, which are then used to build so-called summary descriptions. These, in turn, are the result of the perception process. That is, comprehension does not make use of abstract phonological representations as in abstractionist models but of sets of distinct features. Another model based on features was introduced by Lahiri & Marslen-Wilson (1991). Their approach assumes that there is a single underlying phonological representation per lexical item, which is compatible with all phonologically permissible variants of it in a given context. Entailed in such representations is only marked information, i.e. phonetic features. Concerning the perception of subphonemic durational differences, then, one may regard the two aforementioned models as inconclusive. If subphonemic segmental durational differences are accounted for as a meaningful feature, i.e. if it is assumed to be marked information, perception of durational differences in word-final /s/ can be accounted for. Then, such differences can be used in comprehension. If, however, duration is only considered a feature where it distinguishes between phonemes, then perception of subphonemic durational differences is uncalled for. As a consequence, subphonemic durational differences cannot be used in comprehension.

Exemplar-based models of speech perception (e.g. Goldinger 1996) also rely on features. They assume that individuals draw on a multitude of exemplars per word form, which are all stored in their mental lexicon. Exemplars contain detailed phonetic information, which gives space to information on subphonemic detail. In this regard, exemplar based models account for the perceptibility of subphonemic durational differences, as such differences are stored in exemplars and made use of in perception and comprehension.

However, previous research has shown that effects attributed to exemplars are not consistently found (e.g. Hanique, Aalders, et al. 2013). Such findings are the motivation for hybrid models. One such hybrid model introduced by Pierrehumbert (2002) assumes abstract generalisations as well as exemplars associated with phonological units, that is phonemes, phoneme sequences, and words. While speech production makes use of both abstract representations and exemplars, comprehension mainly relies on the exemplars. Another hybrid model, Polysp (Polysystemic Speech Perception), has been introduced by Hawkins & Smith (2001). Their model assumes that the analysis of acoustic input does not necessarily rely on its transformation into its linguistic units. Rather, it is situation-dependent whether the abstract phonological form of a word or one of its phonetic variants is accessed for comprehension. As phonetic detail is stored in hybrid models, such models can account for the perception of subphonemic differences and the usage of such differences in comprehension.

The final group of approaches to speech perception and comprehension consists of computational models. One such approach is DIANA, an end-to-end computational model of human word comprehension (ten Bosch et al. 2015; ten Bosch & Boves 2021). The implementation of DIANA supports not only the use of abstract units but also takes exemplars, i.e. phonetically rich information, as input for the modelling of comprehension. Thus, it avoids the assumption of a segmental prelexical layer between acoustic signal and the lexical layer, i.e. perception and comprehension. Similarly, linear discriminative learning (Baayen, Chuang, Shafaei-Bajestan, et al. 2019; see also Sections 2.1 and 3.3) does not assume a segmental representation layer for acoustic input. Instead, it makes use of Frequency Band Summary Features (FBSFs; Arnold et al. 2017) as representations. FBSFs consist of detailed information of small time intervals of the signal, containing, for example, information on minimum, maximum, median, initial, and final intensity values. The FBSFs of the complete set of acoustic input then are the result of perception, which is a detailed representation of the perceived phonetic signal. This representation is used in comprehension modelling. In sum, both computational approaches, DIANA and linear discriminative learning, assume detailed phonetic information to be the result of perception, and this information is then used for comprehension. Thus, such models can account for the perception of subphonemic differences and their usage in comprehension.

Based on the approaches laid out above, two hypotheses about the perception of subphonemic durational differences were formulated. H PERC$_1$, the *Abstractionist Hypothesis*, arises from models of speech perception that have an abstract phonological representation as output of perception. If all fine-grained phonetic detail is lost in perception, one needs not perceive it to begin with. H PERC$_2$, the *Phonetic Detail Hypothesis*, takes exemplar and hybrid models as well as the aforementioned computational models as a starting point to account for the perception of subphonemic detail. Based on the assumption of storage and usage of detailed phonetic information, subphonemic durational differences can be stored and should thus be perceptible. A hypothesis based on models which make use of features alone is not considered in this book, as predictions on the perceptibility of subphonemic durational differences by such approaches are inconclusive and thus not testable in the current setup.

H PERC$_1$: *Abstractionist Hypothesis*
Listeners are not sensitive to subphonemic durational differences between different types of word-final /s/.

H PERC$_2$: *Phonetic Detail Hypothesis*
Listeners are sensitive to subphonemic durational differences between different types of word-final /s/.

Finally, H COMP, the *Mismatch Hypothesis*, emerges as a consequence of the prior two hypotheses. That is, if fine-grained phonetic detail is perceptible, listeners may make use of it in comprehension. Thus, comprehension should be affected if subphonemic detail does not match its intended meaning or context. This influence may be visible in behavioural data, such as reaction times and mouse trajectories. This hypothesis is supported by the exemplar-based, hybrid, and computational approaches.

H COMP: *Mismatch Hypothesis*

If listeners make use of subphonemic durational differences in the comprehension of different types of word-final /s/, then a mismatch of subphonemic detail and intended meaning leads to
a) slowed down comprehension processes.
b) deviated mouse trajectories.

The perception study presented in Chapter 6 aims to establish whether durational differences in word-final /s/ are perceptible. The two comprehension studies of Chapters 7 and 8, then, investigate whether subphonemic detail is made use of in comprehension.

2.3 Summary

To summarise, this book aims at investigating three main areas potentially affected by subphonemic detail: production, perception, and comprehension. Previous findings and relevant theoretical accounts were illustrated in the present chapter.

The five subsequent chapters will each discuss one study. In Chapter 4, I will report on the production study that investigates whether durational differences between different types of word-final /s/ are also found in pseudowords. For this study, the following hypotheses are relevant:

H PROD$_1$: *Feed-Forward Hypothesis*

There is no durational difference between word-final non-morphemic /s/, plural /s/, and auxiliary clitic /s/.

H PROD$_2$: *Prosodic Hypothesis*

There are durational differences between different types of word-final /s/: non-morphemic /s/ is shorter than plural /s/, plural /s/ is shorter than auxiliary clitic /s/.

H PROD$_3$: *Emergence Hypothesis*

There are durational differences between different types of word-final /s/ (non-morphemic, plural, and auxiliary clitic).

Chapter 5 will present the implementation of a linear discriminative learning network that was used to analyse the data on non-morphemic and plural /s/ elicited in the aforementioned production study. This study comes without specific hypotheses. Rather, it was used to further investigate H PROD3, that is to explore how the discriminative learning approach might account for durational differences of different types of word-final /s/.

The third study, which constitutes Chapter 6, investigated the perception of durational differences in word-final /s/. The hypotheses derived for this study are the following:

H PERC$_1$: *Abstractionist Hypothesis*
Listeners are not sensitive to subphonemic durational differences between different types of word-final /s/.

H PERC$_2$: *Phonetic Detail Hypothesis*
Listeners are sensitive to subphonemic durational differences between different types of word-final /s/.

Finally, I will report on the two comprehension studies in Chapters 7 and 8. The first comprehension study used real words with non-morphemic and plural /s/ in isolation as stimuli, while the second comprehension study used pseudowords with plural and clitic /s/ embedded into real word contexts as stimuli. For both studies, this is the relevant hypothesis:

H COMP: *Mismatch Hypothesis*
If listeners make use of subphonemic durational differences in the comprehension of different types of word-final /s/, then a mismatch of subphonemic detail and intended meaning leads to
a) slowed down comprehension processes.
b) deviated mouse trajectories.

While each study comes with its individual methodological details, they also share some general methodology. In the next chapter, I will outline this general method applied across all studies, including the sets of stimuli and the foundations of the statistical analyses.

3 General method

The studies of this book share parts of their methodology: That is, the production study (Chapter 4), the linear discriminative learning implementation (Chapter 5), the perception study (Chapter 6), and the comprehension study on plural and clitic /s/ (Chapter 8) all make use of the same set of pseudowords. Pseudowords as a type of item are described in Section 3.1.1, before Section 3.1.2 explains how the pertinent pseudowords were created. The perception study (Chapter 6) and the comprehension study on non-morphemic and plural /s/ (Chapter 7) use sets of real words as stimuli. These sets are also presented in Section 3.1.2.

While each type of study comes with its own specific needs concerning its statistical analysis, a general introduction of statistical methods used in this book is given in Section 3.2. I will explain which types of regression analysis were used, and for what reason different types of regression analysis were applied across the studies of this book. Finally, Section 3.3 introduces the general rationale and mathematics of linear discriminative learning. This foundation is then used and further specified in Chapter 5, the linear discriminative learning implementation itself.

3.1 Stimuli

3.1.1 Pseudowords as items

Ever since Berko Gleason (1958) created the *Wug Test* to investigate if children already have productive knowledge of morphological rules, pseudowords have been the stimuli of choice in a multitude of studies in a wide variety of linguistic areas: morphology and morpho-phonology (e.g. Albright 2002; Albright & Hayes 2003; Pierrehumbert 2006; Dabrowska 2008; Krämer 2009; Kawahara 2012; Gouskova & Becker 2013), the mental lexicon (e.g. Rubenstein et al. 1970; Anshen & Aronoff 1988; Prasada & Pinker 1993; Vitevitch & Luce 1998; Eddington 2000; Shatzman & McQueen 2006a; Meunier & Longtin 2007), language acquisition (e.g. Dollaghan 1985; Singson et al. 2000; Friedrich & Friederici 2005; van de Vijver & Baer-Henney 2014), phonetics and phonology (e.g. Turcsan & Herment 2015; Schmitz et al. 2018), written word recognition (e.g. Burani et al. 1999;

McKay et al. 2008), spoken word recognition (e.g. Marslen-Wilson 1984), semantics (e.g. Ozubko & Joordens 2011), and memory performance (e.g. Hulme et al. 1995), among many others.

Pseudowords are commonly assumed to have the advantage of removing storage effects (e.g. Caselli et al. 2016) and frequency effects (e.g. Gahl 2008; Lohmann 2018), as well as effects of lexical relatedness (e.g. Schriefers et al. 1998) from the equation (e.g. Turcsan & Herment 2015). Using pseudowords as stimuli can make a researcher's life easier in that one has to consider fewer interfering factors. Along the same lines, pseudowords are commonly assumed to be semantically *empty shells* (e.g. Günther 1983; Frisch et al. 2000; Turcsan & Herment 2015). Thus, pseudowords assumably reflect the language-related capacity of speakers, e.g. in terms of morphological productivity as in the seminal study by Berko Gleason (1958), without any interferences caused by confounding factors, e.g. effects of storage, frequency, or lexical relatedness. For the studies presented in this book, this assumption provides a major advantage. Without intervening effects of storage, frequency, and lexical relatedness, pseudowords make the perfect type of item for highly controlled experimental setups. Thus, confounds of the aforementioned effects on acoustic duration can be ruled out in a production experiment, and an interaction of such effects with perception and comprehension can be avoided in perception and comprehension experiments.

Yet, there is a growing body of research from different areas that challenges the assumption of semantically empty, autonomous pseudowords. On the sub-word level, research on phonaesthemes demonstrates that certain sound combinations are paired with meanings (Bergen 2004; Kwon & Round 2015). For example, the /tw/ onset in words like *twist*, *twirl*, *tweak*, *twill*, *tweed*, *tweezer*, *twiddle*, *twine*, and *twinge* is associated with the semantics of twisting (Bolinger 1950). Research investigating sound symbolism repeatedly showed that certain sounds are associated with certain shapes. Most prominently, research on the *bouba*/*kiki* phenomenon showed that rounded vowels are matched with rounder shapes, and that unrounded vowels are matched with pointed shapes. This effect holds across different ages, i.e. can also be found in pre-school children (Maurer et al. 2006), as well as across cultures and writing systems (Ćwiek et al. 2022). Another recent example is the /r/ sound, which across a multitude of languages has been claimed to be associated with roughness (Winter et al. 2022).

On the word level, research on onomatopoeia shows that certain combinations of sounds can be used to imitate sound (Pratha et al. 2016). Studies on sound symbolic patterns in Pokémon names show that the number of voiced obstruents correlates with size, weight, evolution levels, and general strength parameters and that vowel height correlates with size and weight (Kawahara et al. 2018). Independent of the individual names being proper nouns, their phonological composition is connected to the object they name. A similar connection is found in nicknames. For example, taller major league baseball players have longer nicknames (Shih & Rudin 2020). Apart from proper nouns, size adjectives in English apparently show comparable observations. Winter & Perlman (2021) found that sound structure is highly predictive of semantic size, most strongly for the phonemes /ɪ, i, ɑ/, and /t/. Finally, the names of villains in fiction literature commonly sound harsher, as they contain more voiceless segments (Elsen 2008). It thus seems unlikely that pseudowords, when being used as stimuli in experiments, somehow circumvent all these potential sub-word and word level factors which may contribute to some sort of meaning.

Indeed, evidence for semantic content of pseudowords has recently been reported by Chuang et al. (2021). In their study, it was shown that the assumption that pseudowords are bare of meaning is most probably wrong. Due to their formal similarity with existing words, pseudowords resonate with the lexicon. As a result, they may in fact carry some sort of meaning. Chuang et al. (2021) implemented a linear discriminative learning network (Baayen, Chuang, Shafaei-Bajestan, et al. 2019; see Section 3.3) to demonstrate that quantitative measures gauging the semantic neighbourhood of pseudowords predict reaction times in lexical decision and the pseudowords' acoustic durations. Hence, pseudowords are not entities independent of real words, but interact with the lexicon.

This, finally, raises one important question for the present book: Can pseudowords be employed as stimuli without taking their semantics into consideration? Recall the major advantage assumed for pseudowords as stimuli given earlier in this section. First, pseudowords are held to be free of storage and frequency effects. This is still true, even with semantically non-empty pseudowords. In very general terms, a pseudoword is a non-lexical word, and thus is neither stored in the lexicon nor does it have a frequency. Second, pseudowords are not affected by lexical relatedness effects. Such effects describe that a word is more easily recognised when it is preceded by a semantically or associatively related word than when it is preceded by an unrelated word (Schriefers et al. 1998). This, again, still holds for pseudowords. As pseudowords are unknown to the individual, no preceding context can make a pseudoword more recognisable. However, while on the one hand these advantages may still hold, the findings of Chuang et al. (2021) on the other hand show that pseudoword semantics influence reaction times and acoustic durations.

In sum, pseudowords can be employed as stimuli, even though they apparently are not semantically *empty shells*. But even as semantically *non-empty shells* they show some advantages over real words as items, as they have no previous entry in the lexicon, have no frequency, and cannot be predicted from their context. Yet, depending on the experiment, one is well advised to take their semantics into consideration. In this book, the results of the production study in which pseudowords were used as items (Chapter 4) are first analysed independently of any pseudoword semantics. In a subsequent implementation of linear discriminative learning (Chapter 5), pseudoword semantics are then considered in an analysis of a subset of the production study data. Further, pseudowords are used in the perception experiment (Chapter 6) and in the second comprehension experiment (Chapter 8). While pseudoword items are not free of meaning, they nonetheless are free of storage effects at the word level which potentially influence perception and production. Thus, pseudowords make good stimuli for such tasks.

3.1.2 Real word and pseudoword stimuli

The individual experiments of this book share parts of their item sets consisting of real words and pseudowords.[1] That is, the production study (Chapter 4), the perception study (Chapter 6), and the comprehension study on plural and clitic /s/ (Chapter 8) use the same set of pseudowords, while the perception study (Chapter 6) and the comprehension study on non-morphemic and plural /s/ (Chapter 7) share parts of a set of real word items.

For the use of pseudowords as items, a set of forty-eight pseudowords was created, following the phonotactic constraints of English (Gontijo et al. 2003).[2] The pseudowords can be grouped into six groups depending on their onset cluster and nucleus: Each group is defined by its particular stop plus approximant onset (/pl, bl, kl, gl, pr/) and its vowel. The vowel was either a short vowel (/ɪ, ʌ/), a long vowel (/iː, uː/), or a diphthong (/aʊ, eɪ/). In each group, eight different pseudowords were created by adding either a single consonant coda, i.e. /p, t, k, f/, or a consonant cluster coda, i.e. /ps, ts, ks, fs/. The set of coda consonants preceding the /s/ was chosen in such a way that the voiceless realisation of the /s/ allomorphs was elicited. Pseudowords with a simple coda were created for morphemic /s/ elicitation, while pseudowords with a complex coda were created for non-morphemic /s/ elicitation.

[1] An earlier version of this section has been published as part of Schmitz, Baer-Henney, et al. (2021).

[2] It only later came to attention that English phonotactics do not allow for /aʊ/ nuclei to be followed by non-coronal coda consonants such as /p, k, f/. However, as variation in pronunciation was expected and accounted for where necessary, this did not influence the results of the experiments which made use of these pseudowords.

One issue when constructing pseudowords is their spelling. For vowels, orthographic representations were chosen following the highest phonotactically legal grapheme-phoneme probabilities (Gontijo et al. 2003; see the supplementary material given in Chapter 11 for the top competitors for nucleus grapheme representations for each pseudoword group). The aforementioned coda consonants, however, showed a variety of possible orthographic representations to choose from. That is, /p/ may be represented by <p> or <pp>, /t/ may be represented by <t> or <tt>, /k/ may be represented by <k>, <c>, or <ck>, and /f/ may be represented by <f>, <ph>, or, exceptionally, by <gh>. When combined with a coda-internal /s/, some additional options can be observed: /ks/ may not only be represented as <ks>, <cs> or <cks> but also as <x>, /ps/ may be represented as <ps>, <pps>, and <pse>, and /ts/ may be represented as <ts>, <tts>, and <tz>. The choice of orthographic representation is important for two reasons. First, when comparing two kinds of words, variable representations add another source of variation of unclear consequences and should be avoided. Second, studies on the influence of number of letters on spoken language production have found that increasing the number of letters to represent a single sound may go together with longer durations in speech (e.g. Brewer 2008). Based on these considerations, the following orthographic representations were chosen for all word-final clusters: /ks/ is represented uniformly as <ks>, /ps/ is represented uniformly as <ps>, /ts/ is represented uniformly as <ts>, and /fs/ is represented uniformly as <fs>. Table 3.1 shows the final set of pseudowords and their orthographic and phonological representations.

Sets of real word items were created for the perception task (Chapter 6) and the comprehension task on non-morphemic and plural /s/ (Chapter 7). All real word items consist of one syllable to exclude a potential influence of stress placement. Items start with a simple onset and end in either non-morphemic or plural word-final /s/ preceded by a voiceless stop, i.e. /p, t, k/. As for the nuclei, an equal distribution of short monophthongs, long monophthongs, and diphthongs was desired to avoid an unwanted potential effect of vowel quality. For this, words were extracted from the British National Corpus (BNC; Davies 2004). Table 3.2 displays all selected real words with non-morphemic word-final /s/; Table 3.3 displays all selected real words with plural word-final /s/.

As can be seen in Table 3.2, it was not possible to find monomorphemic words with an even distribution of short monophthongs, long monophthongs, and diphthongs. More precisely, only one word with a long monophthong and a word-final non-morphemic /s/ preceded by a voiceless stop could be identified using the BNC, i.e. *corpse* /kɔ:ps/. Another monomorphemic word with a short monophthong was used instead.

Table 3.1: Orthographic (*orth.*) and phonological (*phon.*) representations of all pseudowords.

	Group	/glɪ/	/prʌ/	/pli:/	/clu:/	/blaʊ/	/gleɪ/
pseudowords for morphemic /s/	*orth.*	*glip*	*prup*	*pleep*	*cloop*	*bloup*	*glaip*
	phon.	/glɪp/	/prʌp/	/pli:p/	/klu:p/	/blaʊp/	/gleɪp/
	orth.	*glit*	*prut*	*pleet*	*cloot*	*blout*	*glait*
	phon.	/glɪt/	/prʌt/	/pli:t/	/klu:t/	/blaʊt/	/gleɪt/
	orth.	*glik*	*pruk*	*pleek*	*clook*	*blouk*	*glaik*
	phon.	/glɪk/	/prʌk/	/pli:k/	/klu:k/	/blaʊk/	/gleɪk/
	orth.	*glif*	*pruf*	*pleef*	*cloof*	*blouf*	*glaif*
	phon.	/glɪf/	/prʌf/	/pli:f/	/klu:f/	/blaʊf/	/gleɪf/
pseudowords for non-morphemic /s/	*orth.*	*glips*	*prups*	*pleeps*	*cloops*	*bloups*	*glaips*
	phon.	/glɪps/	/prʌps/	/pli:ps/	/klu:ps/	/blaʊps/	/gleɪps/
	orth.	*glits*	*pruts*	*pleets*	*cloots*	*blouts*	*glaits*
	phon.	/glɪts/	/prʌts/	/pli:ts/	/klu:ts/	/blaʊts/	/gleɪts/
	orth.	*gliks*	*pruks*	*pleeks*	*clooks*	*blouks*	*glaiks*
	phon.	/glɪks/	/prʌks/	/pli:ks/	/klu:ks/	/blaʊks/	/gleɪks/
	orth.	*glifs*	*prufs*	*pleefs*	*cloofs*	*bloufs*	*glaifs*
	phon.	/glɪfs/	/prʌfs/	/pli:fs/	/klu:fs/	/blaʊfs/	/gleɪfs/

Table 3.2: Real word items with non-morphemic word-final /s/. Frequency measures are taken from the BNC (Davies 2004).

		Word	Frequency	Vowel	Vowel quality
words used in the first comprehension task	words used in the perception task	*mix*	1669	ɪ	short
		box	8254	ɒ	short
		tax	15627	æ	short
		coax	12	əʊ	diphthong
		hoax	148	əʊ	diphthong
		corpse	754	ɔ	long
		lynx	98	ɪ	short
		flux	494	ʌ	short
		wax	644	æ	short
		fax	997	æ	short
		lapse	251	æ	short
		fox	1418	ɒ	short

Table 3.3: Real word items with plural word-final /s/. Frequency measures are taken from the BNC (Davies 2004).

		Word	Frequency	Vowel	Vowel quality
words used in the first comprehension task	words used in the perception task	*books*	1669	ʊ	short
		steps	8254	ɛ	short
		rights	15627	aɪ	diphthong
		points	12	ɔɪ	diphthong
		groups	148	u	long
		parts	754	ɑ	long
		costs	98	ɔ	short
		crusts	494	ʌ	short
		rates	644	eɪ	diphthong
		notes	997	əʊ	diphthong
		sports	251	ɔ	long
		cheats	1418	i	long

3.2 Statistical analysis

The statistical analyses for all studies were conducted using the software environment R (R Core Team 2020) in the integrated development environment RStudio (RStudio Team 2020). The main analyses of all studies consisted of different forms of regression modelling. In the following sections, I will give a general introduction to the types of regression models fitted. Additionally, I will discuss issues pertinent to the individual types of regression models and how they were dealt with. The details of the individual models as well as the issues encountered while developing them will be discussed in the respective chapters.

3.2.1 Linear mixed-effects regression

The analyses of the production study data and of the linear discriminative learning implementation data (see Sections 4.2 and 5.2) make use of linear mixed-effects regression models (henceforth LMER models). LMER models as such are an extension of multiple linear regression models. Multiple linear regression has long been an established method to analyse linguistic data (e.g. Baayen 2008; Winter 2019). As the name suggests, multiple linear regression can model a dependent variable in the presence of multiple independent variables at once. One can, for example, investigate whether the morphological makeup of a word-final /s/ significantly influences its duration, while also taking into account the effects other variables might show. While this in itself is a promising statistical tool, multiple linear regression falls short in one important aspect. It does not differentiate between highly regular and predictable variables such as *speaking rate* on the one hand, and highly irregular and virtually unpredictable variables such as *experimental participant* on the other hand.

Such irregular and unpredictable variables, and the *experimental participant* variable in particular, are the prototypical case of so-called random effects in linear mixed-effects regression. In general, random effects are factors with levels randomly sampled from a larger population (Baayen 2008). That is, a random effect is not repeatable, as the set of possible levels for a repeatable factor is fixed, with each level being repeatable itself. Taking the example of the *experimental participant* variable, it makes a lot of sense to classify this variable as a random effect: Participants of a study are a random sample of a larger population; if one was to repeat a study, one would recruit other randomly sampled participants; subjects may behave differently, i.e. unpredictably, on a day to day, and maybe even hour to hour basis. This notion of random effects follows the definition introduced by Green & Tukey (1960), and while there are other competing definitions (see e.g. Kreft & de Leeuw 1998; Searle et al. 2009; Snijders & Bosker 2011; McElreath 2015), this is the definition I will adhere to. The counterpart of random effects are fixed effects. These show repeatable levels and, in most cases, make up the variable(s) of interest in a mixed-effects regression model (Baayen 2008).

LMER models were fitted as implemented by the packages `lme4` (Bates et al. 2015), `lmerTest` (Kuznetsova et al. 2017), and `LMERConvenienceFunctions` (Tremblay & Ransijn 2020). Following the standard backward stepwise selection process (e.g. Baayen 2008), the first model for each analysis contained the whole set of pertinent independent variables as fixed and random effects, adhering to the aforementioned concept of effect structures. The whole set of variables here refers to the set of variables after taking measures to avoid issues of collinearity (see Section 3.2.3). By starting with a full set of theoretically justified random variables, I followed the *keep it maximal* policy of Barr et al. (2013) for results that are most generalisable. Interactions of fixed effects were included where motivated by theory.

Such a full model was then continuously reduced through step-wise exclusion of non-significant factors using the step function for linear mixed-effects regression models introduced by the `lmerTest` package (Kuznetsova et al. 2017). This function starts with the backward elimination of random-effect terms, followed by the backward elimination of fixed-effect terms. The result of this step-wise exclusion is a model which contains only variables with significant effects on the dependent variable.

At the last stage of the model fitting process, the resulting model's residuals were trimmed (e.g. Baayen & Milin 2010). Data points with residuals larger than 2.5 standard deviations were removed, ensuring a satisfactory distribution of residuals.

The final model was then analysed in terms of its R^2 values which were computed with the `MuMIn` package (Barton 2020; for marginal and conditional R^2 value computation see Nakagawa et al. 2017). The marginal R^2 value of a model indicates the percentage of variation in the data explained by the fixed effects of that model. The variance explained by the entire model is given by its conditional R^2 value.

Lastly, the predictor strength of individual variables was checked by taking the respective final model as template. For each predictor variable, a model was fitted lacking a particular variable. This resulted in a number of models, each lacking a different predictor. Then, marginal R^2 values were computed for these models and finally compared. The variable leading to the highest decrease in marginal R^2 value as compared to the final model is thus the variable showing the highest predictor strength. This procedure was implemented using the `predictor_strength` function of the `SfL` package (Schmitz & Esser 2021).

3.2.2 Generalised additive models

As one specific type of linear regression model, LMER models assume effects of numeric predictors to be strictly linear. This assumption is no longer met when working with numeric predictors which show non-linear effects. Modelling a non-linear variable as if it were linear results in inaccurate predictions, leading to unreliable coefficients and probability values (Baayen & Linke 2020).

Thus, linear mixed-effects regression is no longer a suitable statistical tool if such variables are to be involved. Instead, generalised additive models (henceforth GAMs) may be used as an appropriate tool, and indeed have been used in various linguistic research already (see e.g. Wieling et al. 2011; Linke et al. 2017; Milin, Divjak, et al. 2017; Tomaschek, Tucker, Fasiolo, et al. 2018). GAMs take a number of different arguments; however, I only need to consider four of them for the present purposes. First, categorical variables can be included in GAMs straightforwardly. In GAMs, the effects of categorical variables are most often reported under the term of *parametric effects*; a term I will use in the pertinent sections. Second, numeric variables are included in GAMs as so-called *smooth* or *smoother* terms. A numeric variable's smooth term expresses the estimated effect of that variable on the dependent variable. Smooth terms, in stark contrast to effects predicted by linear regression, can take the form of wiggly curves. Such wiggly curves are the weighted sum of their basis functions. I will come back to the specifications of basis functions in the description of the modelling process itself. Third, interactions of predictor terms are included as so-called *tensor product interactions*. Fourth, GAMs can incorporate random effects. GAMs containing random effects are called generalised additive mixed models (henceforth GAMMs). Including adequate random effects may help the interpretability of the model output as it protects against overly wiggly curves (Baayen & Linke 2020).

While general Gaussian GAMMs such as described above have not been used in the analyses of data presented in this book, three further specialised types of GAMMs, which rely on the same basic structures, have: GAMMs for beta distributed data (henceforth BGAMMs; Wood 2017), piece-wise additive mixed models (henceforth PAMMs; Bender & Scheipl 2018), and additive quantile regression models (henceforth QGAMs; Fasiolo et al. 2021). BGAMMs integrate the mathematical assumptions of beta regression in GAMMs. They can be used to adequately model data for which observations are limited to the open interval (0,1) (Ferrari & Cribari-Neto 2004; Smithson & Verkuilen 2006). While the first choice for modelling beta distributed data in R commonly is the `betareg` package (Cribari-Neto & Zeileis 2010), this package cannot integrate random effects into its model calculations. As beta regression was used for the analysis of a subject-specific measure, a random effect for individual subjects seemed worthwhile. I thus used BGAMMs instead of common beta regression models (Chapter 6). PAMMs have been developed for time-to-event analyses in the GAMM framework. They offer insight into the temporal dynamics of predictor effects. Thus, they are the tool of choice for the analyses of the reaction time data of the comprehension study on non-morphemic and plural /s/ (Chapter 7). QGAMs, on the other hand, provide an adequate tool to analyse data with a high level of autocorrelation. Timeseries of changing coordinates are characterised by strong correlations between the positions at time t and at $t-1$. Such autocorrelation is an issue if unaddressed, as model predictions become less reliable with higher levels of autocorrelation. QGAMs, however, show a high prediction accuracy even in the presence of high autocorrelation (Fasiolo et al. 2021). Using QGAMs, individual GAMs or GAMMs are fitted for any given conditional quantile of the response distribution (Tomaschek et al. 2021). As such, QGAMs are the appropriate tool to analyse the mouse-track coordinate data obtained by the comprehension studies (Chapters 7 and 8).

Depending on the specific type of GAMM, suitable packages were used for modelling. BGAMMs were fitted with the `mgcv` package (Wood 2017), PAMMs were fitted with the `pammtools` package (Bender & Scheipl 2018), and QGAMs were fitted with the `qgam` package (Fasiolo et al. 2021). As a stepwise selection process is uncommon in research literature using GAMMs, only one model was created.

This model was then tested for concurvity issues (see Section 3.2.3) using the `concurvity` function of the `mcgv` package. In case a variable showed a high concurvity value, this variable was excluded. The model was then re-fit without the excluded variable, and again checked for concurvity issues.

The last step of the model fitting process consisted of a check of basis functions. That is, for smooth terms of the fitted models I checked whether the number of basis functions was sufficient. This is indicated by the so-called k-index as reported by the `gam.check` function of the `mgcv` package. The further below 1 this value is, the more likely it is that there is a missed pattern left in the residuals and the number of basis functions in the model specification is too low (Wood 2017). In that case, the model was re-fit with a higher number of basis functions. The adjustment of the number of basis functions was done in small increments as to consider two points. First, on a theoretical note, models should not be more complex due to more basis functions than absolutely necessary, following the reasoning of Occam's razor. Second, on a mathematical note, the number of basis functions should be lower than the number of a variable's distinct values (Baayen & Linke 2020).

3.2.3 Collinearity and concurvity

One issue to address when fitting a linear model to a multitude of conceptually similar or potentially interrelated covariates is collinearity (Tomaschek, Hendrix, et al. 2018). Collinearity is a threefold issue. First, it may lead to unexpected and uninterpretable model estimates. Second, the model fit to the data may be unstable, i.e. the removal or addition of just few data points may change the model estimates drastically. Third, it may overestimate the effect of predictors, in that on its own a variable shows no significant effect on the dependent variable, while in combination with collinear variables it does. To avoid these issues, before each modelling process, variables were tested for their correlation coefficients.

For highly correlated variables, i.e. with correlation coefficients of $|rho| \geq 0.5$, one of two strategies was adopted. While there is no "one correct" way to deal with collinearity (Tomaschek, Hendrix, et al. 2018), these are two of the most commonly used strategies. The first strategy consisted of the competitive exclusion of one of two highly correlated variables. That is, for each pair of highly correlated variables, two linear mixed-effects models, each containing only one of two variables, were created and compared with a log-likelihood test. Each of these models contained the same variable as dependent variable, one of the highly correlated variables as fixed effect, and subject as random intercept. This procedure was done manually, and the results were checked with the `predictor_competition` function of the `SfL` package. This procedure allowed me to decide which of the covariates under discussion was a stronger predictor for the dependent variable. This covariate was then kept while the other one was no longer used.

Depending on the number of highly correlated variables, the first strategy may lead to a significant loss of predictor variables. Thus, in such cases a second strategy was adopted: Principal Component Analysis (PCA; e.g. Venables & Ripley 2002; Baayen 2008; Tomaschek, Hendrix, et al. 2018). In a PCA, the dimensionality of the data is reduced by transforming the included variables into principal components. These transformations result in linear combinations of the predictors that are orthogonal to each other. Thus, the resulting principal components are not correlated. PCAs for sets of only numeric variables were carried out using the `prcomp` function of the `stats` package (R Core Team 2020); PCAs for sets of numeric and factor variables were carried out using the `PCAmix` function of the `PCAmixdata` package (Chavent et al. 2017), which allows the simultaneous integration of continuous and discrete variables. A PCA computes as many principal components as variables were specified as input. The next step of the PCA is to determine how many of these principal components are meaningful and thus should be retained for further use. For this decision, several rules of thumb were followed (cf. O'Rourke et al. 2005; Baayen 2008). First, any component that displays an eigenvalue greater than 1 accounts for a greater amount of variance than had been contributed by one variable. Such a component is therefore potentially meaningful. Second, one should retain enough components so that the cumulative percentage of variance explained is equal to some minimal value. Following other implementations of principal component analyses, a value of 80% was aimed at (e.g. O'Rourke et al. 2005). Third, only interpretable components are to be retained. That is, each component is made up out of loadings, i.e. parts of the variables included in the PCA's computation represented by correlation coefficient values. If none of these variables is strongly represented in a component, the interpretability of that component is extremely low, rendering the component of small interest for further analyses. Thus, this strategy to avoid issues of collinearity is only meaningful as long as the resulting new variables are interpretable. If they were indeed not, the aforementioned first strategy was used instead.

Finally, all final models were checked for their variance inflation factors (VIFs). VIF values equal to or greater than 3 indicate the risk of introducing collinearity (e.g. Zuur et al. 2010). If a predictor with a high VIF value was identified, the model was re-fit after the exclusion of that predictor. Then, VIFs were computed again to make sure all potentially harmful variance inflation factor values were dealt with.

While collinearity is an issue in linear models, such as in LMER models, a similar issue is at stake for non-linear models, such as in GAMMs. In the GAMM setting, this issue is referred to as concurvity. Concurvity is the nonparametric analogue of collinearity, and may lead to the same issues in GAMMs as collinearity does in LMERs (e.g. Ramsay et al. 2003). Thus, GAMMs were checked for issues of concurvity during the fitting process.

3.3 Linear discriminative learning

Linear discriminative learning (henceforth LDL; e.g. Baayen, Chuang, Shafaei-Bajestan, et al. 2019) as a computational model implements a discriminative view of learning.[3] In contrast to deep learning models that have multiple hidden layers based on non-linear functions, LDL networks are very simple two-layer networks and are linguistically transparent and interpretable. In LDL, the mental lexicon consists of five high-dimensional numeric matrices, each of which represents a different subsystem: the visual matrix, retina; the auditory matrix, cochlea; the speech matrix, speaking; the spelling matrix, typing; and the semantic matrix. For the current implementation, the semantic and the speech matrix are most important.

With regard to the mappings between vectors, linear mappings are implemented. These mappings are estimated using the linear algebra of multivariate regression. Thus, each mapping is defined by a matrix A that transforms the row vectors in a matrix X into the row vectors of a matrix Y, i.e. $Y = XA$. Then, $A = X'Y$, where X' is the generalised inverse of X. I will return to the mapping of matrices later in this section, and refer the interested reader to Baayen, Chuang, Shafaei-Bajestan, et al. (2019) for an introduction to the mathematical details, as well as to Milin, Feldman, et al. (2017) for a detailed discussion on the restrictions and possibilities of linear mappings.

Another important feature of LDL is its notion of lexomes, i.e. basic semantic units corresponding to words or morphological functions. As outlined in Chuang et al. 2021, lexomes fall into two groups: content lexomes, and inflectional and derivational lexomes. Content lexomes can be morphologically simple or complex forms, i.e. *cat* and *cats*. Inflectional lexomes represent inflectional functions, e.g. number, tense, or aspect. Derivational lexomes represent derivational functions, e.g. morphological categories such as -NESS, -LESS, or UN-. Each lexome is

[3] An earlier version of this section has been published as part of Schmitz, Plag, et al. (2021).

paired with a vector of the aforementioned five subsystems. That is, for the semantic matrix, each lexome is paired with a semantic vector, making each lexome a pointer to a semantic vector on the one hand (Milin, Feldman, et al. 2017), and a location in a high-dimensional space on the other hand. For monomorphemic words, the semantic vector is identical to the semantic vector of the corresponding lexome. Thus, the semantic vector of the word *cat*, $\overrightarrow{cat}$, is identical to the vector of the lexome CAT. For complex words, the semantic vector is the sum of its corresponding lexome vectors. Accordingly, the semantic vector of the word *cats*, $\overrightarrow{cats}$, is the sum of the semantic vectors of the lexomes CAT and PLURAL, $\overrightarrow{cat} + \overrightarrow{\textsc{plural}}$.

In LDL, form can be represented by different units. The study presented in Chapter 5 uses triphones to represent form, as previous studies (Milin, Feldman, et al. 2017; Baayen, Chuang, Shafaei-Bajestan, et al. 2019; Chuang et al. 2020) have shown that triphones capture the variability of neighbouring phonological information well for English. Triphones are sequences of three phones within a word form. They overlap and can be understood as proxies for phonetic transitions. The cue matrix C encodes the forms of words in a binary fashion, giving information on which triphones are part of which word. In each word's individual form vector $\vec{c}$, the presence of a triphone is marked with 1, while the absence is marked with 0. The cue vectors of all words of a set of words constitute its C matrix and each row in such a C matrix represents a word form, while the columns of the C matrix represent all triphones of its underlying word set.

Meaning is contained within the semantic matrix S, which consists of semantic vectors of word forms on basis of their corresponding lexomes. Thus, the semantic vector $\vec{s}$ in S for a simplex word is identical to its corresponding lexome, while the semantic vector $\vec{s}$ in S for a complex word is the sum of its corresponding lexomes, e.g. $\overrightarrow{apple} + \overrightarrow{\textsc{plural}}$ for *apples* (Baayen, Chuang, Shafaei-Bajestan, et al. 2019). Semantic vectors of lexomes can be derived in different ways (e.g. Landauer & Dumais 1997; Jones & Mewhort 2007; Shaoul & Westbury 2010; Mikolov et al. 2013).

Once matrices for form and meaning are established, one can make use of linear mappings to compute comprehension and production. In LDL, comprehension refers to a model that has form vectors as input and semantic vectors as output. I illustrate the C matrix of a set of words with a toy lexicon containing the words *cat*, *bus*, and *eel* in Equation 3.1. Here, the DISC keyboard phonetic alphabet (the "Distinct Single Character" representation introduced by Burnage 1988) is used for triphone representation. Word boundaries are marked by the # symbol.

$$C = \begin{array}{c} \\ cat \\ bus \\ eel \end{array} \begin{array}{c} \begin{array}{cccccccc} \#k\{ & k\{t & \{t\# & \#bV & bVs & Vs\# & \#il & il\# \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array} \tag{3.1}$$

For the same toy lexicon, suppose that the semantic vectors for these three words are the row vectors of the following S matrix:

$$S = \begin{array}{c} \\ cat \\ bus \\ eel \end{array} \begin{array}{c} \begin{array}{ccc} cat & bus & eel \end{array} \\ \begin{pmatrix} 1.0 & 0.2 & 0.5 \\ 0.4 & 1.0 & 0.1 \\ 0.2 & 0.3 & 1.0 \end{pmatrix} \end{array} \tag{3.2}$$

To map forms onto meanings one needs a transformation matrix F, such that

$$CF = S \tag{3.3}$$

The transformation matrix F is straightforward to obtain. Let C' denote the Moore-Penrose generalised inverse[4] of C, available in R as the `ginv` function of the `MASS` package (Venables & Ripley 2002). Then,

$$F = C'S \tag{3.4}$$

For the toy lexicon example,

$$F = \begin{array}{c} \\ \#k\{ \\ k\{t \\ \{t\# \\ \#bV \\ bVs \\ Vs\# \\ \#il \\ il\# \end{array} \begin{array}{c} \begin{array}{ccc} cat & bus & eel \end{array} \\ \begin{pmatrix} 0.33 & 0.06 & 0.16 \\ 0.33 & 0.06 & 0.16 \\ 0.33 & 0.06 & 0.16 \\ 0.13 & 0.33 & 0.03 \\ 0.13 & 0.33 & 0.03 \\ 0.13 & 0.33 & 0.03 \\ 0.10 & 0.15 & 0.50 \\ 0.10 & 0.15 & 0.50 \end{pmatrix} \end{array} \tag{3.5}$$

[4] The inverse of a matrix needs not exist, rendering such a matrix a singular one. Most matrices used in LDL implementations are singular matrices. Thus, an approximation of the inverse must be used instead of an inverse itself. One such approximation is the Moore-Penrose generalised inverse (Moore 1920; Penrose 1955).

with CF being exactly equal to S in this simple example. That is, taking form vectors as input for the prediction of semantic vectors as output, i.e. solving $\hat{S} = CF$, this toy example correctly predicts 100% of all (three) words' semantics, i.e. $\hat{s}_i = s_i$. In more complex cases, semantic vectors are only approximately identical, thus, for a word i and its predicted semantic vector s_i, comprehension is successful if $\hat{s}_i$ shows the highest correlation with the targeted semantic vector s_i (Baayen, Chuang, Shafaei-Bajestan, et al. 2019). Following this method, one can report the percentage of comprehension accuracy.

Production as modelled in LDL takes semantic vectors as input and delivers form vectors as output. Using the same toy lexicon as before, I adapt its C matrix, i.e. I borrow the notation by Baayen, Chuang, Shafaei-Bajestan, et al. (2019) and henceforth call it T as it contains the Targeted triphones. For production, the transformation matrix G is of interest. Similar to F for comprehension, it is straightforward to obtain. Let S' denote the Moore-Penrose generalised inverse of S. Then,

$$G = S'T \tag{3.6}$$

Given G, one can then predict the triphone matrix $\hat{T}$ from the semantic matrix S by solving

$$\hat{T} = SG \tag{3.7}$$

For the toy lexicon example, the G transformation matrix is

$$G = \begin{array}{c} \\ cat \\ bus \\ eel \end{array} \begin{array}{cccccccc} \#k\{ & k\{t & \{t\# & \#bV & bVs & Vs\# & \#il & il\# \\ \end{array}$$

$$G = \begin{array}{r} \\ cat \\ bus \\ eel \end{array}\!\!\begin{array}{c} \begin{array}{cccccccc} \#k\{ & k\{t & \{t\# & \#bV & bVs & Vs\# & \#il & il\# \end{array} \\ \begin{pmatrix} 1.14 & 1.14 & 1.14 & -0.06 & -0.06 & -0.06 & -0.56 & -0.56 \\ -0.44 & -0.44 & -0.44 & 1.05 & 1.05 & 1.05 & 0.12 & 0.12 \\ -0.09 & -0.09 & -0.09 & -0.30 & -0.30 & -0.30 & 1.08 & 1.08 \end{pmatrix} \end{array} \tag{3.8}$$

As this is a toy example, SG is identical to T. For more complex cases, $\hat{T}$ will not be virtually identical to T "but will be an approximation of it that is optimal in the least squares sense" (Baayen, Chuang, Shafaei-Bajestan, et al. 2019: 21). Triphones with the strongest support are expected to be the triphones making up a word's form. As triphones are not ordered, it is also checked whether the sequence of phones can be constructed correctly. Both checking triphone support and sequence are conveniently done by the functions of the `WpmWithLdl` package (Baayen, Chuang & Heitmeier 2019). Following this method, one can report the percentage of production accuracy.

Figure 3.1 summarises the mapping between form and meaning by the *F* and *G* transformation matrices for comprehension and production modelling.

$$\begin{tikzcd}[column sep=huge]
C = \begin{pmatrix}1 & 0\\ 0 & 1\\ 1 & 0\end{pmatrix} \arrow[r, shift left=2, "{F = \begin{pmatrix}0.5 & 1.0\\ 0.1 & 0.2\end{pmatrix}}"] & S = \begin{pmatrix}0.5 & 1.0\\ 0.1 & 0.2\\ 0.9 & 0.7\end{pmatrix} \arrow[l, shift left=2, "{G = \begin{pmatrix}-1.22 & -0.24\\ 1.57 & 0.31\end{pmatrix}}"]
\end{tikzcd}$$

Figure 3.1: Illustration of mapping between *C* and *S* matrix via *F* (i.e. comprehension), and *S* and *C* matrix via *G* (i.e. production). Note: In production, *C* is referred to as *T*.

4 Production of word-final /s/

As explained in detail in Section 2.1, the present production study investigates the potential durational differences between three types of word-final /s/: non-morphemic /s/, plural /s/, and clitic /s/ (with the auxiliaries *is* and *has*).[1] Pseudowords are used as items to prevent potential lexical effects to confound findings (see Section 3.1.1). Three hypotheses derived from theories and models of speech production are examined. H PROD$_1$, the *Feed-Forward Hypothesis*, assumes that there is no durational difference between different types of word-final /s/. According to H PROD$_2$, the *Prosodic Hypothesis*, non-morphemic /s/ is shorter than plural /s/, and plural /s/ is shorter than auxiliary clitic /s/. H PROD$_3$, the *Emergence Hypothesis*, assumes that there are durational differences between different types of word-final /s/, but does not indicate what the nature of these durational differences is.

4.1 Methodology

4.1.1 Speakers and recordings

Forty native speakers of Southern British English took part in the experiment. Their mean age was 28.7 years, ranging from 19 to 58. Eight speakers were bi- or multilingual, and twenty-five speakers were from London while the other fifteen speakers were from other places in South Britain. None of the participants had a background in linguistics.

The recordings took place at Chandler House, University College London. The acoustic data were recorded on a computer with a Røde NT1-A microphone using an RME Fireface UC audio interface and sampled at 44.1 kHz, 16 bit.

4.1.2 Materials

For the production experiment, the pseudoword paradigm by Berko Gleason (1958) was adopted. Following her reasoning, it was assumed that phonetic effects

[1] An earlier version of this chapter has been published as part of Schmitz, Baer-Henney, et al. (2021).

found in pseudoword paradigms mirror linguistic reality. The pseudowords used in the production experiment consist of the full set of pseudowords discussed in detail in Section 3.1.2. For reasons of convenience, Table 4.1 lists these pseudowords once more.

Table 4.1: Orthographic representations of all pseudowords.

gli-	pru-	plee-	cloo-	blou-	glai-
glip	*prup*	*pleep*	*cloop*	*bloup*	*glaip*
glit	*prut*	*pleet*	*cloot*	*blout*	*glait*
glik	*pruk*	*pleek*	*clook*	*blouk*	*glaik*
glif	*pruf*	*pleef*	*cloof*	*blouf*	*glaif*
glips	*prups*	*pleeps*	*cloops*	*bloups*	*glaips*
glits	*pruts*	*pleets*	*cloots*	*blouts*	*glaits*
gliks	*pruks*	*pleeks*	*clooks*	*blouks*	*glaiks*
glifs	*prufs*	*pleefs*	*cloofs*	*bloufs*	*glaifs*

To elicit the pertinent types of /s/ under investigation, i.e. non-morphemic, plural, and *is*- and *has*-clitic /s/, 48 contexts and accompanying questions for /s/ elicitation were created. The verbs directly following the pseudowords in these contexts were chosen in such a way that out of twelve verbs in total, three each started with a voiceless plosive (/pl/, /k/), a vowel (/ɑ/, /iː/, /ʌ/, /eɪ/), a nasal (/m/, /n/), and an approximant (/w/, /l/, /r/). This was done to control for possible coarticulatory effects of these segmental classes with the preceding /s/. Examples are given in (1) to (4) with pseudowords and verbs in italics (see the supplementary material given in Chapter 11 for all contexts).

(1) Every day, the *glips plays* with the cloops.

(2) Two days ago, the *glips ate* their lunch together.

(3) Tonight, the *glip's meeting* the cloop for a drink.

(4) The *glip's written* a love letter to the cloop.

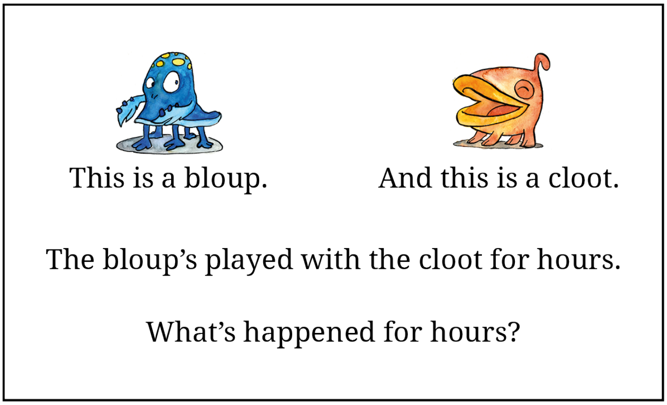
To keep priming effects to a minimum, pseudowords were split into two groups. Each group consisted of 24 pseudowords, with 12 pseudowords used for morphemic /s/ elicitation and 12 pseudowords used for non-morphemic /s/ elicitation. This way it was ensured that no single participant encountered a phonologically identical pseudoword as both mono- and multi-morphemic, i.e. no participant was to encounter /glɪps/ as both singular and plural or clitic item. Participants were distributed equally across both groups. Each participant was supposed to produce 12 tokens for each of the four types of /s/ (non-morphemic, plural, *is*-clitic, *has*-clitic; 48 tokens overall).

To ensure that each pseudoword was elicited within each context, i.e. with each verb for each type of /s/, twelve pseudorandomised lists were created. The same twelve lists were used for both groups to keep them comparable. Additionally, types of /s/ were alternated in such a way that no type of /s/ was elicited twice in a row. This was done to keep priming effects to a minimum.

4.1.3 Procedure

First, participants were introduced to the idea of a recently discovered far away planet. They were told that the inhabitants of this planet at first might appear bizarre, but engage in activities known to the participants, and not to worry about the unfamiliar names of the creatures. Second, the trial structure was explained, i.e. for each slide there would be pictures and names of alien creatures, a short explanation of a situation, and a question relevant to the situation which was to be answered aloud. Participants were then told to proceed in a natural pace and to take as much time as necessary to read and understand the aliens' names as well as the situations. To avoid possible confusion due to the simplicity of the task at hand, participants were made to believe that they were part of a control group of an experiment originally designed for children. Before starting practice trials, participants were reminded to use the aliens' names instead of pronouns when answering. Then, a practice set of four contexts (see the supplementary material given in Chapter 11) was used to familiarise the participants with the experimental procedure itself.

For each trial, the screen proceeded similarly (see Figure 4.1 as well as examples (5) to (8)): First, the relevant pseudowords were introduced. In the stimuli testing the plural, one pseudoword (in its plural form) was introduced, while in the other three conditions two different pseudowords were introduced. In either case, two images (van de Vijver & Baer-Henney 2014) representing the pseudowords were used to create familiarity with the items under investigation. In all cases but plural, two images of different creatures were given, while in plural contexts two images of the same creature were used. The pseudowords and images were paired randomly across lists to rule out possible confounding effects of appearance, e.g. due to the *bouba*/*kiki* effect (e.g. Köhler 1929; Fort et al. 2015). Second, a context was introduced. Third, a question was given to elicit an answer with the pertinent type of /s/ while the context slowly faded out. The fading out of the question forced the participants not to rely on the reading-aloud of the given context. This open format was chosen in order to elicit speech that is as natural as possible. By



This is a bloup.

And this is a cloot.

The bloup’s played with the cloot for hours.

What’s happened for hours?

Figure 4.1: Item, context, and question display during the production experiment.

choosing such an open format one obviously runs the risk of eliciting a large proportion of responses that do not contain the desired forms. This drawback of the experimental design was countered by having a large number of trials and participants. This strategy resulted in a sufficient number of observations. The experiment was carried out in a self-paced fashion; participants were instructed to progress in a contextually appropriate manner and at a speaking rate they considered to be normal.

(5) non-morphemic context
Introduction: This is a glaits. # And this is a pleeps.
Context: Every day, the glaits plays with the pleeps.
Question: What happens every day?
Answer: The glaits plays with the pleeps.

(6) plural context
Introduction: This is a glait. # And this is another one.
Context: Two days ago, the glaits ate their lunch together.
Question: What happened two days ago?
Answer: The glaits ate their lunch together.

(7) *is*-clitic context

Introduction:	This is a glait. # And this is a pleep.
Context:	Tonight, the glait's meeting the pleep for a drink.
Question:	What's happening tonight?
Answer:	The glait's meeting the pleep for a drink.

(8) *has*-clitic context

Introduction:	This is a glait. # And this is a pleep.
Context:	The glait's written a love letter to the pleep.
Question:	What's happened?
Answer:	The glait's written a love letter to the pleep.

4.1.4 Labels and measurements

In a first step, all recordings were manually transcribed on the utterance level. Using the freely available WebMAUS Basic system (Schiel 1999; Kisler et al. 2017), a phonetic transcription and segmentation based on the manual transcription was created. This automated segmentation was then manually checked by six trained annotators using the software Praat (Boersma & Weenink 2019). Boundaries marking the beginning of an item or /s/ were moved to the nearest zero crossing where both spectrogram and waveform indicated the initiation of the gesture for the respective segment, following laid out segmentation criteria based on features of specific sounds as described in the phonetic literature (e.g. Ladefoged 2003). In the case of /s/, the boundaries were set to the zero crossing closest to the onset and offset of the friction visible in the waveform (see Figure 4.2). If a pause followed the /s/, the boundary was set to the point where the friction of the /s/ dropped to silence.

The reliability of the segmentation criteria was verified by trial segmentations, in which it was ensured that all annotators placed boundaries with only very small variations. Each annotator worked on a disjoint set of items; segmentation criteria were regularly re-verified in meetings of the annotators. After the segmentation process, a Praat script was used to extract the item, its phonetic transcription, and its duration, as well as the /s/ duration itself. If applicable, the duration of the following pause was also extracted. Additionally, the preceding and the following word were extracted as well.

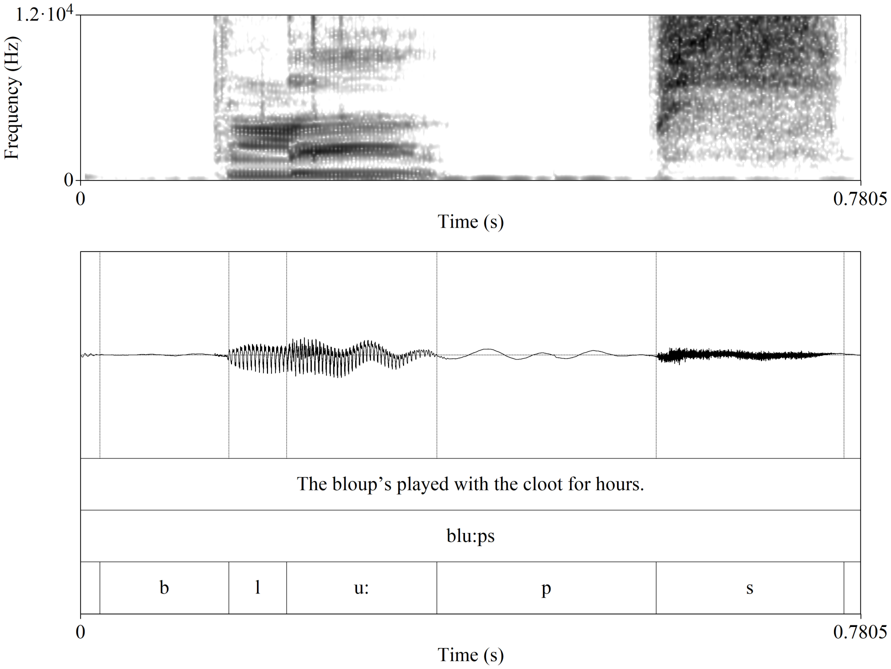


Figure 4.2: Example acoustic analysis of the item *bloup’s*.

4.1.5 Pre-processing

A part of the 1,920 (40 participants × 48 utterances) recorded data points had to be excluded from analysis for one or more of the following reasons. If an utterance did not include a word-final /s/, this utterance was discarded (n = 599). A high number of failures to produce final /s/ was expected especially with the clitics since participants could use a different tense form, or the full form of the auxiliary. It was also expected that participants would produce wrong pronunciations (including those with the final /s/) of the newly encountered written word-forms, as the participants had to retrieve them from short-term memory after the fading out of the context. Additionally, utterances containing stutter or hesitation (n = 29) or replacement of pseudowords by pronouns (n = 15) were excluded as well. Some utterances were ungrammatical (n = 9), while other utterances contained pseudowords that were not part of the original set of pseudowords (n = 8). Cases where the interpretation of the final /s/ was ambiguous presented another problem (n = 114). An example of such a case is given in (9) where a *has*-clitic was expected. Note that two pseudowords without a non-morphemic word-final /s/

were introduced, while either a non-morphemic /s/ or a *has*-clitic /s/ was produced for the item under investigation, and most likely a non-morphemic word-final /s/ for the second pseudoword. As for regular inflected verbs there was no way to decide which type of /s/ had been produced in such cases, such utterances were discarded.

(9) ambiguous case example
Introduction: This is a glait. # And this is a pleep.
Context: The glait's attended concerts with the pleep many times.
Question: What's happened many times?
Answer: The glaits attended many concerts with the pleeps many times.

After exclusions, 1,146 data points (approx. 60%) remained in the final data set. The final data set as well as the analysis and results discussed in the following sections can be found in the supplementary material given in Chapter 11.

4.2 Analysis

4.2.1 Covariates

The set of covariates chosen for the present study is similar to that of other studies on phonetic effects of morphological structure (Pluymaekers et al. 2005a,b; Hanique, Ernestus, et al. 2013; Plag et al. 2017). In the following, covariates used as fixed effects are described first. Then, variables used as random effects are introduced.

baseDurLog. Indicating a more local speaking rate (e.g. Plag et al. 2017), base duration was measured. Base duration in this case is equal to the summed duration of all word-internal segments preceding the /s/ under investigation. That is, the base of multi-morphemic items and the segmental string without the final /s/ of mono-morphemic items is henceforth considered the base. The base duration was log-transformed and centred (Robinson & Schumacker 2009; Afshartous & Preston 2011; Winter 2019). This variable is called baseDurLog.

biphoneProb. A potential problem with using pseudowords is their phonotactics. Pseudowords created for this book are mostly phonotactically legal (see Section 3.1.2 and the relevant footnote therein), and their final consonant clusters (with /s/ as the second consonant) are not uncommon in multi-morphemic words. However, in mono-morphemic words these clusters are rarer, or, in the case of

/fs/, even unattested (e.g. in CELEX, Baayen et al. 1995). The different phonotactic probabilities of these clusters could potentially influence the pronunciation of /s/ in the pseudowords, especially when spoken in the contexts where these words receive a mono-morphemic interpretation. To address this concern, the probability of the final biphones /fs/ (0), /ks/ (0.00427), /ps/ (0.00058), and /ts/ (0.00072) in mono-morphemic words was included as a covariate. biphoneProb was computed on the basis of the transcriptions of all mono-morphemic words in CELEX.

biphoneProbSum & biphoneProbSumBin. A potential factor influencing the duration of a word in running speech is its predictability in context. The more predictable, the shorter the duration (Pluymaekers et al. 2005a; Bell et al. 2009; Torreira & Ernestus 2009). Such a word bigram frequency, however, is not applicable to pseudowords for obvious reasons. Instead, the summed biphone probability was used analogously as a comparable measure. The summed biphone probability for each pseudoword and its phonological variants was calculated using the Phonotactic Probability Calculator (Vitevitch & Luce 2004). Additionally, a binary covariate based on the summed biphone probability was created. The threshold for low versus high summed biphone probability for biphoneProbSumBin was the mean of the continuous covariate. That is, all values below the mean were considered to be low, while all values above the mean were taken as high.

folSeg & folType. To account for potential effects of the following word on the duration of /s/ (cf. Klatt 1976; Umeda 1977), these were included in regard to their onset segment adjacent to the word-final /s/. This segment was included in its phonological representation in folSeg (e.g. k for the onset of *cooked*) as well as in its segmental class by folType (i.e. approximant APP for *listen*, fricative F for *find*, nasal N for *know*, plosive P for *cook*, vowel V for *eat*).

gender / location / monoMultilingual. Participants' gender and whether they had grown up in London or elsewhere in South Britain (location) were included as well as they may influence phonetic realisations. Additionally, participants who were early bilinguals (i.e. the L2 was/the L2s were acquired as a pre-school child) were categorised as multilingual, while all other participants were categorised as monolingual in monoMultilingual.[2]

[2] Psycholinguistic experiments are standardly done with monolingual speakers (mostly of English, and mostly in the US). In the multicultural context of a large European city like London, experiments with student populations necessarily involve speakers that are multilingual (with varying degrees of competence). To control for this potential confound, the variable monoMultilingual was added. While there are studies of phonetic duration in bilingual speech (e.g. Mack 1982; Lee & Iverson 2012) the effect of mono-/multilingualism on the duration of word-final /s/ has not been explored yet.

NEIGHBOURHOODDENSITY & NEIGHBOURHOODFREQUENCY. The densities and frequencies of neighbourhoods were included as covariates as the number of neighbours may influence phonetic reduction (e.g. Gahl et al. 2012). Both neighbourhood measures were taken from the CLEARPOND database (Marian et al. 2012). That is, NEIGHBOURHOODDENSITY describes the number of words differing in one segment from the item in question (Marian et al. 2012: 3), while NEIGHBOURHOODFREQUENCY describes the mean frequency (per million) of these neighbouring words.

PAUSEDUR & PAUSEBIN. In order to account for final-lengthening effects, all stretches of silence between the offset of the word-final /s/ and the onset of the following word were measured. Silence of 50 ms and above was considered as pause (Lee & Oh 1999; see also Zvonik & Cummins 2003 and Krivokapić 2007 on short pause duration in between short phrases). The closure durations of following plosives were taken into account by subtracting the mean closure duration of the pertinent plosive (mean values for /p, t, k/ adopted from Yao 2007) from the measured stretch of silence. It was considered a pause only if the resulting duration was above the aforementioned threshold. Pause measurements were included as the continuous variable PAUSEDUR as well as the binary variable PAUSEBIN (with the levels pause and no_pause).

PREC. It has been shown that the consonant preceding word-final /s/ may influence the duration of word-final /s/ (e.g. Umeda 1977). In particular, Umeda (1977: 853) finds that /s/ becomes shorter after plosives, and longer after the fricative /θ/ (and this presumably also holds for /s/ after the fricative /f/). The consonant preceding the final /s/ was therefore included as a covariate, PREC.

SPEAKINGRATE. As speaking rate is a self-evident variable affecting segment durations, this was controlled for. The speaking rate was computed as the number of syllables in an utterance divided by the duration of the utterance. For the statistical analysis, SPEAKINGRATE was centred (Robinson & Schumacker 2009; Afshartous & Preston 2011; Winter 2019). The computation was done automatically in Praat (de Jong & Wempe 2008). This way of computing speaking rate is similar to that utilised in previous studies (e.g. Plag et al. 2017).

ITEM & TRANSCRIPTION. Pseudowords were sometimes produced with varying segmental make-up. Therefore, both the orthographic representation of the pseudoword and a phonological transcription of the word as spoken were included as variables. These covariates were labelled ITEM and TRANSCRIPTION.

LIST & SLIDENUMBER. To account for possible durational differences due to priming and similar effects, the list number (1 to 12) and the point of occurrence during the experiment of the individual item were also included.

speaker / age. speaker ID was included to account for inter-speaker differences in production. age was included as well, as it may show an influence on phonetic realisations.

4.2.2 Overview of the data

An overview of all variables and their distribution is given in Table 4.2 and Table 4.3.

Table 4.2: Summary of categorical predictors and the explanatory variable of interest in the final data set.

Categorical predictors	Levels
item	48
transcription	67
NeighbourhoodDensity	0: 419 1: 238 2: 165 3: 107 4: 14 5: 114 6: 32 7: 30
pauseBin	no: 777 yes: 342
biphoneProbSumBin	low: 856 high: 263
list	24
slideNumber	48
preC	f: 273 k: 292 p: 281 t: 273
folSeg	18
folType	APP: 229 F: 12 N: 230 P: 300 V: 278
speaker	40
gender	2
location	London: 636 elsewhere: 483
monoMultilingual	monolingual: 871 multilingual: 248
Explanatory variable	Levels
typeOfS	nm: 308 pl: 373 is: 284 has: 154

4.2.3 Collinearity

As described in Section 3.2.3, one issue to address when fitting a model to a multitude of similar covariates is collinearity (e.g. Tomaschek, Hendrix, et al. 2018). To avoid such issues, covariates were tested for correlation issues. High correlation coefficients, i.e. $|rho| \geq 0.5$, were found for item and transcription ($rho = 0.82$, $p < 0.001$, Spearman), pauseDur and pauseBin ($rho = 0.87, p < 0.001$, Spearman), neighbourhoodDensity and neighbourhoodFrequency ($rho = 0.86$, $p < 0.001$, Spearman), biphoneProbSum and biphoneProbSumBin ($rho = 0.87$,

Table 4.3: Summary of the dependent variable and numerical predictors in the final data set.

Dependent variable	Mean	St. Dev.	Min	Max
sDurLog	0.002	0.388	-1.201	1.098
Numerical predictors	Mean	St. Dev.	Min	Max
speakingRate	0.000	0.899	2.250	3.540
baseDurLog	-1.235	0.240	-1.987	-0.375
pauseDur	0.072	0.193	0.000	3.559
neighbourhoodFrequency	27.345	84.645	0.000	412.027
biphoneProbSum	0.013	0.007	0.005	0.031
biphoneProb	0.001	0.002	0.000	0.004
age	28.740	9.743	19.000	58.000

$p < 0.001$, Spearman), and for folSeg and folType ($rho = -0.74, p < 0.001$, Spearman).

Given the nature of the highly correlated variable pairs, that is both variables tap into very similar features of the given items or utterances, it was decided to make use of the competitive exclusion strategy outlined in Section 3.2.3. This procedure led to the exclusion of item (in favour of transcription), pauseDur (in favour of pauseBin), neighbourhoodFrequency (in favour of neighbourhoodDensity), biphoneProbSum (in favour of biphoneProbSumBin), folSeg (in favour of folType), and biphoneProb (in favour of preC).

4.2.4 Statistical analysis

Differences in consonant duration may play out as differences in absolute duration or as differences in relative duration (e.g. with gemination: Oh & Redford 2012; Ridouane & Hallé 2017; Ben Hedia 2019). Some previous analyses of the duration of /s/ (Plag et al. 2017) have therefore looked at both absolute and relative duration, and the present study will also present these two types of analyses. In the first analysis (Section 4.3.1) absolute duration of /s/ was used as the dependent variable, whereas in the second analysis (Section 4.3.2) the duration of /s/ relative to the duration of the whole word was used as the dependent variable. Relative duration (i.e. the variable proportionOfS) was calculated by dividing the absolute duration of the /s/ by the duration of the whole word.

The dependent variable, duration of /s/, was log-transformed and centred following standard procedures to reduce the potentially harmful effect of skewed distributions in linear regression models (e.g. Winter 2019). The name of this variable is sDurLog. proportionOfS did not have a skewed distribution and no transformation was necessary. Following the modelling procedure for LMER models outlined in Section 3.2.1, models for sDurLog and proportionOfS as dependent variables were fitted, tested for collinearity issues by using variance inflation factors, and finally trimmed. This resulted in a loss of 9 data points (0.8%) for sDurLog and in a loss of 12 data points (1.0%) for proportionOfS, and in both cases led to a satisfactory distribution of the residuals.

4.3 Results

4.3.1 Absolute duration

Figure 4.3 shows the distribution of the observed durations of non-morphemic, plural, *is*-, and *has*-clitic /s/. On average, non-morphemic /s/ duration is 134 ms, which is about 13 ms longer than plural /s/ with a mean duration of 121 ms. The mean duration of the *is*-clitic is 103 ms and the mean duration of the *has*-clitic is 94 ms.

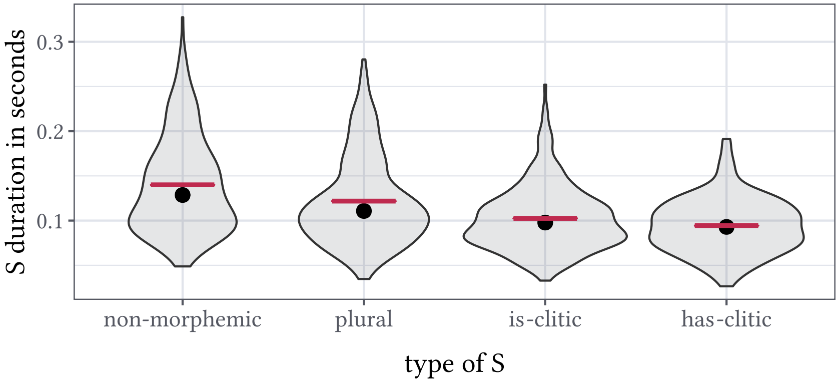


Figure 4.3: Observed durations of non-morphemic, plural, *is*- and *has*-clitic /s/. The dot represents the median, the horizontal line indicates the mean. The violin shapes represent rotated density plots describing the distribution of the data.

Multivariate analyses as described in the previous section were then conducted to control for the many potentially intervening influences of the described covariates listed in Section 4.2.1. In the final model, fitted according to the procedure described above, main effects of type of /s/ (typeOfS), speaking rate (speakingRate), base duration (baseDurLog), pause (pauseBin), preceding consonant (preC), biphone probability sum (biphoneProbSumBin), following segmental type (fol-Type), and mono-/multilingualism (monoMultilingual) were found.

Regarding the random effects, only speaker-specific random intercepts turned out to significantly improve the model fit. The *p*-values for the analysis of variance of the final model are given in Table 4.4.

Table 4.4: *p*-values of fixed effects in the final model, fitted to the log-transformed durations of /s/.

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr (F)
typeOfS	5.312	1.771	3	1089.66	33.338	0.000
speakingRate	0.230	0.230	1	1117.09	4.324	0.038
baseDurLog	9.466	9.466	1	1079.58	178.220	0.000
pauseBin	6.970	6.970	1	1110.28	131.235	0.000
biphoneProbSumBin	0.398	0.398	1	1082.26	7.492	0.006
preC	0.623	0.208	3	1080.29	3.910	0.009
folType	2.677	0.669	4	1081.55	12.598	0.000
monoMultilingual	0.345	0.345	1	37.37	6.498	0.015

The final model was then analysed in terms of its R^2 values which were computed with the `MuMIn` package (Barton 2020; for marginal and conditional R^2 value computation, see Nakagawa et al. 2017). The marginal R^2 value of a model indicates the percentage of variation in the data explained by the fixed effects of that model. The variance explained by the entire model is given by its conditional R^2 value. The marginal R^2 value of the model is 0.46, that is, fixed effects explain 46% of the variation in the data. The variance explained by the entire model is 61% as obtained by the conditional R^2 value of 0.61.

The estimates of the final model and their *p*-values are given in Table 4.5. The reference levels for the categorical predictors are: for typeOfS it is `non-morphemic` /s/, for pauseBin it is `no-pause`, for biphoneProbSumBin it is `low`, for preC it is `t`, for folType it is `approximant`, and for monoMultilingual it is `monolingual`. All coefficients can be interpreted as changes relative to these reference levels.

The predictor strength of individual predictors was checked following the method outlined in Section 3.2.1, that is by fitting models that lacked a particular predictor and comparing their marginal R^2 values to those of the final model. The results are reflected in the hierarchy given in (10). The decrease in R^2 is greatest when removing baseDurLog, followed by pauseBin, and so forth. Overall, the

Table 4.5: Fixed-effect coefficients and *p*-values as computed by the final model (mixed-effects model fitted to the log-transformed and centred durations of /s/).

	Estimate	SE	df	*t*-value	Pr(|t|)
(Intercept)	-1.321	0.068	550.378	-19.498	0.000
typeOfSpl	-0.114	0.019	1094.00	-6.062	0.000
typeOfSis	-0.178	0.020	1096.00	-8.839	0.000
typeOfShas	-0.196	0.024	1091.00	-8.140	0.000
speakingRate	-0.021	0.010	1117.00	-2.079	0.038
baseDurLog	0.586	0.044	1080.00	13.35	0.000
pauseBinpause	0.206	0.018	1110.00	11.456	0.000
biphoneProbSumBinhigh	0.047	0.017	1082.00	2.737	0.006
preCf	0.061	0.020	1081.00	-3.044	0.003
preCk	0.055	0.020	1082.00	-0.303	0.006
preCp	0.050	0.020	1079.00	2.522	0.012
folTypeF	0.012	0.070	1084.00	0.171	0.864
folTypeN	-0.036	0.021	1079.00	-1.764	0.078
folTypeP	-0.045	0.019	1080.00	-2.384	0.017
folTypeV	-0.136	0.020	1082.00	-6.85	0.000
monoMultilingualmultilingual	-0.152	0.059	37.37	-2.549	0.015

morphological status of an /s/ appears to be a strong predictor of its acoustic duration.

(10) baseDurLog » pauseBin » typeOfS » monoMultilingual » folType » speakingRate » biphoneProbSumBin » PreC

Figure 4.4 shows the effect of the numerical variables included in the final model on /s/ duration. The estimated values of the dependent variable and the base duration are back-transformed into seconds. Speaking rate and base duration show effects in the expected direction. With faster speech, /s/ becomes shorter (Panel A), while longer base durations also come with longer /s/ durations (Panel B).

The partial effects of the categorical variables included in the final model are illustrated in Figure 4.5. /s/ duration is longer if the /s/ is followed by a pause (Panel A), which can be interpreted as a clear case of phrase-final lengthening (e.g. Cooper & Danly 1981). Higher biphone probability sum leads to longer /s/ durations (Panel B). There is also an effect of the preceding consonant: The plosive /t/ is followed by significantly shorter /s/ durations than are /k/ and /f/ (Panel C). /s/ duration is significantly shorter when followed by a vowel, while all other differences between following consonants are minor in nature (Panel D). Lastly,

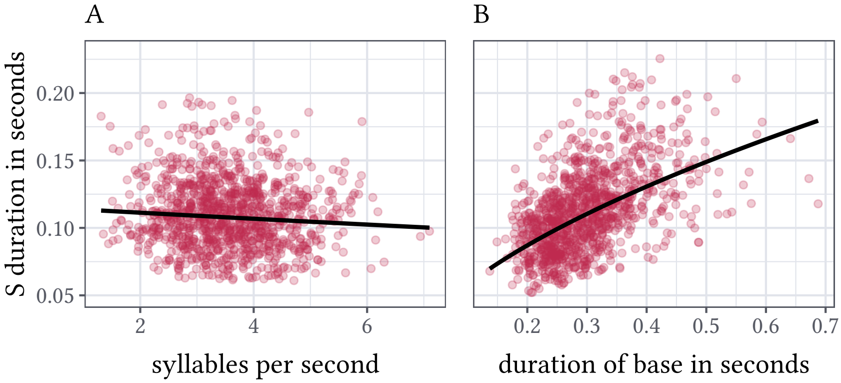


Figure 4.4: Partial effects of the numerical variables SPEAKINGRATE (Panel A) and BASEDURLOG (back-transformed, Panel B) included in the final model, fitted to the log-transformed values of duration of /s/.

monolingual speakers produce longer /s/ durations than multilingual speakers (Panel E).

The effect of the variable of interest, i.e. TYPEOFS, is plotted in Figure 4.6. As above, the values of the dependent variable are back-transformed into seconds.

One can see that there are durational differences between the different types of /s/. The results of pair-wise comparisons of the predicted means using Tukey contrasts (as implemented by the `SfL` package, Schmitz & Esser 2021) are summarised in Table 4.6.

Table 4.6: Multiple comparisons of means of duration of /s/ (Tukey contrasts). Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

			Estimate	SE	z-value	Pr(|z|)	
plural	-	non-morphemic	-0.114	0.019	-6.062	0.001	***
is-clitic	-	non-morphemic	-0.188	0.020	-8.839	0.001	***
has-clitic	-	non-morphemic	-0.196	0.024	-8.140	0.001	***
is-clitic	-	plural	-0.064	0.019	-3.294	0.005	**
has-clitic	-	plural	-0.082	0.023	-3.503	0.003	**
has-clitic	-	*is*-clitic	-0.018	0.023	-0.766	0.868	

Based on the Tukey tests, the comparison of the different types of /s/ yields the significant contrasts shown in Table 4.7. Considering the different durations given in Table 4.8, the following hierarchy emerges: non-morphemic /s/ > plural /s/ > *is*-/*has*-clitic /s/.

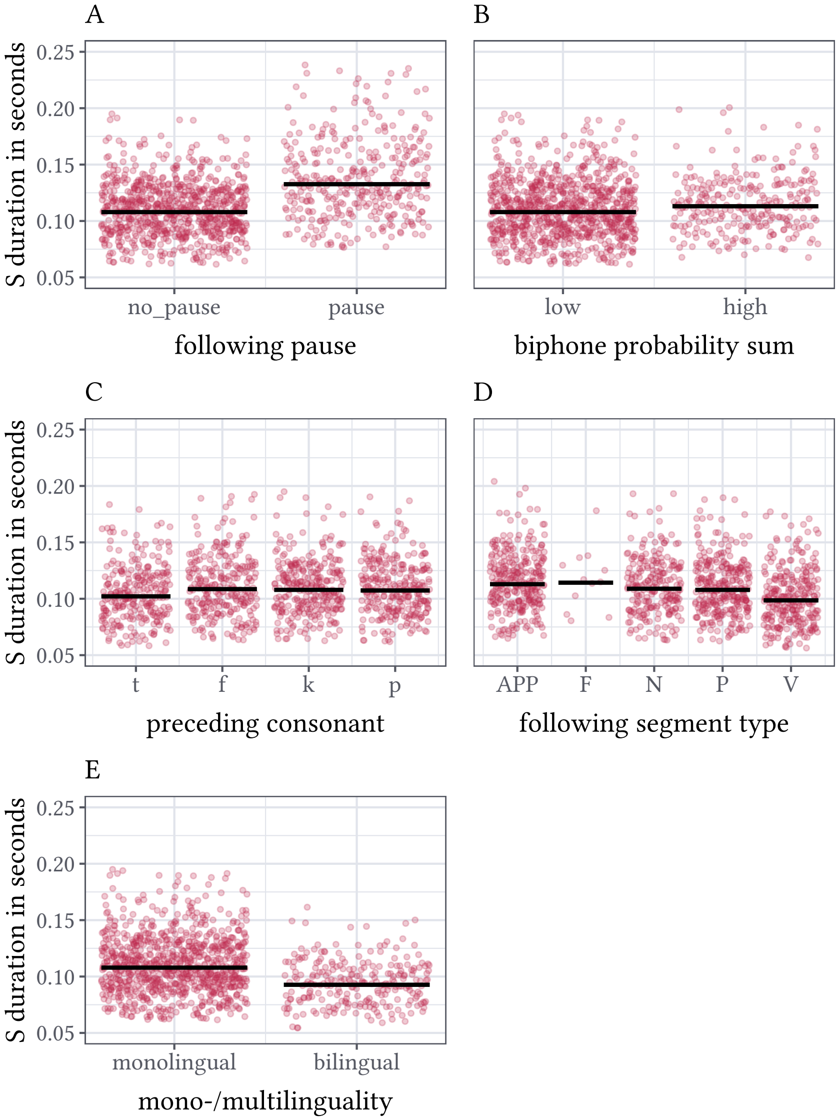


Figure 4.5: Partial effects of the categorical variables pauseBin (Panel A), biphoneProbSumBin (Panel B), preC (Panel C), folType (Panel D), and monoMultilingual (Panel E) included in the final model, fitted to the log-transformed values of duration of /s/.

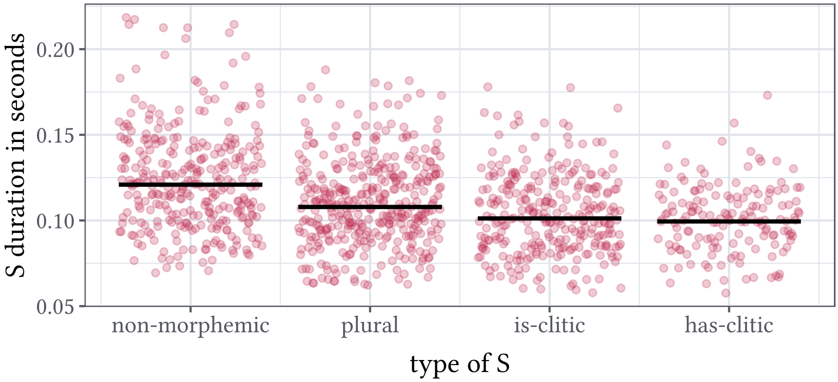


Figure 4.6: Partial effect of TYPEOFS in the final model, fitted to the log-transformed values of duration of /s/.

Table 4.7: Significant contrasts in duration between different types of /s/. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	nm	pl	*is*	*has*
non-morphemic	n.a.	***	***	***
plural		n.a.	**	**
is-clitic			n.a.	
has-clitic				n.a.

Table 4.8: /s/ durations as estimated by the final model using non-centred data. All values are back-transformed to seconds. Values given are estimated for items without following pause, high biphone sum probability, monolingual speakers, and across all preceding and following segment types.

TYPEOFS	Mean
non-morphemic	0.224
plural	0.200
is-clitic	0.187
has-clitic	0.184

To summarise, the durational differences between non-morphemic and other types of /s/, as well as the durational difference between plural and the clitics are significant, while there is no significant durational difference between the two clitics. Non-morphemic /s/ is longest in duration, followed by plural /s/, which in turn is followed by clitic /s/.

4.3.2 Relative duration

The results for relative duration are very similar to those of absolute duration. The *p*-values for the analysis of variance of the final model are given in Table 4.9. Table 4.10 shows the coefficients for the final model. All effects go in the same direction as in the analysis of absolute duration. The only predictors that have lost significance when compared to the model for absolute duration are preC and speakingRate. The differences in the means show the same pattern as in the analysis of absolute duration, as can be seen in Table 4.11.

Table 4.9: *p*-values of fixed effects in the final model, fitted to the relative durations of /s/.

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(F)
typeOfS	0.161	0.054	3	1070.68	25.510	0.000
pauseBin	0.186	0.186	1	1101.26	88.518	0.000
biphoneProbSumBin	0.015	0.015	1	36.32	6.917	0.012
folType	0.071	0.018	4	1063.31	8.389	0.000
monoMultilingual	0.010	0.010	1	37.81	4.561	0.039

4.4 Discussion

Following in the footsteps of previous studies on durational differences between different types of /s/, this study tested whether the morphological category of word-final /s/ has an influence on its acoustic duration in speech production. In order to avoid imbalanced data as in the case of corpus studies, speech material elicited by the means of highly controlled contexts of a production task was used. For the first time in this context, pseudowords instead of real words were used to minimise potentially confounding lexical effects. It was found that there are significant durational differences between non-morphemic and morphemic types of word-final /s/, with morphemic types of /s/ being significantly shorter in duration than non-morphemic /s/. Also, there are significant durational differences between the plural suffix and the *is*- and *has*-clitic /s/, with plural /s/ being

Table 4.10: Fixed-effect coefficients and p-values as computed by the final model (mixed-effects model fitted to the relative durations of /s/).

	Estimate	SE	df	t-value	Pr(|t|)
(Intercept)	0.299	0.007	89.73	45.827	0.000
TYPEOFSpl	-0.019	0.004	1085.00	-5.157	0.000
TYPEOFSis	-0.031	0.004	1070.00	-7.651	0.000
TYPEOFShas	-0.035	0.005	1067.00	-7.260	0.000
PAUSEBINpause	0.033	0.004	1101.00	9.408	0.000
BIPHONEPROBSUMBINhigh	0.013	0.005	36.32	2.630	0.012
FOLTYPEF	0.001	0.014	1068.00	0.086	0.931
FOLTYPEN	-0.006	0.004	1061.00	-1.409	0.159
FOLTYPEP	-0.007	0.004	1056.00	-1.708	0.088
FOLTYPEV	-0.022	0.004	1063.00	-5.568	0.000
MONOMULTILINGUALmultilingual	-0.024	0.011	37.81	-2.136	0.039

Table 4.11: Multiple comparisons of means of duration of /s/ (Tukey contrasts). Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

			Estimate	SE	z-value	Pr(|z|)	
plural	-	non-morphemic	-0.019	0.004	-5.157	0.001	***
is-clitic	-	non-morphemic	-0.031	0.004	-7.651	0.001	***
has-clitic	-	non-morphemic	-0.035	0.005	-7.260	0.001	***
is-clitic	-	plural	-0.011	0.004	-2.936	0.017	*
has-clitic	-	plural	-0.015	0.005	-3.300	0.005	**
has-clitic	-	*is*-clitic	-0.004	0.005	-0.854	0.827	

significantly longer than clitic /s/ and with no significant difference between the two clitics. Hence, the type of /s/ emerged as a strong, significant predictor of segmental duration.

The differences between different types of /s/ in the present study are completely in line with previous studies that were based on speech corpora and on different varieties of English (Plag et al. 2017 and Tomaschek et al. 2019 on North American English; Zimmermann 2016 on New Zealand English). In those studies the same pattern of differences was found. Turning to previous experimental studies, differing results were found. The results of both prior experimental studies (Walsh & Parker 1983; Seyfarth et al. 2017) are subject to potentially confounding effects of the lexical and contextual properties of the items under investigation. Their finding of non-morphemic /s/ being shorter than morphemic /s/ may well be an artefact of such properties. The items used in the present study, however, are much less prone to be subject to such effects as they are pseudowords with no established representations in the speakers' mental lexicons. The results on the duration of clitic /s/ cannot be compared to previously reported ones by other experimental studies, as none of the previously conducted experimental studies investigated clitic /s/ production.

No previous studies have used pseudowords either, so before turning to the theoretical interpretation of the results of the present study, a few words are in order on whether using pseudowords might have had an undesired impact on the results. While the use of pseudowords in phonetic experiments comes with a number of benefits (see Section 3.1.1), it also raises some questions. First, there is the issue of phonotactic probability raised in Section 4.2.1. Two measures concerned with phonotactics (one describing the phonotactic probability of the whole word, the other taking into consideration the consonant preceding the word-final /s/) were included in the statistical analysis to address this issue. It turned out that phonotactic probability influences the production of pseudowords, as it does for real words. Crucially, there was no interaction between the type of /s/ and the consonant preceding it in mono-morphemic words. This means that speakers produced these clusters in the same way, no matter whether the cluster occurred in the mono-morphemic words or whether the cluster straddled the morphemic boundary between the base and the /s/. The main effects of the phonotactic variables turned out to be rather weak and, crucially, were properly controlled for in the regression analysis. In sum, the phonotactics of the final cluster does not seem to have unduly influenced the results.

Second, there might have been a problem with another aspect of the phonological structure of the pseudowords in the experiment, i.e. long-distance agreement of phonological features (Coetzee 2009). Such effects of the Obligatory Contour Principle (OCP; Coetzee 2005) might have arisen with pseudowords such as *pleep* (in which initial /p/ and final /p/ share all features) or *glik* (in which the initial and final sounds /g/ and /k/ share the dorsal feature). Following the findings by Coetzee (2009), a new variable was coded to test this effect post-hoc as an additional covariate and as an interacting term of typeOfS with the following levels: `not well-formed` for pseudowords in which the initial and final consonant share all features (n = 836), `moderately well-formed` for pseudowords in which the initial and final consonant share the dorsal feature (n = 147), and `well-formed` for all remaining pseudowords (n = 145). There was no significant main effect of this variable on the duration of /s/, nor a significant interaction with typeOfS. OCP effects thus cannot explain the present results.

Third, after having carried out the experiments, it came to attention that some of the pseudowords have real word relatives that are spelled differently but are phonologically identical: *pleet(s)* corresponds to *pleat(s)*, *glits* corresponds to *glitz* (and no word corresponding to *glit*), and *glik* corresponds to the surname *Glick* (and no surname corresponding to *gliks*), whereas *glif(s)* corresponds to *glyph(s)*, which has a very low frequency and thus may constitute a pseudoword for most of the participants. These words might have unduly influenced the results and should perhaps not have been included into the statistical analysis. To check whether these items had any influence on the results, a data set was created containing all data but the four potentially offending items. Fitting the final model (as done in Section 4.2.4) to this new dataset resulted in basically the same findings, i.e. TYPEOFS was still a significant predictor for /s/ duration showing the same significant differences between non-morphemic, plural, and clitic items as presented in Table 4.7.

It has recently been shown that the notion of pseudoword is problematic in a more general way (also see Section 3.1.1). The notion of pseudoword itself is usually based on the idea of the lexicon as a community construct. When talking about the mental lexicon, however, it is clear that what is an existing word and what is an unknown pseudoword is a matter of the individual speaker's mental lexicon. All participants of the present experiment denied knowing any of the pseudowords used in this experiment when asked afterwards. At the community level, Google frequencies of pseudowords have been shown to be a robust predictor of reaction times in lexical decision tasks (e.g. Hendrix & Sun 2020). To test whether Google frequency had an effect on the present results, the covariate GOOGLEFREQ was created containing the number of Google search hits for each pseudoword. The addition of this covariate as either fixed effect or interacting term to TYPEOFS resulted in its exclusion during the model simplification procedure.

Finally, let us turn to the theoretical implications of the present results. What do these results mean for the three hypotheses that were tested? H PROD$_1$, the *Feed-Forward Hypothesis*, states that there is no durational difference between word-final non-morphemic /s/, plural /s/, and auxiliary clitic /s/. This hypothesis is rejected as carefully controlled evidence was provided that shows that the duration of /s/ varies by morphological category. This is an effect that present feed-forward models cannot accommodate, unless they would be refined in such a way that post-lexical processes can arise from certain kinds of lexical information. At present, no such refinement is available.

H PROD$_2$, the *Prosodic Hypothesis*, states that there are durational differences between different types of word-final /s/, with non-morphemic /s/ being shorter than plural /s/, and plural /s/ being shorter than the auxiliary clitic. While it is true that there are durational differences between the categories, the observed differences pattern in the opposite direction. The more integrated the /s/ is with the stem, the longer its duration. The *Prosodic Hypothesis* is correct in positing that the two auxiliary clitics should show no difference in duration. Overall, however, the *Prosodic Hypothesis* must be rejected, as the prosodic structure does not explain the most important patterning of the data.

Finally, H PROD3, the *Emergence Hypothesis*, states that there are durational differences between the different types of word-final /s/ under investigation. The fact that such differences were found means that these differences might emerge through the mechanisms posited by the theories underlying this hypothesis.

As mentioned in Section 2.1, Tomaschek et al. (2019) found that stronger support for a morphological function leads to a longer duration, that is as for the present findings, non-morphemic /s/ showed the longest duration, auxiliary clitic /s/ showed the shortest durations, and plural suffix /s/ duration was in between. This effect seems to run counter to the predictions of information-theoretic accounts and probabilistic theories, according to which words and segments are realised shorter when they are less informative (Aylett & Turk 2004; Jaeger 2010; Cohen Priva 2015). However, the enhancement effects are in line with studies showing that duration increases with increasing paradigmatic certainty (Kuperman et al. 2007; Cohen 2014; Bell et al. 2021; Tucker, Sims, et al. 2019). For instance, Kuperman et al. (2007) found that the duration of a given interfix in Dutch compounds increases with increasing probability of this interfix (as against its competitors) in the left constituent family of the compound.

Overall, it seems that simplistic approaches can neither explain the existence, nor the patterning of the durational differences one finds attested. The *Feed-Forward Hypothesis* is rejected because durational differences were in fact observed. The *Prosodic Hypothesis* is rejected because the observed durational differences pattern in a direction that is opposite to the one predicted. The *Emergence Hypothesis* is supported by the present findings as it proposes that durational differences of some nature should emerge between different types of /s/.

The results of the present study may bring up further questions. First, how can the aforementioned effects of morphological support, informativity, and paradigmatic probability be reconciled? This question is addressed further in Chapter 5, making use of linear discriminative learning (Baayen, Chuang, Shafaei-Bajestan, et al. 2019; Chuang et al. 2021). Second, assuming the durational differences found here and in previous studies are indeed systematic, one would also like to know whether language users are able to perceive them. This automatically leads to the question of whether all differences are perceptible or only some of them, given the knowledge on the perception of differences in fricative durations, i.e. that the threshold for perceptible durational differences appears to be at 25 ms (Klatt & Cooper 1975). This question is further investigated in Chapter 6. Third, if the durational differences are perceptible, another question naturally suggests itself: Do users of a language not only perceive but also make use of such differences in comprehension? This question is addressed in Chapters 7 and 8.

5 Modelling word-final /s/ with linear discriminative learning

The aim of the linear discriminative learning implementation presented in this chapter is to further investigate H PROD$_3$, the *Emergence Hypothesis*.[1] For the production study of Chapter 4, this hypothesis delivered a rather weak prediction: There are durational differences between different types of word-final /s/. Using an LDL implementation, the nature of these differences is further examined. That is, this study investigates whether measures derived from such an implementation are capable of explaining durational differences between different types of word-final /s/. If so, such measures will potentially provide insight into the underlying effects which lead to such durational differences.

5.1 Methodology

The methodology of the present investigation consists of two main stages. First, the implementation of the LDL network itself, including the selection of data to train the network (Sections 5.1.1 and 5.1.2) and the implementation of required matrices (Sections 5.1.3 to 5.1.5). Second, the extraction of several measures derived from the LDL implementation (Section 5.1.6), which are then used in the statistical analysis (Section 5.2).

5.1.1 The semantics of pseudowords

The present study follows the implementational basics outlined in Section 3.3. However, as /s/ durations in pseudowords (and not in real words) are to be modelled, there are a number of complications. The most important complication arises from the widely shared belief that pseudowords do not have meaning (see Section 3.1.1 for a more detailed discussion). So how can one map form and meaning with forms that have no, or at least no a priori specified, meaning? In a recent study (Chuang et al. 2021) it was shown that the assumption that pseudowords

[1] An earlier version of this chapter has been published as part of Schmitz, Plag, et al. (2021).

are void of meaning is most probably wrong. Due to their formal similarity with existing words, pseudowords resonate with the lexicon. As a result, they may in fact carry meaning. Chuang et al. (2021) demonstrated that quantitative measures gauging the semantic neighbourhoods of pseudowords predict reaction times of lexical decision and acoustic durations. The present study is inspired by these results and implements a similar architecture. To model resonance of pseudowords with the lexicon, both real words and pseudowords must be included in the network. The following sections will detail the combined LDL implementation of real words and pseudowords.

5.1.2 Sets of pseudowords and real words

The pseudowords and their phonetic realisations that this study is based on are taken from the study of word-final /s/ production presented in Chapter 4. As linear discriminative learning (e.g. Baayen, Chuang, Shafaei-Bajestan, et al. 2019) in its current implementation does not offer the option to integrate clitics, the pseudoword set for the present study was limited to two types of /s/: non-morphemic and plural /s/. Recall that some pseudowords showed a number of different realisations by the participants in the production experiment, e.g. *prups* was sometimes produced as /pɹʌps/ and sometimes as /pɹups/. Thus, not 48 (i.e. the number of pseudowords in their orthographic representation) but 78 different phonological forms were included in the pseudoword data set. Table 5.1 gives an overview of all pseudowords and their phonological forms.

Table 5.1: Overview of all pseudowords and their phonological forms used in the LDL implementation. Transcriptions are given in the DISC keyboard phonetic alphabet (Burnage 1988).

Pseudoword		Phonological form	Pseudoword		Phonological form
blou-	fs	blufs	glai-	fs	gl1fs
	ks	bl{ks; bluks; blVks		ks	gl1ks; gl{ks
	ps	blups		ps	gl1ps; gl{ps
	ts	bl6ts; bluts		ts	gl1ts; gl{ts; gl2ts
cloo-fs; -ks; -ps; -ts		klufs; kluks; klups; kluts	plee-fs; -ks; -ps; -ts		plifs; pliks; plips; plits
gli-fs; -ks; -ps; -ts		glIfs; glIks; glIps; glIts glifs; gliks; glips; glits	pru-fs; -ks; -ps; -ts		prVfs; prVks; prVps; prVts; prufs; pruks; prups; pruts;

The second set of words contained real words and their phonetic realisations. Following Chuang et al. (2021), these words were extracted from the MALD corpus (Tucker, Brenner, et al. 2019). While the MALD corpus contains 26,793 real words, only a subset of 8,285 words was used for a number of reasons. First, some 7,577 words in the corpus contain multiple affixes. As it was unclear how to handle such words, these were excluded. Second, only words for which there were semantic vectors could be used, leading to the exclusion of 6,828 further words. Third, only words with transcriptions available in the CELEX corpus (Baayen et al., 1995) were retained, i.e. there was no transcription available for 818 words. Fourth, 3,285 words showed ambiguities regarding their morphology, e.g. walks as a third-person singular verb versus the plural of a noun. As huge numbers of words lead to extensive computation times, it was decided to exclude such cases as well. The final set of real words contained 6,165 simple and 2,120 complex word forms.

5.1.3 Cue matrices

As introduced in Section 3.3, cue matrices are coded in binary form, giving information on which triphones are part of which word. For the current implementation, two such cue matrices were created using the `WpmWithLdl` package's (Baayen, Chuang & Heitmeier 2019) `make_cue_matrix` function. First, C_{rw}, the real word cue matrix, was created for the set of real words. Then, a second cue matrix, C_{pw}, was created for the set of pseudowords. C_{pw} is a lot smaller than C_{rw} as there were only 78 phonological forms for pseudowords, but more than 8,000 for real words. C_{rw} was of dimension 8,285 × 7,610, while C_{pw} was of dimension 78 × 78.

5.1.4 Semantic matrices

To introduce semantics, i.e. semantic vectors, for the present set of real words, a pre-built semantic matrix A from Baayen, Chuang, Shafaei-Bajestan, et al. (2019) was used. These authors derived semantic vectors based on the TASA corpus (Ivens & Koslin 1991). For this, words were parsed into their lexomes, i.e. inflected words were represented by their base and sense-disambiguated labels for their respective inflectional functions. Ambiguous forms, e.g. walks, were disambiguated using part of speech tagging (Schmid 1999). Derived words were assigned a lexome for their base and a lexome for derivational function. Then, following Baayen et al. (2016) and Milin, Feldman, et al. (2017), naive discriminative learning (henceforth NDL; Baayen et al. 2011; Sering et al. 2018) was used to build semantic vectors. The Rescorla-Wagner update rule (Rescorla & Wagner 1972; Wagner & Rescorla 1972; Rescorla 1988) was applied incrementally to the sentences of the TASA corpus. That is, for each sentence the algorithm was given the task to predict the lexomes in that sentence from all lexomes of that sentence. This resulted in a 23,562 × 23,562 weight matrix A. This matrix lists all lexomes as rows and columns. Thus, each row i represents the association strengths of its corresponding lexome with all other lexomes as are represented by the columns of the matrix. In this state of the A matrix, lexomes predict themselves. Thus, the diagonal of the A matrix is set to zero (see Baayen, Chuang, Shafaei-Bajestan, et al. 2019 for a discussion on this procedure). Lastly, columns which mostly contained zeros, i.e. no information, and showed small variances ($\sigma < 3.4 * 10^{-8}$) were removed. The resulting A matrix is of dimension 23,562 × 5,030. Following the method outlined in Section 3.3, a semantic matrix for real words S_{rw} can be constructed based on A. That is, the semantic vector $\vec{s}$ in S_{rw} for a simplex word is identical to its corresponding lexome, while the semantic vector $\vec{s}$ in S_{rw} for a complex word is the sum of its corresponding lexomes. That is, the semantic vector of *apple* is $\overrightarrow{apple}$, while the semantic vector of *apples* is the sum of the vectors of the lexomes APPLE and PLURAL, i.e. $\overrightarrow{apples} = \overrightarrow{apple} + \overrightarrow{\text{PLURAL}}$. As a set of real words was used, S_{rw} contained only semantic vectors for this set of real words (instead of, e.g., all word forms of the TASA corpus). The final real word semantic matrix S_{rw} was of dimension 8,285 × 5,487.

While this procedure is rather straightforward, the creation of a pseudoword semantic matrix S_{pw} is not. Due to the nature of pseudowords, their lexomes are not contained within any corpus or the A matrix, for that matter. Instead, one can estimate a pseudoword's semantic content by utilising the semantic and phonological information on real words, i.e. their C and S matrix (Chuang et al. 2021). That is, the same transformation matrix F that is used for mapping real word cues onto predicted real word meanings (see Section 3.3) can be used to map pseudoword cues onto their estimated semantics. That is, one must first solve

$$F = C'_{rw}S_{rw} \tag{5.1}$$

to obtain F. Then, one can make use of the pseudoword cue matrix C_{pw} and estimate pseudoword semantics, as

$$S_{pw} = C_{pw}F \tag{5.2}$$

with S_{pw} denoting the originally estimated semantic matrix for pseudowords. In this semantic matrix, pseudowords of identical segmental makeup show identical semantics, as semantics are calculated only based on triphone occurrence, i.e. the semantics of *pleeps*$_{singular}$ is identical to the semantics of *pleeps*$_{plural}$. To differentiate between singular and plural pseudowords, the semantic vector of the PLURAL lexome is added to all plural pseudowords in the S matrix. Similarly, the semantic vectors of ALIEN and CREATURE are added to all pseudoword semantic vectors as participants in the original production experiment were told that pseudowords describe alien creatures. As explained in Section 4.1, the pairing of the pictures with pseudowords representing the alien creatures was randomised during the experiment. A particular pseudoword thus only contained the semantics of "alien creature" as a constant part of its own semantics, while other factors such as appearance, e.g. colour, shape, or number of eyes, differed across participants. One may assume that in the course of the experiment, participants gradually came to realise that the looks of these alien creatures, i.e. colour, shape, etc., are not relevant to their label names. Thus, participants were just aware of the fact that these are all alien creatures, without paying much attention to their individual features. Please see the supplementary material given in Chapter 11 for a detailed implementation.

5.1.5 Comprehension and production

Pseudoword comprehension and production were not computed and evaluated in isolation, but in combination with real words, simulating a real person's lexicon in a pseudoword comprehension and production situation, respectively. For this, a cue matrix C_{comb} was created based on a combined set of words, containing all aforementioned real words and pseudowords. In total, 8,440 word forms were part of this set of words. A combined semantic matrix S_{comb} was created by attaching S_{pw} to S_{rw}, and reordering its rows to reflect the same order of words as found in C_{comb} using the `LDLConvFunctions` package (Schmitz 2021a).

Then, using the `WpmWithLdl` package (Baayen, Chuang & Heitmeier 2019), a comprehension model was trained and checked for accuracy. That is, taking form vectors as input for the prediction of semantic vectors of output, $\hat{S}_{comb} = C_{comb}F$ is solved. Comprehension is successfully modelled for a word i if its predicted semantic vector $\hat{s}_i$ is most highly correlated with its targeted semantic vector s_i. This is true for 74.41% of cases (i.e. 6,165 word forms) in the comprehension model. In total, 25.59% of cases (i.e. 2,120 word forms) were incorrectly predicted, with 1,912 simple and 208 complex word forms. None of the incorrectly predicted word forms was a pseudoword.

Similarly, a production model was trained and checked for accuracy using functions of the aforementioned R package. Thus, semantic vectors were provided as input to predict form vectors as output, i.e. to solve $\hat{T}_{comb} = S_{comb}G$. Production was successfully modelled for a word *i* if its predicted triphones are those triphones present in its targeted cue vector in the correct sequence (possible sequences of triphones will be referred to below as *paths*). This was true for 97.3% of cases (i.e. 8,061 word forms) in the production model. In total, 2.7% of cases (i.e. 224 word forms) were incorrectly predicted, with 98 simple and 126 complex word forms. None of the incorrectly predicted word forms was a pseudoword.

5.1.6 Measures

In order to explore the potential of different measures emerging from the network to predict phonetic duration, a whole range of measures, based on the measures introduced by the `WpmWithLdl` package (Baayen, Chuang & Heitmeier 2019) and by Chuang et al. (2021), were extracted. The measures introduced by Chuang et al. (2021) were extracted using the `LDLConvFunctions` package (Schmitz 2021a). Please see the supplementary material given in Chapter 11 for exploratory analyses of individual measures.

In the following, the semantic measures are described first. Then, the phonetic measures are introduced.

L1NORM and L2NORM. The L1NORM is the sum of the absolute values of vector elements of a given word's predicted semantic vector $\hat{s}$, i.e. its city-block distance. The L2NORM is the square root of the sum of the squared values of a given word's predicted vector $\hat{s}$, i.e. its Euclidean distance. For both variables, higher values imply more strong links to many other lexomes. Thus, both measures may be interpreted as semantic activation diversity.

DENSITY. For DENSITY, the correlation values of a word's predicted semantic vector $\hat{s}$ and its eight nearest neighbours' semantic vectors $s_{n1}...s_{n8}$ are taken into consideration. The mean of these eight correlation values describes DENSITY, with higher values indicating a denser semantic neighbourhood.

ALC. The Average Lexical Correlation is the mean value of all correlation values of a pseudoword's estimated semantic vector as contained in S_{pw} with each of the real word semantic vectors as contained in S_{rw}. Higher ALC values indicate that a pseudoword's semantics are part of a denser semantic neighbourhood. Thus, ALC may be interpreted as a measure of semantic activation diversity for pseudowords.

EDNN. This variable describes the Euclidean Distance of a pseudoword's estimated semantic vector s and its Nearest semantic real word or pseudoword

Neighbour. Thus, higher values indicate a larger distance to the nearest semantic neighbour. EDNN may be regarded as a measure of semantic neighbourhood density.

NNC. The Nearest Neighbour Correlation is computed by taking a pseudoword's estimated semantic vector as given in S_{pw} and checking it for the highest correlation value against all real word semantic vectors as given in S_{rw}. This highest correlation value is taken as NNC value. Thus, higher values indicate that a pseudoword is semantically close to a real word. Additionally, one can tell which real word a pseudoword's semantics are closest to. This measure may be interpreted as a measure of similarity between pseudo- and real words, indicating the co-activation of a real word when confronted with a pseudoword.

SUPPORT. This measure describes the amount of support the word-final triphone (i.e. fs#, ks#, ps#, ts#) obtains for each pseudoword. The value of SUPPORT is extracted from $\hat{T}$. Higher values of this variable indicate a higher semantic support for the word-final triphone which includes the segment of interest, i.e. word-final /s/.

PATH_COUNTS. PATH_COUNTS describes the number of paths, i.e. possible sequences of triphones, detected for the production of a word by the production model. PATH_COUNTS may be interpreted as a measure of phonological activation diversity, as higher values indicate the existence of multiple candidates (and thus paths) in production.

PATH_SUM. PATH_SUM describes the summed support of paths for a predicted form. PATH_SUM may be interpreted as a measure of phonological certainty, with higher values indicating a higher certainty in the candidate form.

PATH_ENTROPIES. PATH_ENTROPIES contains the Shannon entropy values that are calculated over the path supports of the predicted form in $\hat{T}$. Thus, PATH_ENTROPIES may be interpreted as a measure of phonological uncertainty, with higher values indicating a higher level of disorder, i.e. uncertainty.

ALDC. The Average Levenshtein Distance of all Candidate productions is the mean of all Levenshtein distances of a word and its candidate forms. That is, for a word with only one candidate form, the Levenshtein distance between that word and its candidate form is its ALDC. For words with multiple candidates, the mean of the individual Levenshtein distances between candidates and targeted form constitutes the ALDC. Thus, higher values indicate that a word's candidate forms are very different from the intended pronunciation. ALDC may be interpreted as a measure of phonological neighbourhood density as it takes into account real word neighbourhoods for pseudowords, i.e. large values indicate sparse real word neighbourhoods.

5.2 Analysis

Recall that the data set of the production study (Chapter 4) contains non-morphemic, plural, and clitic word-final /s/ as final segment of a pseudoword. As mentioned in Section 5.1.2, the present LDL implementation does not include information on clitics. Thus, only durational data on non-morphemic and plural /s/ for the present study are considered. A subset of 666 data points remains, with 303 observations with non-morphemic /s/ and 363 observations with plural /s/. Due to some variable pronunciations requiring triphones not included in the present LDL implementation, 13 data points had to be excluded, resulting in a final data set with non-morphemic and plural /s/ durations of 653 data points, i.e. 300 entries on non-morphemic /s/ and 353 entries on plural /s/. The data set and the following analysis can be found in the supplementary material given in Chapter 11.

5.2.1 Covariates

Besides the aforementioned variables extracted and computed from the LDL implementation itself (see Section 5.1.6), the following covariates adopted from the production experiment (see Section 4.2.1) were included in the analysis. The main reason for this is to allow for the comparison of the performance of these predictors with the performance of LDL predictors. LDL measures often correlate with traditional measures (such as lexical frequencies, transitional probabilities, or neighbourhood densities), but the traditional measures have no clear correlating mechanisms in learning or processing.

There are, however, also covariates that do not tap into lexical properties, but that control for other influences, such as speech rate, the speaker, gender, the order of stimuli in an experiment, etc. These will be referred to as ”non-lexical covariates” and they will also be included in regression models.

For reasons of convenience, I will repeat the covariates adopted from the production experiment and their definitions in a shortened version in the following. See Section 4.2.1 for a detailed account.

typeOfS. This is the explanatory variable of the production study. As the present data set contains only two types of word-final /s/, this binary variable codes whether the pertinent pseudoword is a singular or plural form. It takes the value `nm` for pseudowords with a non-morphemic word-final /s/ and `pl` for pseudowords with a plural word-final /s/.

speakingRate. The speaking rate was computed as the number of syllables in an utterance divided by the duration of the utterance.

baseDurLog. Indicating a more local speaking rate, base duration was measured. The base duration in this case is equal to the summed duration of all word-internal segments preceding the /s/ under investigation. The base duration was log-transformed and centred. This variable is called baseDurLog.

pauseBin. In order to account for final-lengthening effects, all stretches of silence between the offset of the word-final /s/ and the onset of the following word were measured. Silence of 50 ms and above was considered as pause. The closure durations of following plosives were taken into account. Following the results of the production study, pause information was included as binary variable with the values `pause` versus `no_pause`.

transcription. As some pseudowords were produced with multiple pronunciations, their transcription was incorporated as a categorical variable.

biphoneProbSumBin. A binary covariate based on the summed biphone probability was used as a measure of contextual predictability.

list & slideNumber. To account for possible durational differences due to priming and similar effects, the list number (1 to 12) and the point of occurrence during the experiment of the individual item were also included.

preC. It has been shown that the consonant preceding word-final /s/ may influence the duration of word-final /s/. The consonant preceding the final /s/ was therefore included as a covariate, preC.

biphonePron. The probability of the final biphones /fs/, /ks/, /ps/ and /ts/ in monomorphemic words is included as covariate to account for potential effects of phonotactics.

folType. To account for potential effects of the following word on the duration of /s/, the following word was included in regard to its onset segment adjacent to the word-final /s/. This information was included in form of its segmental class in folType.

speaker / age. speaker ID was included to account for inter-speaker differences in production. age was included as well as it may show an influence on phonetic realisations.

gender / location / monoMultilingual. Participants' gender and whether they had grown up in London or elsewhere in South Britain (location) were included as well as they may influence phonetic realisations. Additionally, participants who were early bilinguals (i.e. the L2 was/the L2s were acquired as a pre-school child) were categorised as multilingual, while all other participants were categorised as monolingual in monoMultilingual.

Finally, one additional covariate was introduced, following the discussion of the production experiment.

REAL. Some of the pseudowords used here and in the production experiment have an orthographically different, but phonologically identical real word counterpart (see Section 4.4). The variable REAL was introduced to control for this potential confound. This variable is TRUE for pseudowords with such a real word counterpart, and FALSE for those without. The following pseudowords were considered to show such counterparts: *pleet(s)* corresponds to *pleat(s)*, *glits* corresponds to *glitz*, and *gliks* corresponds to the plural of the surname *Glick* (as in *the Glicks live next door*), whereas *glif(s)* corresponds to *glyph(s)*, which has a very low frequency and thus may constitute a pseudoword for most of the participants.[2]

All of the following analyses make use of the following non-lexical covariates: BASEDURLOG, SPEAKINGRATE, SLIDENUMBER, and PAUSEBIN as variables concerning speech rate and continuity, PREC and FOLTYPE accounting for coarticulatory effects, LIST taking into consideration potential priming effects, MONOMULTILINGUAL, GENDER, LOCATION, AGE, and SPEAKER to account for speaker-individual differences, and REAL to include effects of real word counterparts.

5.2.2 Overview of the data

An overview of all variables is given in Table 5.2 and Table 5.3.

5.2.3 Modelling strategy

Three kinds of models were devised. First, a baseline model with the traditional predictor variables (plus the non-lexical covariates). Second, a model with LDL predictors that also includes TYPEOFS as a covariate (plus the non-lexical covariates). Third, a model that contains only the LDL predictors (plus the non-lexical covariates).

The three kinds of models will allow answering the given research question. Recall that the ultimate goal is to understand how systematic durational differences emerge between words of different, but homophonous morphological categories. Traditional lexical variables are predictive but cannot explain how morphology can make its way into durational differences. But these models can show that such differences exist by looking at the effect of the variable TYPEOFS. This is the baseline model. As an alternative, a model that uses LDL measures is

[2] Note that in Schmitz, Plag, et al. (2021) a slightly different set of pseudowords was considered to have real word counterparts, i.e. *pleets*, *glits*, *glaiks* (instead of *glik*), and *glifs*. The analysis presented here uses the set of pseudowords given in the main text. Results reported here and in Schmitz, Plag, et al. (2021) do not differ significantly; all effects show into the same directions.

Table 5.2: Summary of the dependent variable and the numerical variables used in the modelling processes.

Dependent variable	Mean	St. Dev.	Min	Max
sDurLog	-2.116	0.388	-3.361	-1.221
Numerical variables	Mean	St. Dev.	Min	Max
speakingRate	3.566	0.927	1.310	7.100
baseDurLog	-1.203	0.232	-1.987	-0.375
biphoneProb	0.001	0.002	0.000	0.004
age	28.470	9.323	19.000	58.000
Component1	0.000	1.975	-17.748	2.509
Component2	0.000	1.959	-2.832	11.989
Component3	0.000	1.488	-4.312	2.983
Component.woA.1	0.000	1.973	-18.860	2.178
Component.woA.2	0.000	1.957	-10.011	2.894
Component.woA.3	0.000	1.487	-4.175	2.928
Component.woA.4	0.000	1.269	-3.608	3.076

Table 5.3: Summary of categorical predictors and the explanatory variable of interest in the final data set.

Categorical variables	Levels				
typeOfS	nm: 300	pl: 353			
pauseBin	no: 412	yes: 241			
transcription	38				
biphoneProbSumBin	high: 161	low: 492			
list	12				
slideNumber	48				
preC	f: 156	k: 169	p: 164	t: 164	
folType	APP: 190	F: 11	N: 106	P: 165	V: 181
speaker	40				
gender	2				
location	London: 392	elsewhere: 261			
monoMultilingual	monolingual: 532	multilingual: 121			
real	FALSE: 542	TRUE: 111			

implemented. If these measures are predictive, they offer an explanation of the morphologically induced phonetic differences: They emerge as a by-product of the association of form and meaning in the mental lexicon, and this association is the outcome of discriminative learning. By having a model that also includes typeOfS as an additional predictor, one can see whether the LDL measures completely capture the morphological effect, or whether there is a residue of morphological information that is predictive of duration but is still not captured by the LDL measures.

5.2.4 Model A: Traditional measures

This model is meant to resemble those in previous studies on word-final /s/ duration (e.g. Plag et al. 2017), with a special focus on the model found in the production study (see Section 4.2.4). Thus, an LMER model was fitted with similar variables and similar effect structures: typeOfS, biphoneProbSumBin, and biphoneProb, as well as those control variables included in all analyses of this study. None of these covariates showed high correlation coefficients. Hence, no cautionary measures regarding collinearity were required before an initial full model was constructed. Following standard procedures to reduce the potentially harmful effect of skewed distributions in linear regression models (e.g. Winter 2019), the dependent variable, duration of /s/, was log-transformed. The name of this variable is sDurLog. The model selection process proceeded as explained in Section 3.2.1. That is, non-significant variables were excluded in a controlled step-wise fashion.

Then, variance inflation factors (VIFs) were checked. The covariates biphoneProb and preC showed high VIF values (i.e. 46.53 and 46.88, respectively), indicating potential overfitting of the model (e.g. Zuur et al. 2010; Fox & Weisberg 2019). Consequently, preC was removed from the model as it showed the highest VIF value, following the procedure described by Zuur et al. (2010). Re-fitting the model without preC and re-checking the new variance inflation factor values revealed only non-problematic values.

Finally, the resulting model's residuals were trimmed, following the reasoning given in Section 3.2.1. This procedure led to a loss of 4 data points, i.e. 0.61% of all data points.

5.2.5 Model B: LDL measures and typeOfS specification

This model makes use of all LDL measures as well as of the typeOfS variable. Additionally, the non-lexical covariates are included. When fitting a model with such a multitude of variables, collinearity is an issue to consider. Following the procedure given in Section 3.2.3, all covariates were checked for correlation using the `SfL` package (Schmitz & Esser 2021). This correlation check resulted in eight correlation coefficients indicating a high degree of correlation, for which the threshold was assumed to be $|rho| \geq 0.5$. The pairs of correlated covariates as well as their correlation coefficients are given in Table 5.4.

Table 5.4: Correlated variables and their correlation coefficients.

Variables		*rho*	Variables		*rho*
l1norm	l2norm	0.98	typeOfS	NNC	-0.89
path_counts	path_entropies	0.95	path_counts	support	-0.65
path_counts	ALDC	0.89	path_sum	support	0.73
path_entropies	ALDC	0.90	path_entropies	support	-0.63

Due to the high number of correlated variables, a principal component analysis was used (PCA; see Section 3.2.3 for further details) to address collinearity issues. In a PCA, the dimensionality of the data is reduced by transforming the included variables into principal components. These transformations result in linear combinations of the predictors that are orthogonal to each other. Thus, the resulting principal components are not correlated. All variables given in Table 5.4 were included in the computation of the principal component analysis, which yielded nine principal components.

The next step of the PCA is to determine how many of these principal components are meaningful and thus should be retained for further use. Following the criteria given in Section 3.2.3, the following was found. First, any component that displays an eigenvalue greater than 1 accounts for a greater amount of variance than had been contributed by one variable. Such a component is therefore potentially meaningful. This is true for components 1, 2, and 3. Second, one should retain enough components so that the cumulative percentage of variance explained is equal to at least 80%. This, again, is true for components 1, 2, and 3. Third, only interpretable components are to be retained. This, once again, is true for components 1, 2, and 3. Therefore, components 1 to 3 are retained for further analysis, all of which show an eigenvalue greater than 1, account for more than 80% of variance, and contain strong representations of variables in their loadings.[3] But what do these principal components mean? The highest loadings

[3] In addition, a cluster analysis was performed. This analysis revealed clusters which align well with the retained components of the principal component analysis. The cluster analysis can be found in the supplementary material given in Chapter 11.

of the principal components, i.e. the correlation of the original variables to the pertinent component, are given in Table 5.5.

Table 5.5: Loadings of original predictor variables in the three retained principal components of the principal component analysis for model B.

	Component1	Component2	Component3
l1norm		0.397	0.348
l2norm		0.405	0.363
path_counts	0.813		
path_entropies	0.828		
path_sum	-0.430		
ALDC	0.710		
NNC		0.698	
support	-0.650		
typeOfS		0.421	0.517

Component1. Component1 is most strongly positively correlated with path_counts, path_entropies, and ALDC, while it is most strongly negatively correlated with path_sum and support. For path_counts, higher values indicate the existence of multiple candidates (and thus paths) in production. It hence functions as an indicator of phonological uncertainty. Values of path_entropies relate to the level of uncertainty concerning the path supports of the predicted candidate form, with higher values indicating a higher level of uncertainty. For ALDC, higher values mean that a word's candidate forms are very different from the intended pronunciation, indicating uncertainty in production. path_sum describes the summed support of paths for a predicted form, with higher values indicating a higher certainty in the candidate form. Higher values for support suggest more certainty in the choice of the word-final triphone. Component1 can thus be described as a dimension that represents phonological or articulatory certainty.

Component2. Component2 is most strongly correlated with l1norm, l2norm, NNC, and typeOfS. l1norm and l2norm both imply more strong links to many other lexomes with higher values indicating a higher semantic activation diversity. Higher values of NNC suggest a close real word neighbour, which leads to higher levels of co-activation of that real word when confronted with the pseudoword, also leading to higher semantic activation diversity. As for typeOfS, Component2 is positively correlated with the presence of non-morphemic /s/ data points.

Component3. Component3 is similar to Component2 as it is also strongly correlated with l1norm, l2norm, and typeOfS. Again, for l1norm and l2norm higher values indicate higher semantic activation diversity. typeOfS is positively correlated for plural /s/ data points. I will come back to the interpretation of this correlation in Section 5.3.2.

In a next step, LMER models were fitted following the procedure given in Section 3.2.1. As in Section 5.2.4, the dependent variable, duration of /s/, was log-transformed to reduce the potentially harmful effect of skewed distributions in linear regression models. Following the backward step-wise selection process for model selection, a first model containing all remaining variables is created. That is, Component1, Component2, Component3, density, ALC, EDNN, baseDurLog, speakingRate, pauseBin, folType, preC, and real were included as fixed effects. The remaining variables, gender, location, monoMultilingual, age, list, and speaker, were included as random intercepts.

This full model was then continuously reduced through step-wise exclusion of non-significant variables. Then, variance inflation factors (VIFs) were computed. For the present model, all variance inflation factor values were below 3. Thus, no action was necessary. Finally, the resulting model needed trimming of its residuals. This resulted in a loss of 6 data points (0.92%).

5.2.6 Model C: LDL measures only

This model uses all LDL measures but does not incorporate the typeOfS covariate. As in the previous model, there was a high number of highly correlated variables (see Table 5.4 with the exception of the correlation of typeOfS and NNC, as typeOfS is not included in this analysis). I therefore again computed a principal component analysis, following the procedure outlined in Section 3.2.3. Following the first two criteria, two principal components are to be retained. However, considering the third criterion, it is found that the two components are not readily interpretable as they show relatively high positive or negative correlations with all or almost all variables, without indicating a clearly discernible dimension underlying the patterns of correlations. I thus turned to the procedure of competitive exclusion to reduce collinearity issues as introduced in Section 3.2.3. This procedure led to the exclusion of l2norm, path_counts, path_entropies, and path_sum.

Linear mixed-effects regression models were fitted according to the procedure given in Section 3.2.1. That is, an initial full model was fitted with the following variables: l1norm, ALDC, support, density, ALC, EDNN, NNC, baseDurLog, speakingRate, pauseBin, folType, preC and real. As for random effects, random intercepts for GENDER, LOCATION, MONOMULTILINGUAL, AGE, LIST, and SPEAKER were included. The dependent variable, duration of /s/, again was log-transformed.

This full model was then continuously reduced through step-wise exclusion of non-significant variables, following the aforementioned procedure. Then, variance inflation factors were computed, resulting only in non-problematic values. Finally, the resulting model needed trimming of its residuals. This procedure led to a loss of 8 data points, i.e. 1.2% of all data points.

5.3 Results

5.3.1 Model A: Traditional measures

The final model of traditional measures included effects of the following variables: type of /s/ (TYPEOFS), speaking rate (SPEAKINGRATE), log-transformed base duration (BASEDURLOG), pause (PAUSEBIN), following segmental type (FOLTYPE), and the summed biphone probability (BIPHONEPROBSUMBIN). As for random effects, random intercepts for SPEAKER and random slopes for TYPEOFS are included. The *p*-values of the analysis of variance of the final model are given in Table 5.6.

Table 5.6: *p*-values of fixed effects in model A, fitted to the log-transformed durations of /s/.

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(F)
TYPEOFS	0.711	0.711	1	37.90	13.845	0.001
SPEAKINGRATE	0.163	0.163	1	604.07	3.165	0.076
BASEDURLOG	6.278	6.278	1	572.80	122.247	0.000
PAUSEBIN	5.430	5.430	1	635.92	105.722	0.000
BIPHONEPROBSUMBIN	0.646	0.646	1	596.28	12.580	0.000
FOLTYPE	2.199	0.550	4	605.15	10.703	0.000

The marginal R^2 value of the model is 0.43, i.e. fixed effects explain 43% of variation in the data (see Section 3.2.1 for details on R^2 values). Taking random effects into account as well, the conditional R^2 value is 0.62. That is, the model explains 62% of data variation in total. The R^2 values are similar to the values found for the final model of the production experiment (see Section 4.3.1).

The estimates of the final model and their *p*-values are given in Table 5.7. The reference levels for the categorical predictors are: for TYPEOFS it is `nm`, for PAUSEBIN it is `no_pause`, for BIPHONEPROBSUMBIN it is `high`, and for FOLTYPE it is `APP`.

Table 5.7: Fixed-effect coefficients and *p*-values as computed for model A (mixed-effects model fitted to the log-transformed duration of /s/).

	Estimate	SE	df	t-value	Pr(|t|)
(Intercept)	-1.202	0.083	407.927	-14.520	0.000
typeOfSpl	-0.087	0.023	37.896	-3.721	0.001
speakingRate	-0.022	0.012	604.072	-1.779	0.076
baseDurLog	0.635	0.057	572.805	11.057	0.000
pauseBinpause	0.234	0.023	635.917	10.282	0.000
biphoneProbSumBinlow	-0.076	0.021	596.279	-3.547	0.000
folTypeF	-0.001	0.073	610.436	-0.007	0.994
folTypeN	-0.004	0.028	600.528	-0.134	0.893
folTypeP	-0.027	0.025	599.182	-1.107	0.269
folTypeV	-0.145	0.025	610.241	-5.852	0.000

The predictor strength of individual covariates was checked by taking the final model as template. For each predictor variable, a model was fitted lacking the particular variable. For each of these models, R^2 values were computed and compared following the method outlined in Section 3.2.1. The variable leading to the highest decrease in R^2 value as compared to the final model is thus the variable showing the highest predictor strength. The results of this comparison are reflected in the hierarchy given in (1). The decrease in R^2 is greatest when removing baseDurLog, followed by pauseBin, and so forth. The resulting order is identical to the one found in the analysis of production experiment for the complete data set (see Section 4.3.1).

(1) baseDurLog » pauseBin » typeOfS » folType » speakingRate » biphoneProbSumBin

5.3.2 Model B: LDL measures and typeOfS specification

In the final model including LDL measures as well as the typeOfS covariate as parts of the individual components resulting from the principal component analysis and fitted according to the procedure described in Section 3.2.1, one finds effects of the first principal component (Component1), the third principal component (Component3), density, ALC, base duration (baseDurLog), following pause (pauseBin), following segmental type (folType), and preceding consonant (preC). Regarding random effects, only a speaker-specific random intercept turned out to significantly improve model fit. The *p*-values of the analysis of variance of the final model are given in Table 5.8.

Table 5.8: *p*-values of fixed effects in model B, fitted to the log-transformed durations of /s/.

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(F)
Component1	0.376	0.376	1	618.06	6.970	0.008
Component3	1.340	1.340	1	627.71	24.819	0.000
baseDurLog	6.751	6.751	1	620.55	125.080	0.000
pauseBin	5.805	5.805	1	642.19	107.568	0.000
folType	2.093	0.523	4	617.98	9.695	0.000
preC	0.702	0.234	3	615.33	4.334	0.005
density	0.219	0.219	1	621.79	4.067	0.044
ALC	0.293	0.293	1	623.25	5.425	0.020

The marginal R^2 value of the final model is 0.42, thus fixed effects explain 42% of the variation in the data. The conditional R^2 value of the final model is 0.60, that is fixed and random effects taken together explain 60% of variation.

The estimates of the final model and their *p*-values are given in Table 5.9. The reference levels for the categorical predictors are: for pauseBin it is `no_pause`, for folType it is `APP`, and for preC it is `f`.

Table 5.9: Fixed-effect coefficients and *p*-values as computed for model B (mixed-effects model fitted to the log-transformed duration of /s/).

	Estimate	SE	df	t-value	Pr(|t|)
(Intercept)	-1.106	0.124	635.215	-8.952	0.000
Component1	0.014	0.005	618.057	2.640	0.008
Component3	-0.041	0.008	627.708	-4.982	0.000
baseDurLog	0.652	0.058	620.548	11.184	0.000
pauseBinpause	0.237	0.023	642.193	10.371	0.000
folTypeF	-0.014	0.075	621.463	-0.180	0.857
folTypeN	-0.006	0.029	614.760	-0.198	0.843
folTypeP	-0.028	0.025	615.172	-1.126	0.261
folTypeV	-0.141	0.025	620.352	-5.612	0.000
preCk	-0.023	0.027	614.436	-0.835	0.404
preCp	-0.040	0.027	614.491	-1.475	0.141
preCt	-0.095	0.028	615.916	-3.414	0.001
density	-0.241	0.119	621.790	-2.017	0.044
ALC	-5.302	2.277	623.246	-2.329	0.020

As described in Section 3.2.1, the predictor strength of individual covariates was checked by taking the final model as template. The result of this procedure is reflected in the hierarchy in (2). The decrease in R^2 is greatest when removing baseDurLog, followed by pauseBin, and so forth. In sum, variables containing measures obtained by the LDL analysis appear to be meaningful predictors of /s/ duration.

(2) baseDurLog » pauseBin » Component3 » folType » ALC » density » Component1 » PreC

Figure 5.1 shows the effect on /s/ duration of the numerical variables included in the model. The estimated values of the dependent variable sDurLog, i.e. /s/ duration, and baseDurLog, i.e. base duration, are back-transformed into seconds. For Component1, higher values lead to longer /s/ durations (Panel A), while for Component3, higher values lead to shorter /s/ durations (Panel B). Higher values of density (Panel C) and ALC (Panel D) come with shorter /s/ durations. Longer bases come with longer /s/ durations (Panel E).

The partial effects of the categorical variables included in the final model are illustrated in Figure 5.2. Pauses lead to longer /s/ durations (Panel A), which is most likely a case of phrase-final lengthening (e.g. Cooper & Danly 1981). There is also an effect of the following segment type, with /s/ being shorter when followed by a vowel (Panel B). This difference is significant for all consonant types being compared against vowels with the exception of fricatives. However, as there is only a small number of fricative cases in the data, this non-significant difference is potentially not meaningful. Lastly, there is an effect of preceding consonant on /s/ duration (Panel C). /s/ duration is significantly longer if preceded by a voiceless labiodental fricative /f/ or a voiceless velar stop /k/ as compared to cases where /s/ is preceded by a voiceless alveolar stop /t/. All other comparisons are non-significant.

Let us turn to the variables of interest, i.e. those derived from the LDL network. Component1 acts as a general measure of phonological certainty. High values of Component1 come with high values of path_counts, path_entropies, and ALDC, indicating a high level of phonological uncertainty. At the other end of the Component1 dimension, high values of path_sum and support indicate a high level of phonological certainty. Higher uncertainty appears to lead to longer /s/ durations, while higher certainty appears to lead to shorter /s/ durations.

Recall from Section 5.2.5 that Component3 relates to semantic activation diversity and to the presence of the plural suffix. Higher values of Component3 indicate a higher level of semantic activation diversity. Higher levels of activation diversity then lead to shorter /s/ durations (see Panel B of Figure 5.1). High values of Component3 are positively correlated with the presence of plural /s/.

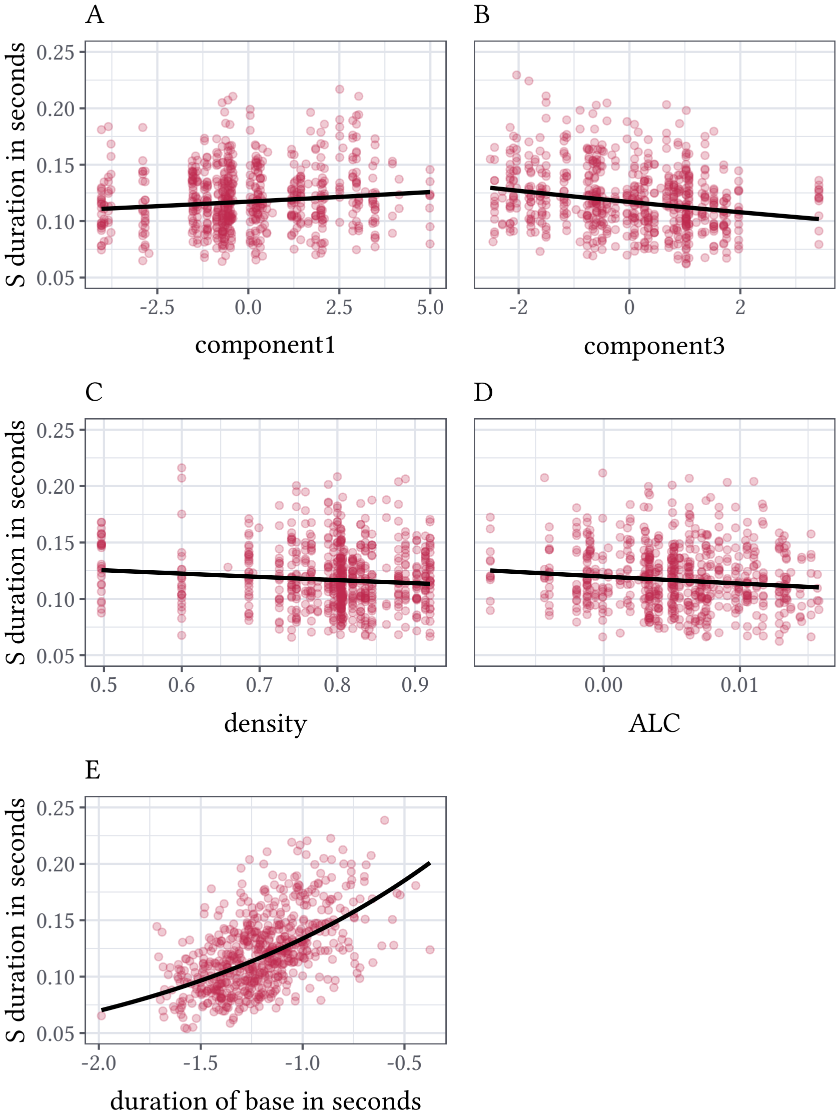


Figure 5.1: Partial effects of the numerical variables Component1 (Panel A), Component3 (Panel B), density (Panel C), ALC (Panel D), and baseDurLog (back-transformed, Panel E) included in model B, fitted to the log-transformed values of duration of /s/.

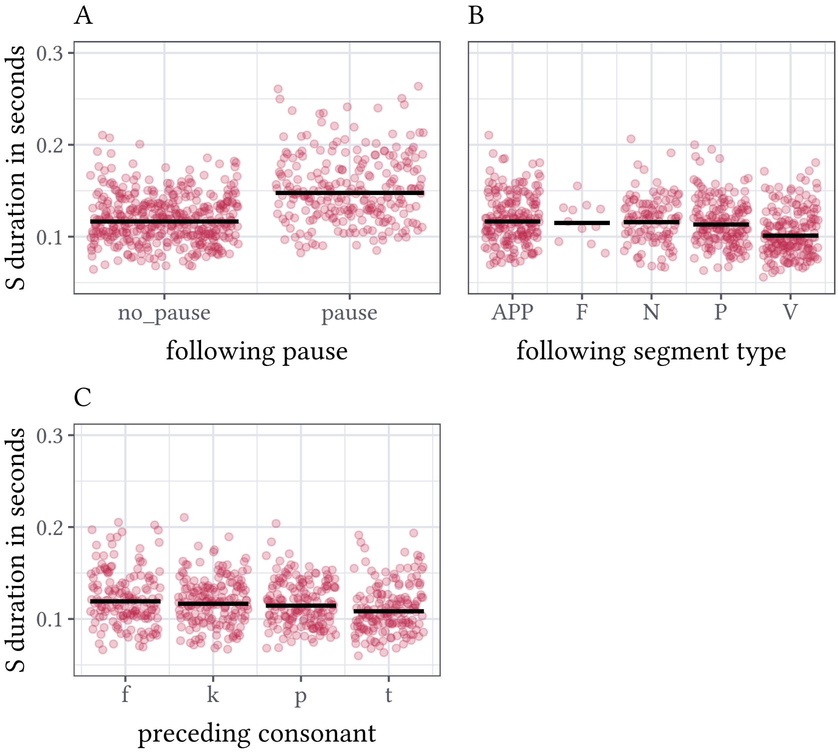


Figure 5.2: Partial effects of the categorical variables pauseBin (Panel A), folType (Panel B), and preC (Panel C) included in model B, fitted to the log-transformed values of duration of /s/.

It appears that the presence of plural makes words semantically more similar to each other as they share this meaning component. Hence, it is to be expected that plural words live in a space of greater semantic activation diversity. Component3 is not only a measure of semantic activation diversity, but also indicates that plural pseudowords show a tendency of having a higher degree of semantic activation diversity as compared to monomorphemic pseudowords in general. density and ALC also tap into the semantics of pseudowords. That is, similar to Component3, higher values indicate higher levels of semantic activation diversity. These higher levels then lead to shorter /s/ durations.

5.3.3 Model C: LDL measures only

The final model of LDL measures only was fitted with effects of the following variables: l1norm, ALC, NNC, log-transformed base duration (baseDurLog), pause (pauseBin), following segmental type (folType), and preceding consonant (preC). The speaker variable was included as random intercept. The *p*-values of the analysis of variance of the final model are given in Table 5.10.

Table 5.10: *p*-values of fixed effects in model C, fitted to the log-transformed durations of /s/.

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr(F)
l1norm	0.685	0.685	1	611.07	13.473	0.000
baseDurLog	6.047	6.047	1	627.51	118.901	0.000
pauseBin	5.440	5.440	1	632.72	106.956	0.000
folType	2.056	0.514	4	610.10	10.105	0.000
preC	0.761	0.254	3	607.96	4.985	0.002
ALC	0.534	0.534	1	615.51	10.504	0.001
NNC	0.778	0.778	1	619.67	15.296	0.000

With a marginal R^2 value of 0.41, the fixed effects of this model explain 41% of variation within the data. The conditional R^2 value of the model is 0.61, that is the complete model accounts for 61% of variation.

The coefficients of the final model and their *p*-values are given in Table 5.11. The reference levels for the categorical covariates are: for pauseBin it is `no_pause`, for folType it is `APP`, and for preC it is `f`.

Table 5.11: Fixed-effect coefficients and *p*-values as computed for model C (mixed-effects model fitted to the log-transformed duration of /s/).

	Estimate	SE	df	t-value	Pr(|t|)
(Intercept)	-2.334	0.320	625.440	-7.301	0.000
l1norm	-0.044	0.012	611.066	-3.671	0.000
baseDurLog	0.624	0.057	627.514	10.904	0.000
pauseBinpause	0.233	0.022	632.719	10.342	0.000
folTypeF	-0.019	0.073	613.088	-0.267	0.790
folTypeN	-0.005	0.028	607.324	-0.195	0.845
folTypeP	-0.023	0.024	607.817	-0.950	0.343
folTypeV	-0.140	0.025	611.952	-5.693	0.000
preCk	-0.029	0.027	607.726	-1.058	0.291
preCp	-0.053	0.027	607.478	-1.950	0.052
preCt	-0.101	0.028	608.068	-3.632	0.000
ALC	-6.663	2.056	615.511	-3.241	0.001
NNC	1.221	0.312	619.671	3.911	0.000

As for both other final models, the predictor strength of the individual predictors was checked. This procedure resulted in the hierarchy of predictor strength given in (3). That is, the decrease in R^2 is greatest when removing BASEDURLOG, followed by PAUSEBIN, and so forth.

(3) BASEDURLOG » PAUSEBIN » FOLTYPE » NNC » L1NORM » ALC » PREC

Figure 5.3 displays the effect on /s/ duration of the numerical variables included in the model. Base duration shows an identical effect as compared to model B in Section 5.3.2, i.e. longer base durations come with longer /s/ durations.

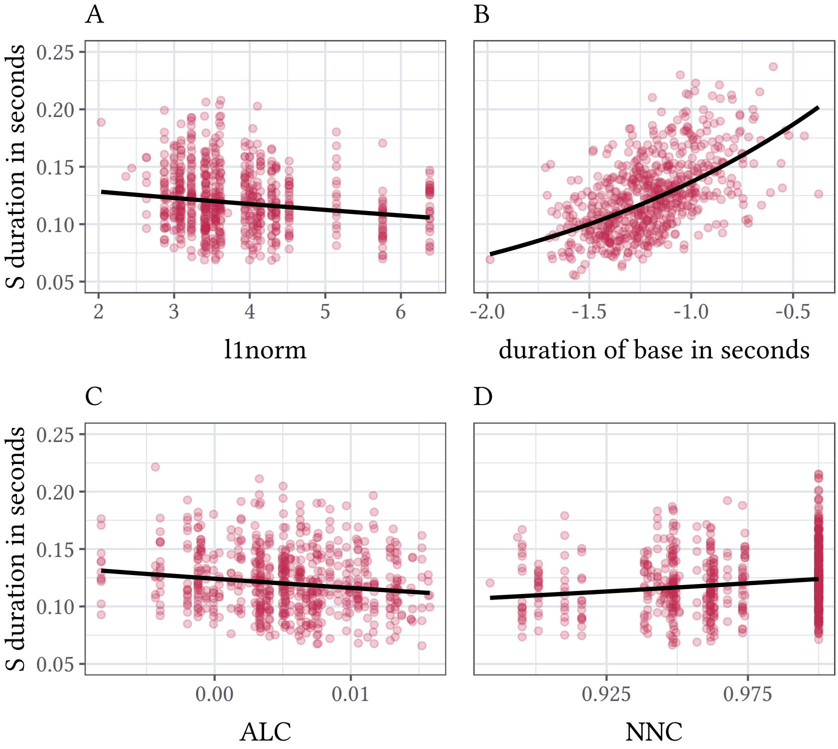


Figure 5.3: Partial effects of the numerical variables L1NORM (Panel A), BASEDURLOG (back-transformed, Panel B), ALC (Panel C), and NNC (Panel D) included in model C, fitted to the log-transformed values of duration of /s/.

Figure 5.4 shows the effect on /s/ duration of the categorical variables included in the model. Pauses again come with longer /s/ durations, and /s/ is shorter if followed by a vowel. There is also an effect of the preceding consonant, with /s/ duration being significantly longer if preceded by a voiceless labiodental fricative /f/ or a voiceless velar stop /k/ as compared to cases where /s/ is preceded by a voiceless alveolar stop /t/. These results are generally in line with those by the analysis in the previous section.

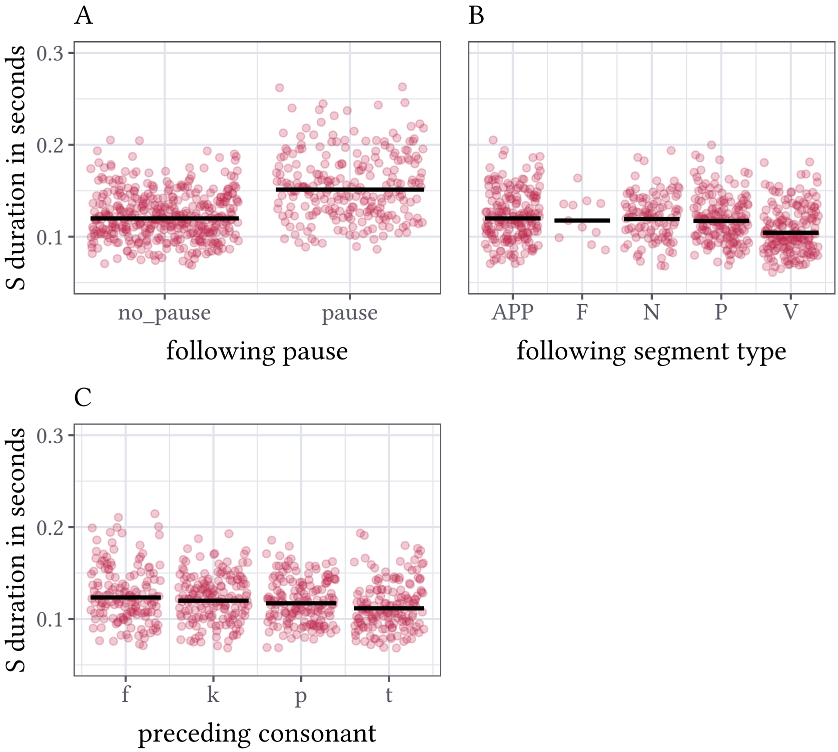


Figure 5.4: Partial effects of the categorical variables pauseBin (Panel A), folType (Panel B), and preC (Panel C) included in model C, fitted to the log-transformed values of duration of /s/.

Taking a closer look at the variables of interest, one finds that higher values of l1norm and ALC, i.e. higher semantic activation diversity, lead to shorter /s/ durations. As in model B, higher levels of semantic activation diversity come with shorter /s/ durations. For NNC, it is found that /s/ duration is longer if a pseudoword is semantically similar to a real word.

5.4 Discussion

The production study presented in Chapter 4 of this book as well as previous studies (Zimmermann 2016; Plag et al. 2017; Seyfarth et al. 2017; Tomaschek et al. 2019; Plag et al. 2020) reported that there are significant differences in the acoustic duration between different types of word-final /s/ in English. Such durational differences challenge established feed-forward theories of morphology-phonology interaction (e.g. Chomsky & Halle 1968; Kiparsky 1982) as well as theories of psycholinguistics (e.g. Levelt et al. 1999; Roelofs & Ferreira 2019; Turk & Shattuck-Hufnagel 2020). The present study investigated whether measures derived on the basis of a discriminative learning theory are predictive of /s/ durations in pseudowords. In particular, LDL networks that model the production of a word based on its relation to the rest of the lexicon were implemented.

The predictive possibilities of LDL measures were explored by fitting three different models: a) a model based on the traditional predictors as used in previous studies (Plag et al. 2017; Tomaschek et al. 2019) and most importantly in the production study reported in this book; b) a model with LDL measures and a variable typeOfS specifying the presence or absence of an affix; and c) a model with LDL measures but without a variable specifying the presence or absence of an affix. Both models with LDL measures show that such measures are predictive of /s/ durations. This result is the most important of the present study. While traditional variables such as lexical frequencies, bigram frequencies, transitional probabilities, or neighbourhood densities measure important lexical properties, it is unclear why they would manifest themselves in a particular morphological effect in speech production. In LDL such effects can emerge through the mapping of form and meaning in a clearly defined process of discriminative learning.

All regression models showed a similar hierarchy of predictor strength for the variables included in the models. For the traditional model A, typeOfS is the third-strongest predictor of /s/ duration and for model B this spot is taken by Component3, while there is no comparable variable included in model C. Comparing the variance explained by the fixed effects of the different models, one finds that the traditional model accounts for most variation, i.e. 43%, while the LDL model including the typeOfS variable accounts for 42%, and the LDL model without the typeOfS variable accounts for 41% of variation. Thus, in terms of marginal R^2 values, all three models are close to each other. To check whether these differences in marginal R^2 values are of significance, the three models were refitted to the untrimmed data set and then compared with a likelihood-ratio test. The results suggest that there is no significant difference between the traditional model and the LDL model including the typeOfS variable. However, the LDL model without the typeOfS variable shows a significantly worse fit ($p < 0.01$). This seems to indicate that the LDL measures do not capture the full amount of the variance that is captured by the variable typeOfS. This means that there is still something about the morphological function that translates into duration and that is not properly modelled by the associative measurements of the learning network. The same problem holds, incidentally, for the traditional model (model A), in which the usual lexical measures (such as lexical frequencies, neighbourhood densities, etc.) and phonetic covariates (such as pauses, speech rate, etc.) are also not able to cover all durational variance. The morphological residue in both types of analysis remains a conundrum that calls for more sophisticated approaches in future research.

The LDL measures included in the final models are either concerned with semantic activation diversity (Component3, ALC, & density in model B; l1norm & ALC in model C), semantic similarity (NNC in model C) or with phonological certainty (Component1 in model B).

Higher degrees of semantic activation diversity come with shorter /s/ durations. This effect is similar to the one which was reported by Tucker, Sims, et al. (2019) in a study on stem vowels and Tomaschek et al. (2019) in their NDL study on /s/ duration. A higher degree of activation diversity makes it "more difficult to discriminate the targeted outcome from its competitors" (Tomaschek et al. 2019: 27). As for production, a prolongation of the acoustic signal is dysfunctional if the prolongation maintains or increases the discrimination problem instead of contributing to resolving it (Tomaschek et al. 2019).

In the model without typeOfS as predictor variable, NNC (i.e. a pseudoword's semantic similarity to its closest semantic real word neighbour) emerges as significant (see model C). Why so? As reported in Section 5.2.5, the typeOfS variable and NNC are strongly negatively correlated ($rho = -0.89$). Post-hoc analysis shows that plural /s/ has significantly lower NNC values as compared to non-morphemic /s/ (Wilcoxon test, $p < 0.001$). It therefore appears that NNC takes over the role of differentiating between plural and non-morphemic /s/ in model C.

As for phonological certainty, one finds that higher phonological certainty comes with shorter /s/ durations, while higher phonological uncertainty comes with longer /s/ durations. Shorter durations in contexts of high phonological certainty may be related to effects of frequency, i.e. highly frequent forms are produced with higher certainty and are thus shorter.

The results of the present study may bring up further questions. First, are the predictive measures found for word-final /s/ duration in pseudowords also predictive for word-final /s/ duration in real words? The NDL implementation of Tomaschek et al. (2019) suggests that they are, but LDL networks still need to be implemented. It would be especially interesting to model those data sets that have yielded seemingly contradictory effects. Second, taking into account that the specification of typeOfS in the modelling process leads to a significantly better model fit, one may ask what the underlying reasons for this significant effect are. This then automatically leads to another question: Is it possible to catch the effect of the typeOfS specification in terms of (new) LDL measures?

To summarise, this study was the first to investigate durational differences between different types of word-final /s/ (non-morphemic versus plural /s/) in pseudowords by means of an LDL implementation, measures, and resulting statistical analyses. The findings yielded important evidence on the question of how such durational differences come to be, i.e. they can be predicted based on their pseudoword's relations to the lexicon. It was demonstrated that durational differences emerge from the pseudoword's resonance with the lexicon by way of differing degrees of semantic activation diversity and phonological uncertainty. These manifestations of the relations to other words in the lexicon in turn are the result of discriminative learning.

6 Perception of word-final /s/

As introduced in detail in Section 2.2, the perception study presented here investigates whether subphonemic durational differences in word-final /s/ are perceived by listeners. Two hypotheses derived from theories and models of speech perception are examined. H perc$_1$, the *Abstractionist Hypothesis*, assumes that listeners are not sensitive to subphonemic durational differences. H perc$_2$, the *Phonetic Detail Hypothesis*, predicts that subphonemic durational differences are perceptible. Subsequently, listeners are assumed to be sensitive to such differences. The two hypotheses are tested by analysing the results of a same-different task.

6.1 Methodology

6.1.1 Participants

Forty native speakers of New Zealand English took part in the same-different task. One participant had to be excluded right away as they did not respond in any trial. The mean age of the remaining 39 subjects was 23.0 years, ranging from 18 to 39. Six participants identified as multilingual. The experiment took place at the University of Canterbury, Christchurch, New Zealand, from December 2020 to March 2021.

6.1.2 Materials

The speech materials consisted of pseudowords as well as of real words and real word filler items. As the aim of the present experiment was to study the perception of word-final /s/, only those pseudowords with word-final /s/ were used. The 24 pseudowords used as stimuli were introduced in Section 3.1.2. For reasons of convenience, Table 6.1 lists these pseudowords once more.

The set of twelve real words used in this experiment was also introduced in Section 3.1.2. Recall that words were taken from the British National Corpus (Davies 2004), following a number of criteria. That is, words had to have a word-final /s/ as part of a voiceless stop plus sibilant coda; they had to be either singular

Table 6.1: Orthographic (*orth.*) and phonological (*phon.*) representations of the pseudowords used in the same-different task.

	/glɪ/	/prʌ/	/pli:/	/clu:/	/blaʊ/	/gleɪ/
orth.	*glips*	*prups*	*pleeps*	*cloops*	*bloups*	*glaips*
phon.	/glɪps/	/prʌps/	/pli:ps/	/klu:ps/	/blaʊps/	/gleɪps/
orth.	*glits*	*pruts*	*pleets*	*cloots*	*blouts*	*glaits*
phon.	/glɪts/	/prʌts/	/pli:ts/	/klu:ts/	/blaʊts/	/gleɪts/
orth.	*gliks*	*pruks*	*pleeks*	*clooks*	*blouks*	*glaiks*
phon.	/glɪks/	/prʌks/	/pli:ks/	/klu:ks/	/blaʊks/	/gleɪks/
orth.	*glifs*	*prufs*	*pleefs*	*cloofs*	*bloufs*	*glaifs*
phon.	/glɪfs/	/prʌfs/	/pli:fs/	/klu:fs/	/blaʊfs/	/gleɪfs/

or plural nouns with one syllable; and the number of short monophthong, long monophthong, and diphthong nuclei had to be equally distributed across words for both singular and plural nouns. For singular /s/, it was not possible to fully meet the final criterion as there was only one word with a long monophthong nucleus. Another monomorphemic word with a short monophthong was used instead. The set of real words is given in Table 6.2.

Table 6.2: Real words used in the same-different task.

Non-morphemic /s/		Plural suffix /s/	
Item	Vowel quality	Item	Vowel quality
mix	short	*books*	short
box	short	*steps*	short
tax	short	*rights*	diphthong
coax	diphthong	*points*	diphthong
hoax	diphthong	*groups*	long
corpse	long	*parts*	long

Additionally, twelve filler items were employed. All filler items were singular nouns consisting of a single syllable with either a short monophthong, a long monophthong, or a diphthong as nucleus. The nucleus type followed the same distribution as for the items described above, i.e. one third of filler items per type of nucleus. Half of the filler items ended in /f/, while the other half ended in /θ/. See Table 6.3 for all filler items.

The recording of the speech materials took place at a soundproof booth of the Department of Linguistics at the University of Tübingen. For this, reading lists

Table 6.3: Filler items used in the same-different task.

Non-morphemic /s/		Plural suffix /s/	
Item	Vowel quality	Item	Vowel quality
riff	short	*death*	short
muff	short	*myth*	short
wife	diphthong	*faith*	diphthong
safe	diphthong	*growth*	diphthong
grief	diphthong	*booth*	long
hoof	long	*path*	long

were created. On these lists, items were embedded within the sentence "He said *item* to me.". A trained native speaker of New Zealand English read the entire reading list aloud for practice before recording the list three times. The recordings were sampled at 44.1 kHz, 16 bit.

For each item the best of the three recordings was chosen by manual inspection. First, all recordings were analysed using Praat following the segmentation conventions laid out in Section 4.1.4. Recordings with production errors, e.g. laughter, stutter or vocal fry, or segmentation difficulties were dismissed. Second, the remaining segmented target and filler items were spliced from their surrounding contexts, resulting in audio files only containing the words of interest. Third, the duration of the items and filler items was measured using a Praat script (de Jong & Wempe 2008) and then analysed in R. The result of this analysis is given as the mean durations presented in Table 6.4. Lastly, for each item the version closest to the mean duration of its nucleus type was chosen for further use in the experiment to keep durational differences between items to a minimum.

Table 6.4: Mean durations of items and filler items across recordings in seconds.

Item type		Short vowel	Long vowel	Diphthong
real words	*mean*	0.576	0.613	0.572
	sd	0.109	0.102	0.062
pseudowords	*mean*	0.521	0.551	0.550
	sd	0.060	0.042	0.046
filler	*mean*	0.455	0.490	0.549
	sd	0.052	0.067	0.071

In a next step, the final /s/ duration of all items was manipulated in such a way that it corresponded to the mean /s/ duration for non-morphemic and plural /s/ found in the reference study by Plag et al. (2017). For example, in the case of *mix* the duration of the final /s/ was changed to 318 ms, while in the case of *books* the duration of the final /s/ was changed to 283 ms. This was done for all items, i.e. real words and pseudowords.

Pseudowords were treated as both singular and plural nouns. That is, pseudowords were equally distributed across four groups as follows. First, each group consisted of at least one pseudoword ending in /ps/, /ts/, /ks/, and /fs/. Second, groups A and B had two additional pseudowords ending in /ps/ and /ts/, respectively, while groups C and D had two additional pseudowords ending in /ks/ and /fs/, respectively. See Table 6.5 for the distribution of pseudowords across groups.

Table 6.5: Pseudoword distribution across the groups A-D used in the same-different task.

	Group A	Group B	Group C	Group D
gli-	ps	ts	ks	fs
plee-	ts	ps	fs	ks
cloo-	ks	ts	ps	fs
pru-	fs	ks	ts	ps
blou-	ps	fs	ks	ts
glai-	ts	ps	fs	ks

The pseudowords in groups A and C were treated as singular nouns with a non-morphemic word-final /s/, while the pseudowords in groups B and D were treated as plural nouns with a plural word-final /s/. Their /s/ durations were changed accordingly. This way of handling type of /s/ across pseudowords was chosen to keep priming effects across pseudowords to a minimum, i.e. no participant was to encounter pseudowords with both singular and plural /s/ durations.

Then, four altered versions of each modified item were created. Each non-morphemic /s/ item was edited in such a way that 10 ms, 20 ms, 35 ms, or 75 ms were subtracted from the word-final /s/ duration, making it gradually more similar to plural word-final /s/ in terms of its duration. For plural /s/ items, 10 ms, 20 ms, 35 ms, or 75 ms were added to the word-final /s/ duration, making it gradually more similar to non-morphemic word-final /s/ in terms of its duration. This resulted in five different /s/ durations per recorded item. See Table 6.6 for all final /s/ durations across non-morphemic and plural /s/ items. Depending on the pertinent item, the duration of the word-final /s/ took up more than half of the total word duration (cf. Table 6.4 and Table 6.6) and was notably longer than the original non-edited /s/ which showed a mean duration of 174 ms. Nonetheless, all items sounded natural. In total, five versions for each of the 12 real word items and for each of the 24 pseudoword items were created, resulting in 180 items. Each participant was to listen to 90 of them, i.e. (12 real words + 6 pseudowords) × 5 versions.

Table 6.6: Durations in milliseconds for non-morphemic and plural /s/ for real word and pseudoword items.

	Mean	± 10 ms	± 20 ms	± 35 ms	± 75 ms
non-morphemic	318	308	298	283	243
plural	283	293	303	318	358

A similar approach was used for the manipulation of filler items. Their final fricative duration was altered as well. For this, the mean duration of word-final /f/ and /θ/ was measured after extracting the recorded filler items from their contexts. It was found that the mean duration of word-final /f/ was 244 ms, while the mean duration of /θ/ was 217 ms. It was therefore decided that the duration of /f/ was treated similarly to that of non-morphemic /s/, i.e. it was shortened, while the duration of /θ/ was treated similarly to that of plural /s/, i.e. it was lengthened. The different durations for both /f/ and /θ/ are given in Table 6.7.

Table 6.7: Durations in milliseconds for /f/ and /θ/ for filler items.

	Mean	± 10 ms	± 20 ms	± 35 ms	± 75 ms
/f/ filler items	244	234	224	209	169
/θ/ filler items	217	227	237	252	292

6.1.3 Procedure

The same-different task was conducted in OpenSesame (Mathôt et al. 2012). First, participants were introduced to the same-different task. They were told that during the following experiment, they were to hear two recordings of the same word at a time and that they had to decide whether these two recordings were identical or different. It was explained that they should decide as quickly as possible and answer by pressing either the *same* or *different* key on the keyboard. The key assigned to *same* was “A”, the key assigned to *different* was “K”. The “A” key was pressed using the left index finger, the “K” key was pressed using the right index finger. Both options were given on screen during the entire experiment as illustrated by Figure 6.1. The participants were also told that if they did not decide on either option within a certain amount of time, the next trial would start automatically. Each participant started with ten practice trials, which consisted of six pseudoword items for familiarisation and four filler items.

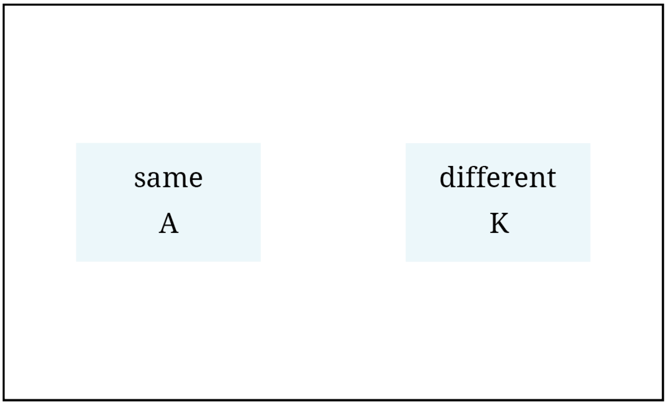


Figure 6.1: Option display during the perception experiment.

Each trial was preceded by a fixation cross and a stretch of silence of 450 ms. Then, both recordings played, with reaction time measurement starting at the onset of the second recording. The word-final /s/ of both recordings was either similar or different in duration, following one of the possible combinations given in Table 6.8, resulting in a trial number of 270, i.e. (12 real word items + 6 pseudoword items + 12 filler items) × 9 combinations. The experiment was split into four main parts to allow for short pauses. Each part consisted of either 67 or 68 trials (67 × 2 + 68 × 2 = 270). Within each of the four parts, target and filler items were distributed evenly but pseudorandomised, i.e. it was prevented that two trials in a row neither contained the same target or filler item nor the same combination of /s/ durations.

Participants were given a 2,000 ms window to react, starting after the offset of the second recording. After that a time-out was recorded. The next trial automatically started 2,500 ms after the offset of the second recording if no reaction was recorded.

Table 6.8: Combinations of /s/ durations used in the same-different task. *mean* is the mean duration found in Plag et al. (2017) for non-morphemic and plural /s/. ± represents a subtraction for non-morphemic /s/ items and an addition for plural /s/ items.

Same/different	/s/ durations of items	Same/different	/s/ durations of items
same	mean vs. mean	different	mean vs. mean ± 10 ms
same	mean ± 10 ms vs. mean ± 10 ms	different	mean vs. mean ± 20 ms
same	mean ± 20 ms vs. mean ± 20 ms	different	mean vs. mean ± 35 ms
same	mean ± 35 ms vs. mean ± 35 ms	different	mean vs. mean ± 75 ms
same	mean ± 75 ms vs. mean ± 75 ms		

6.2 Analysis

Data of same-different tasks are often analysed in terms of their error-rates (e.g. Belke & Meyer 2002; Norris & Kinoshita 2008; Lupker et al. 2018). For example, if a certain condition A shows a significantly higher error rate as compared to another condition B, it is concluded that perception of condition A is significantly worse. Figure 6.2 shows the overall error rates of the present same-different task results.

For a durational difference of 0 ms, the error rate is rather low with about 4%. For the 10 ms difference, the error rate is 96%; for the 20 ms difference, the error rate is 93%; for the 35 ms difference, the error rate is 91%; and for the 75 ms difference, the error rate is 62%. However, the overall results do not take into account inter-subject differences. It may very well be the case that some participants are more sensitive to durational differences or that some participants simply were more motivated to deliver a good performance. Figure 6.3 shows the overall results for all participants.

One can clearly see that some participants outperform others. For example, participant *s035* already improves their error rate at a difference of 20 ms, while participant *s030* shows virtually no correct responses for durational differences between 10 ms and 35 ms, and only some for 75 ms.

One possible way to proceed from these descriptive findings is to fit a statistical model to the data. However, as has become visible, there are clear differences between subjects, which points towards another issue: Individuals may have different levels of conservativity. That is, a more conservative participant will less

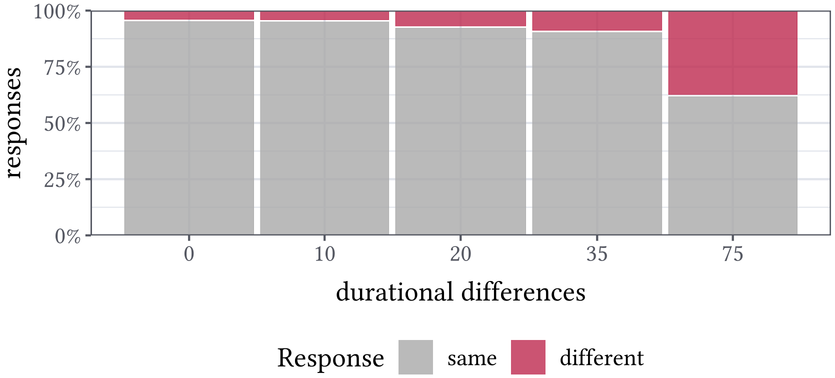


Figure 6.2: Overall error rates for the same-different task for all durational differences and across all subjects. For a durational difference of 0 ms the error rate is represented by the part of the bar corresponding to `different`, while for all other durational differences the error rate is the given part of the pertinent bars corresponding to `same`.

often respond with *different*, while a less conservative participant will more often respond with *different*, irrespective of the stimuli they hear. This intra-subject bias is neglected if one was to use the raw data, as was done in this section thus far.

A common way to factor in this participant bias is to make use of Signal Detection Theory (e.g. Macmillan 1993; Macmillan & Creelman 2005) and its measures. Signal Detection Theory can be applied in the analysis of any experiment in which two possible stimulus types are to be discriminated, i.e. in which error rates are the dependent variable of interest. The different measures of Signal Detection Theory have been used to analyse, among other things, recognition memory, lie detection, personnel selection, jury decision-making, medical diagnosis, industrial inspection, information retrieval, and congenital amusia (e.g. Stanislaw & Todorov 1999; Pfeifer & Hamann 2018). Signal Detection Theory makes use of *all four cells* of discriminative results as illustrated in the toy example in Table 6.9.

To calculate the most commonly used Signal Detection Theory measure, a bias-free measure of subject sensitivity called d', one must first calculate the hit rate H

$$H = \frac{HIT}{HIT + MISS} \tag{6.1}$$

and the false alarm rate F

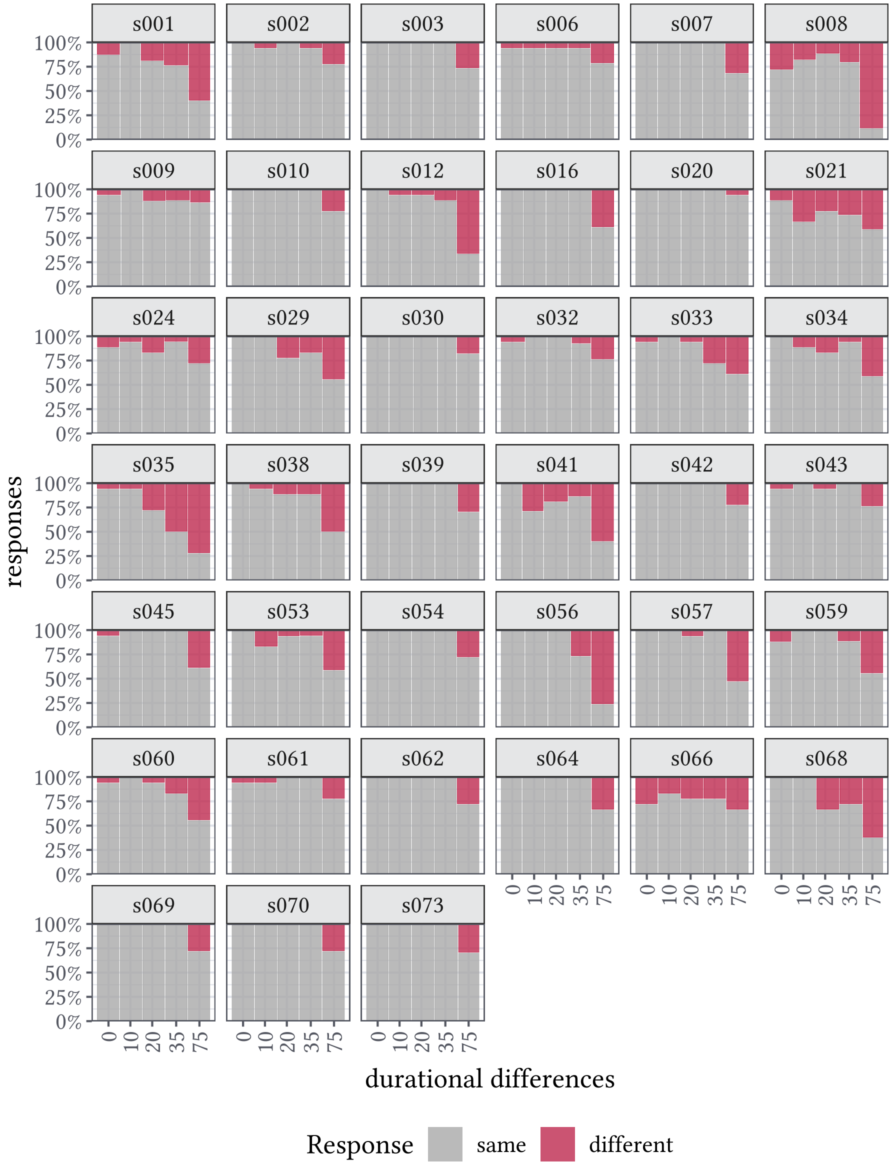


Figure 6.3: Error rates per subject for the same-different task for all durational differences. For a durational difference of 0 ms the error rate is represented by the part of the bar corresponding to `different`, while for all other durational differences the error rate is the given part of the pertinent bars corresponding to `same`.

Table 6.9: Types of results for a discriminative task as described by type of stimulus and type of response. Values illustrate a toy example.

	response: different		response: same	
stimulus: different	HIT	20	MISS	5
stimulus: same	FALSE ALARM	10	CORRECT REJECTION	15

$$F = \frac{FALSE\ ALARM}{FALSE\ ALARM + CORRECT\ REJECTION}\ . \tag{6.2}$$

Then, d' can be computed as

$$d' = z(H) - z(F) \tag{6.3}$$

where $z(.)$ is the Z-transform of either variable. However, d' can only be meaningfully used if two assumptions regarding the decision variable are met (Stanislaw & Todorov 1999). First, the signal and noise distributions are both normal. Second, the signal and noise distributions have the same standard deviation. In the present case, noise is equivalent to trials with two identical stimuli. If one of the assumptions is violated, d' will vary with the response bias (Stanislaw & Todorov 1999). Thus, it was decided to use an alternative measure, A', instead. A' is a nonparametric variant of d' (Pollack & Norman 1964) and its values range between 0 and 1, where higher values indicate higher sensitivity, and 1 indicates perfect performance. H and F, as introduced above, are also used to calculate A':

$$A' = 0.5 + \left[sign(H-F)\frac{(H-F)^2 + |H-F|}{4max(H,F) - 4HF} \right], \tag{6.4}$$

where the term $sign(H - F)$ is +1 if $H - F > 0$, 0 if $H = F$, and −1 otherwise. $max\ (H, F)$ equals either H or F, whichever is greater (Stanislaw & Todorov 1999). For the above toy example given in Table 6.9, A' then is

$$A' = 0.5 + \left[\frac{(0.8-0.4)^2 + |0.8-0.4|}{4 * 0.8 - 4 * 0.8 * 0.4} \right], \tag{6.5}$$

that is, A' has a value of about 0.79. Thus, in the example, sensitivity is quite high.

In the following sections, I will first introduce the covariates used in the analysis of the same-different task data. Then, I will present the analysis of the data, including the calculation of A' values from the raw data, and the statistical modelling of A' as dependent variable.

6.2.1 Covariates

The set of covariates used in the analysis of the subject sensitivity data calculated from the same-different task results is more restricted than other sets of covariates in this book. As A' values are calculated across all trials of a subject, item specific variables such as typeOfS (`non-morphemic` versus `plural`) and typeOfWord (`real word` versus `pseudoword`) cannot be used as covariates. However, analysing the raw data with chi-square tests strongly suggests that no significant difference for these variables were found ($p > 0.05$ for all comparisons; see the supplementary material given in Chapter 11). As sensitivity may very well vary between subjects, an additional covariate on how regularly subjects play musical instruments was introduced. In the following, covariates used in previous studies of this book are described first. For these, definitions are briefly repeated for convenience and adapted to perception where necessary. Then, the newly introduced covariate is given. Finally, the covariate used as random effect is listed.

age. Subjects' age was included as it may show an influence on hearing capabilities, with older subjects often experiencing a loss of hearing (e.g. Lee 2013).

monoMultilingual. To account for potential influences of other L1s besides English, the binary covariate monoMultilingual was introduced.

musicalInstrument. It has been shown that advanced players of musical instruments show an increased performance of phonological perception and of detecting durational differences in speech (e.g. Anvari et al. 2002; Milovanov et al. 2009). Thus, information on how regularly each subject plays a musical instrument was collected.

subject. subject ID was included to account for inter-speaker differences in perception.

Closer inspection of the newly introduced covariate, musicalInstrument, revealed that there was an uneven distribution of subjects across levels. That is, only 5% (n = 8) of trials had very often as value for musicalInstrument, while 41% (n = 64) had never as value. This skewed distribution is maintained by the levels in between, with 8% often (n = 12), 20% sometimes (n = 32), and 26% rarely (n = 40). Due to the skewed distribution and the therefore small amount of data points for some levels, it was decided to drop musicalInstrument as a covariate. If there was an effect of musicalInstrument nonetheless, this should then be indirectly considered as part of the subject random effect.

6.2.2 Overview of the data

An overview of all variables used in the analysis of subject sensitivity and their distribution is given in Table 6.10.

Table 6.10: Summary of the dependent variable and the numerical and categorical predictors in the final data set.

Dependent variable	Mean	St. Dev.	Min	Max
aprime	0.323	0.131	0.226	0.904
Numerical predictors	Mean	St. Dev.	Min	Max
age	23.000	5.235	18.000	39.000
Categorical predictors	Levels			
monoMultilingual	`monolingual`: 128		`multilingual`: 28	
subject	39			
Explanatory variable	Levels			
durDif	`10 ms`: 39	`20 ms`: 39	`35 ms`: 39	`75 ms`: 39

6.2.3 Modelling subject sensitivity

Using the formula for calculating A' as given in Equation 6.4 and as implemented by the `psycho` package for R (Makowski 2018), A' values for all subjects were computed. That is, for each subject, the results for the four durational differences 10 ms, 20 ms, 35 ms, and 75 ms were used to calculate an A' value. This resulted in four A' values per participant. The four durational differences are the predictor of interest in the regression modelling: durDif.

These A' values then entered a regression analysis as dependent variable. As A' assumes values in the standard unit interval $(0, 1)$, regression models such as LMERs or gaussian GAMMs are not sufficient, because such models do not take into account the interval constraint of the dependent variable. As a workaround, one could transform the A' values using, for example, a logit-transformation. However, this comes with several drawbacks (cf. Cribari-Neto & Zeileis 2010). It was thus decided to use beta regression as briefly introduced in Section 3.2.2 as the statistical tool of choice instead. Beta regression models assume that the dependent variable follows a beta distribution, i.e. that it assumes values in the open interval of $(0, 1)$. Commonly, beta regression in R is done using the `betareg` package (Cribari-Neto & Zeileis 2010). However, the `betareg` implementation does not allow for random effects in its model specification. As it was plausible to assume inter-subject differences in the given context, the `mgcv` package (Wood 2017) and its GAMM implementation were made use of instead. While the default for GAMMs is to assume a dependent variable of gaussian distribution, GAMMs can also be specified for dependent variables following a beta distribution. This is what I call BGAMMs (see Section 3.2.2).

A BGAMM was fitted with A' as dependent variable. The predictor of interest, durDif, and the covariate monoMulitlingual were included as parametric effects. age was included as smooth term and subject was specified as random smooth term. Following the procedure introduced in Section 3.2.2, the model was checked for issues of concurvity and of too few basis functions; no issues were found. The final data set as well as the analysis and results discussed in the following sections can be found in the supplementary material given in Chapter 11.

6.3 Results

A significant effect of durDif was found. Neither the effect of monoMulitlingual nor the effect of age reached significance. As anticipated, the random smooth of subject reached significance. This was to be expected due to the vast differences between subjects already found in the raw data. The results of the BGAMM fitted to the A' values are given in Table 6.11. For the parametric terms, I provide the β estimates and the corresponding standard errors (SE), z-values, and p-values. For the smooth terms, the estimated degrees of freedom, the reference degrees of freedom, the χ^2 values, and the p-values are given.

Figure 6.4 shows the partial effect of durDif. Participants show overall little sensitivity towards durational differences of 10 ms and 20 ms. For 35 ms a rather small but nonetheless significant increase in sensitivity is found as compared to the 10 ms difference. A clear increase in sensitivity is found for the durational difference of 75 ms as compared to the other differences. Thus, perceptibility of the 10 ms and 20 ms differences is rather low; the perceptibility of the 35 ms is significantly higher; and the perceptibility of the 75 ms difference is highest.

The overall significant differences in sensitivity are given in Table 6.12. Participants are significantly more sensitive towards the 75 ms difference as compared to all other durational differences.

As shown by the subject-specific A' estimates indicated by points in Figure 6.4, however, inter-subject differences remain high. Especially the biggest durational difference, 75 ms, shows a discernible amount of variation. The raw by-subject

Table 6.11: Summary of the BGAMM fitted to the *A′* values with durDif and monoMultilingual as parametric predictors, age as smooth term, and subject as random smooth term.

Parametric Terms	Estimate	SE	z value	Pr(|z|)
(Intercept)	-1.008	0.079	-12.818	0.000
durDif20	0.125	0.079	1.578	0.114
durDif35	0.206	0.077	2.676	0.007
durDif75	0.993	0.069	14.358	0.000
monoMultilingual	-0.089	0.137	-0.648	0.517
Smooth Terms	edf	Ref.df	Chi.sq	*p*-value
age	1.000	1.000	0.001	0.982
Random Smooth Terms	edf	Ref.df	Chi.sq	*p*-value
subject	28.400	36.000	144.904	0.000

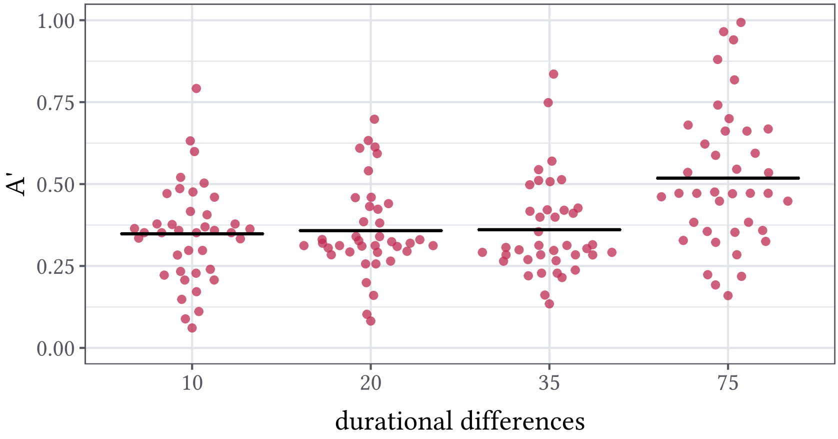


Figure 6.4: Partial effect of durDif as found by the BGAMM. The horizontal lines indicate the estimated *A′* mean for each durational difference; the points illustrate subject-specific estimates.

Table 6.12: Significant contrasts found for the different /s/ durations contrasted in the same-different task. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	10 ms	20 ms	35 ms	75 ms
10 ms	n.a.		**	***
20 ms		n.a.		***
35 ms			n.a.	***
75 ms				n.a.

A' values as illustrated in Figure 6.5 confirm the notion of high inter-subject variability. While some subjects show little increase in sensitivity between the 10 ms and 75 ms differences (e.g. subjects *s020* and *s030*), other subjects show a clear increase in sensitivity (e.g. subjects *s035* and *s056*). Overall, a higher A' value and thus sensitivity can be found for the 75 ms difference for most subjects.

6.4 Discussion

Following previous studies on the perception of subphonemic differences, the present study investigated whether the durational differences between different types of word-final /s/ are perceptible. As such, this is the first study to look into the perception of phonologically identical but morphologically and phonetically different segments. Since real words as well as pseudowords were used as items, potential lexical effects were taken into account. It was found that durational differences in word-final /s/ as small as 10 ms and 20 ms are overall not well perceptible. Durational differences of 35 ms and 75 ms show significantly increased perceptibility, while a durational difference of 75 ms by far shows the greatest perceptibility.

What does this mean for the perceptibility of durational differences found for different types of word-final /s/? The durational differences found in Plag et al. (2017) and in the production study of Chapter 4 are given in Table 6.13. None of the durational differences between the different types of /s/ is as high as 75 ms. However, considering the findings by Plag et al. (2017), one would expect the differences between the non-morphemic /s/ and morphemic types of /s/ to be somewhat perceptible as these differences are all at least equal to or bigger than 35 ms. Taking into account the findings of Chapter 4, only the durational difference between non-morphemic and clitic /s/ should be somewhat perceptible, as only this difference is close to or bigger than 35 ms. Considering both

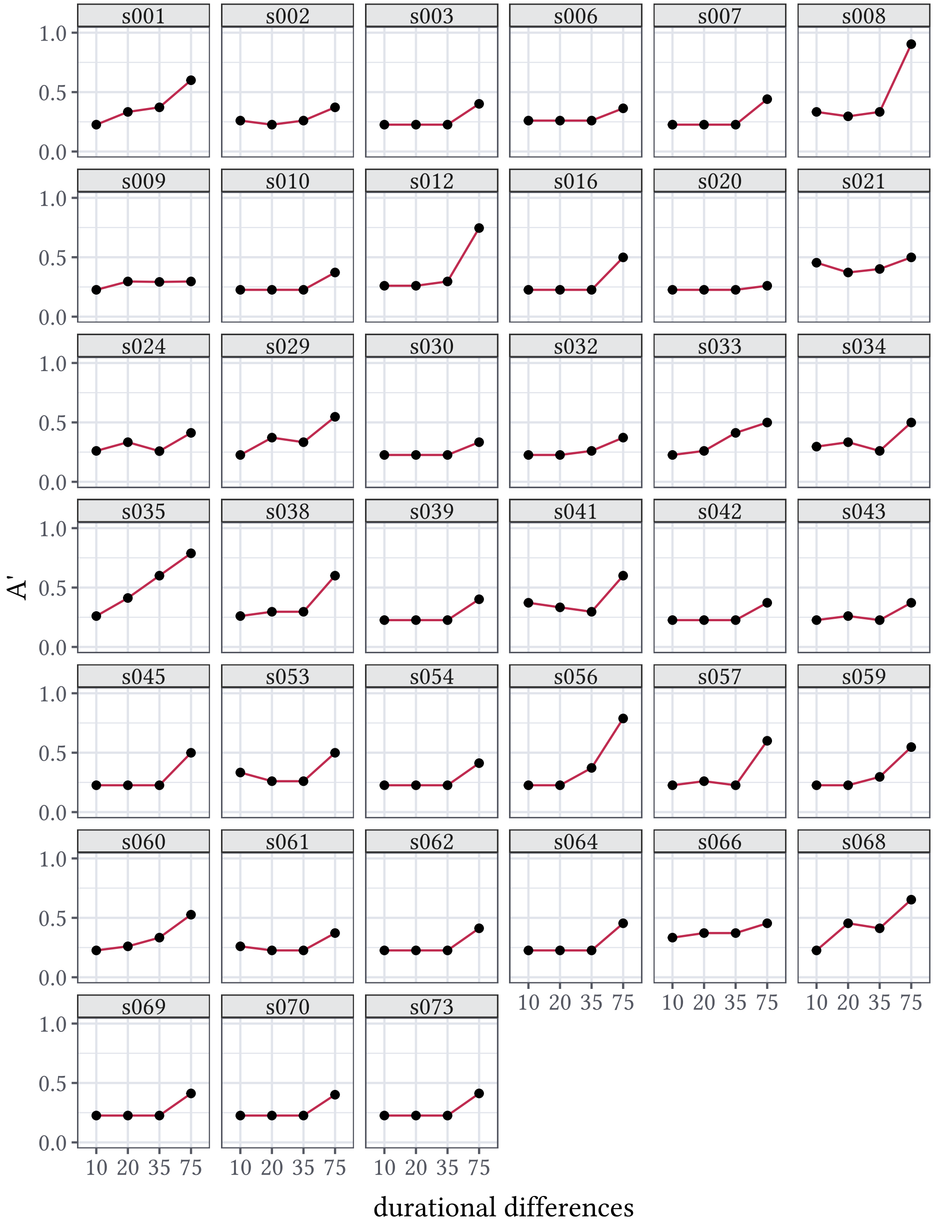


Figure 6.5: By-subject A' values across all durational differences.

studies (Plag et al. 2017 and Chapter 4), the findings indicate that at least some of the durational differences found between different types of /s/ are likely to be perceptible.

Table 6.13: Durational differences between non-morphemic, plural, is-, and has-clitic /s/ in milliseconds found in Plag et al. (2017) and the production study presented in Chapter 4.

		non-morphemic	plural	*is*-clitic	*has*-clitic
non-morphemic	Plag et al.	n.a.	35	57	65
	Chapter 4	n.a.	14	31	37
plural	Plag et al.		n.a.	22	30
	Chapter 4		n.a.	17	23
is-clitic	Plag et al.			n.a.	8
	Chapter 4			n.a.	6
has-clitic	Plag et al.				n.a.
	Chapter 4				n.a.

The significant increase in sensitivity of the 35 ms durational difference found in the present study is more or less in line with the findings by Klatt & Cooper (1975). Recall that in their experiment, the just-noticeable difference to be perceived was 25 ms. That is, a durational difference between the 20 ms and 35 ms difference. The sensitivity between these two durational differences showed a significant increase, thus indicating that the just-noticeable difference to be perceived most likely lies within this range.

However, an overall increase in perceptibility was only found for the durational difference of 75 ms, for which the difference in sensitivity is significant for all comparisons. While I cannot give a definitive answer to the question of why this is the case, I want to propose two considerations. First, fricatives such as /s/ are not only perceived in terms of their duration but also by their centre of gravity, spectral peak location, spectral moments, noise duration, amplitude, and other acoustic features. In the present study, only one of many features – duration – was controlled for and manipulated. Perceptibility might be higher if all acoustic features are manipulated accordingly. Second, in their study, Klatt & Cooper (1975) found that durational differences in word-final position and in fricatives are less well perceptible as compared to other positions and consonants. As the present study investigated differences between fricatives in word-final position, perceptibility was expected to be rather low.

Let us now turn to the theoretical implications of the present results. How do the results relate to the two hypotheses that were tested? H perc$_1$, the *Abstractionist Hypothesis*, assumes that listeners are not sensitive to subphonemic durational differences. As was illustrated, listeners show an increased sensitivity towards a durational difference of 35 ms and such a difference in duration was found between different types of word-final /s/ (e.g. Plag et al. 2017; Chapter 4). Also, none of the tested durational differences distinguishes between phonemes of English: No matter what its acoustic duration within a reasonable range, an /s/ is an /s/. Thus, the *Abstractionist Hypothesis* is rejected.

As listeners were sensitive to subphonemic durational differences, H perc$_2$, the *Phonetic Detail Hypothesis*, can potentially be confirmed. Assuming that fine-phonetic detail is perceived and stored, this hypothesis can most likely account for the present findings. Recent findings in neurobiology (Beach et al. 2021) are especially compatible with the notion of hybrid models, as are part of this hypothesis. That is, brain response patterns in same-different tasks suggest that the perception process does not require loss of subphonemic detail. Instead, the neural representation of perceived speech includes phonemic and subphonemic detail. Yet, a final decision on whether theories underlying this hypothesis can account for the present findings can only be reached with pertinent implementations.

The results of the present study then give rise to a further question: Are durational differences between different types of word-final /s/ made use of in comprehension? This question will be investigated in Chapters 7 and 8.

7 Comprehension of non-morphemic and plural /s/

As illustrated in detail in Section 2.2, two comprehension studies are part of this book. This chapter presents the first of these studies on the comprehension of subphonemic differences in word-final /s/. It makes use of real words in isolation with non-morphemic and plural word-final /s/ as items. Effects on comprehension were tested using a number-decision task in a mouse-tracking paradigm. Considering extant models and approaches of speech perception and comprehension, H COMP, the *Mismatch Hypothesis*, is investigated. That is, if listeners make use of subphonemic durational differences in the comprehension of different types of word-final /s/, then a mismatch of subphonemic detail and intended meaning is predicted to lead to a) slowed down comprehension processes, and b) deviated mouse trajectories.

7.1 Methdology

7.1.1 Participants

Forty native speakers of New Zealand English took part in the experiment. They were the same participants who also participated in the same-different task described in Chapter 6. As was the case for the perception experiment, one participant did not respond in any trials and was therefore excluded. The experiment took place at the University of Canterbury, Christchurch, New Zealand, from December 2020 to March 2021.

7.1.2 Materials

For the present experiment, only real words were used. Recall that words were taken from the British National Corpus (Davies 2004), following a number of criteria. That is, words had to have a word-final /s/ as part of a voiceless stop plus sibilant coda; they had to be either singular or plural nouns with one syllable; and the number of short monophthong, long monophthong, and diphthong nuclei had to be equally distributed across words for both singular and plural nouns. For singular /s/, it was not possible to fully meet the final criterion as there was only one word with a long monophthong nucleus. Another monomorphemic word with a short monophthong was used instead. As such words had already been sampled for the perception experiment in Chapter 6, that set of words was used here as well. Additionally, six new words for both singular and plural nouns were added to increase the overall amount of data without a repetition of items (see Winter & Grice 2021, on why repetitions are not desirable). Table 7.1 gives an overview of the complete set of words. As it was not possible to find more monomorphemic words with an even distribution of short monophthongs, long monophthongs, and diphthongs as nuclei, further monomorphemic words with a short monophthong nucleus were used instead.

Table 7.1: Words used as items in the number-decision task. The upper half of words is identical to the set of words used in the perception experiment. The lower half of words was added for the present comprehension experiment.

Non-morphemic /s/		Plural suffix /s/	
item	vowel quality	item	vowel quality
mix	short	*books*	short
box	short	*steps*	short
tax	short	*rights*	diphthong
coax	diphthong	*points*	diphthong
hoax	diphthong	*groups*	long
corpse	long	*parts*	long
lynx	short	*costs*	short
flux	short	*crusts*	short
wax	short	*rates*	diphthong
fax	short	*notes*	diphthong
lapse	short	*sports*	long
fox	short	*cheats*	long

The set of twenty-four target items was matched with a set of twenty-four filler items. Half of the filler items were high frequency monosyllabic singular words ending in any consonant but /s/. The other half of the filler items were disyllabic plurals ending in /ɪz/. The type of nucleus, i.e. short or long monophthong and diphthong, was distributed equally across both groups of fillers. All fillers used in the present experiment can be found in Table 7.2.

The recording of the speech materials took place at a soundproof booth of the Department of Linguistics at the University of Tübingen. For the recording

Table 7.2: Filler items used in the number-decision task.

High frequency singulars		/ɪz/ plurals	
item	vowel quality	item	vowel quality
end	short	*kisses*	short
fact	short	*fences*	short
head	short	*passes*	short
thing	short	*senses*	short
home	diphthong	*roses*	diphthong
point	diphthong	*houses*	diphthong
way	diphthong	*bases*	diphthong
side	diphthong	*spices*	diphthong
car	long	*classes*	long
world	long	*horses*	long
room	long	*nurses*	long
court	long	*uses*	long

procedure, reading lists were created. On these lists, target items were embedded within the sentence "He said *item* to me.", while filler items were embedded within the sentence "He said *item* again.". The latter sentence was used for filler items as some of them ended in alveolar stops, /d/ and /t/. Thus, the word *to* following the respective filler items would have potentially led to splicing problems later on due to coarticulatory effects, i.e. the omission of one of the two stops, between filler item and the following word. To keep differences due to phrasal context to a minimum, the decision was made to embed all filler items into the second sentence, including those without word-final alveolar stop. Target items were not embedded within the same sentence but within the one mentioned first, as for word-final /s/ a following stop simplifies the segmentation procedure due to the clear cut-off between friction and closure in the acoustic signal. Examples of target and filler items embedded in the pertinent sentences are given in (1) and (2), respectively.

(1) He said *hoax* to me.

(2) He said *world* again.

A trained native speaker of New Zealand English read the entire reading list aloud for practice before recording the list three times. The recordings were sampled at 44.1 kHz, 16 bit.

For each item the best of the three recordings was chosen by manual inspection. First, all recordings were analysed using Praat following the segmentation conventions laid out in Section 4.1.4. Recordings with production errors, e.g. laughter, stutter or vocal fry, or segmentation difficulties, e.g. the absence of a stop release, were dismissed. Second, the remaining segmented items and filler items were spliced from their surrounding contexts, resulting in audio files only containing the words of interest. Third, the duration of the target and filler items was measured using a Praat script (de Jong & Wempe 2008) and then analysed in R. The result of this analysis is given as the mean durations presented in Table 7.3. Lastly, the version closest to the mean duration of its nucleus type was chosen for further use in the experiment to keep durational differences between items to a minimum.

Table 7.3: Mean durations of items and filler items across recordings in seconds.

Item type		Short vowel	Long vowel	Diphthong
target items	*mean*	0.576	0.613	0.572
	sd	0.109	0.102	0.062
singular filler items	*mean*	0.469	0.467	0.523
	sd	0.082	0.035	0.071
plural filler items	*mean*	0.609	0.607	0.613
	sd	0.081	0.056	0.069

Next, for each target the chosen recording was edited so that the word-final /s/ was replaced with another word-final /s/. For recordings of the so-called *matched condition*, this new word-final /s/ was taken from another recording of the same word. If the /s/ was a non-morphemic /s/, its duration was manipulated in such a way that is corresponded to the mean non-morphemic /s/ duration found in Plag et al. (2017). If the /s/ was a plural /s/, its duration was changed to the mean plural /s/ duration found in the same study, accordingly. For recordings of the so-called *mismatched condition*, the new word-final /s/ was taken from a monomorphemic target in case of a plural base and from a plural target in case of a monomorphemic pseudo-base, i.e. the string of segments of a monomorphemic target without the word-final /s/. The duration of the /s/ was then manipulated in such a way that it corresponded to the mean duration found in Plag et al. (2017) for the other type of /s/. That is, a non-morphemic /s/ was changed to the duration of a plural /s/, and a plural /s/ was changed to the duration of a non-morphemic /s/. This procedure resulted in two recordings per target word, one of the matched condition and one of the mismatched condition. For example, for the monomorphemic target word *mix*, there was an audio stimulus with an /s/ duration of

318 ms for the matched condition and an audio stimulus with an /s/ duration of 283 ms for the mismatched condition. For a plural target like *books*, there was an audio stimulus with an /s/ duration of 283 ms for the matched condition and an audio stimulus with an /s/ duration of 318 ms for the mismatched condition. As the (pseudo-)base and the /s/ in both conditions have been spliced together, a one-sided effect of "sounding manipulated" on the experiment's results was ruled out.

In sum, each participant completed 72 trials, i.e. 12 matched non-morphemic /s/ items + 12 matched plural /s/ items + 12 mismatched non-morphemic /s/ items + 12 mismatched plural /s/ items + 12 high frequency singular items + 12 /ɪz/ plural items.

7.1.3 Procedure

The number-decision task was conducted in OpenSesame using the `mousetrap` plugin for mouse-tracking (Kieslich & Henninger 2017). Participants were introduced to the task at hand. They were told that in the following experiment they had to decide whether an audio recording was describing "one" or "two or more" entities. They were told to mouse-click on the corresponding button in the top right or top left corner of the screen as quickly as possible. Figure 7.1 illustrates what participants saw on screen for each trial. The participants were also told that if they did not decide on either option within a certain amount of time, the next trial would start automatically. Each participant started with six practice trials in which recordings of filler items were used (see the supplementary material given in Chapter 11).

Each trial was preceded by a stretch of silence of 450 ms. Then, one of the recordings was played, with reaction time and mouse-tracking measurement starting at the onset of the recording. Participants were given a window of 2000 ms starting after the onset of the recording to react, after that a time-out was recorded. The next trial started automatically 2500 ms after the onset of the recording if no reaction was recorded. Mouse-tracks were recorded with a frequency of 100 Hz.

7.2 Analysis

The data of the mouse-tracking experiment were analysed in terms of reaction times and mouse trajectories. In the following section, covariates used in the analyses are introduced. Section 7.2.2 then presents the analysis of reaction time data. The analysis of the mouse-tracks is given in Section 7.2.3.

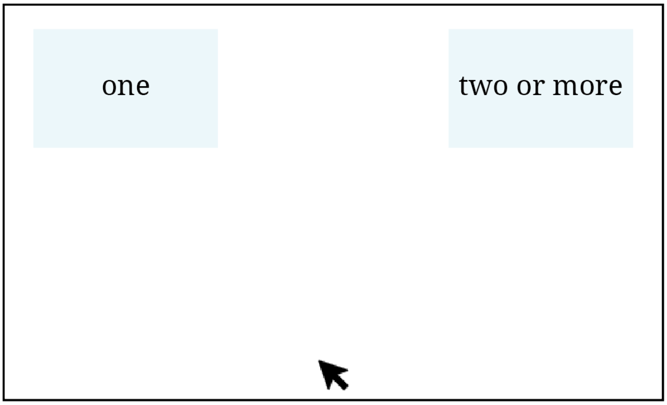


Figure 7.1: Option display during the comprehension experiment. The mouse cursor indicates the position the mouse was reset to in each trial.

7.2.1 Covariates

The set of covariates used in the analyses of the present study is similar to that of other studies on phonetic effects of morphological structure (Pluymaekers et al. 2005a,b; Hanique, Ernestus, et al. 2013; Plag et al. 2017; as well as those used in previous chapters of this book). Additionally, some further covariates, which may either influence perception or reactions based on perception, have been introduced. In the following, covariates based on previous studies on morphological structure are described first. For covariates which have been introduced in detail in Chapters 4 and 6, definitions are briefly repeated for convenience and adapted where necessary. Then, newly introduced covariates are given. Finally, covariates used as random effects are listed.

biphoneProbSum. A covariate based on the summed biphone probability was used as a measure of contextual predictability.

monoMultilingual. To account for potential influences of other L1s besides English, the binary covariate monoMultilingual was introduced.

age. Subjects' age was included as it may show an influence on reaction times, with older subjects generally reacting slower than younger subjects (e.g. Fozard et al. 1994).

neighbourhoodDensity. Neighbourhood densities were included as covariate as the number of neighbours may influence phonetic reduction (e.g. Gahl et al. 2012). The measure was created using the CLEARPOND database (Marian et al. 2012). neighbourhoodDensity describes the number of words differing in one segment from the item in question (Marian et al. 2012: 3).

TRIALNUMER. To account for possible effects of training and fatigue, the number of the trial during the experiment for each of the items per subject was included.

GOOGLEFREQLOG. To account for potential effects of frequency (e.g. Baayen et al. 2006; Keuleers et al. 2010; Brysbaert et al. 2011), Google frequency was included as covariate as it has been shown that Google frequencies are a robust predictor of reaction times (e.g. Hendrix & Sun 2020). The value of GOOGLEFREQLOG is the log-transformed number of Google search hits for each individual item as obtained on July 16, 2021.

TYPEOFS. This binary variable codes whether the pertinent pseudoword is a singular or plural form. It takes the value `nm` for pseudowords with a non-morphemic word-final /s/ and `pl` for pseudowords with a plural word-final /s/.

MUSICALINSTRUMENT. It has been shown that advanced players of musical instruments show an increased performance of phonological perception and of detecting durational differences in speech (Anvari et al. 2002; Milovanov et al. 2009). Thus, information on how regularly each subject plays a musical instrument was added as covariate.

CONDITION. The CONDITION variable is the explanatory variable of interest. Its levels are matched and mismatched and refer to the *matched* and *mismatched* conditions introduced by the creation of the audio stimuli. Recall from Section 7.1.2 that in matched stimuli (pseudo-)base and duration of the word-final /s/ match up, while there is a discrepancy of (pseudo-)base and word-final /s/ duration for mismatched stimuli.

CORRECT. CORRECT is a binary variable coding whether the answer clicked on by the subject in the relevant trial is the correct answer regarding the stimulus' (pseudo-)base.

DOMINANTHAND. Reaction times between the dominant and the non-dominant hand may differ (Gignac & Vernon 2004). The information of which hand was dominant in each subject was added as a covariate, as all participants used the same hand (i.e. their right hand) to use the mouse.

ORDER. This variable codes the order of X and Y coordinates, i.e. their chronological order in the observed mouse-tracks. ORDER was incorporated as a variable to account for the natural sequence of coordinates, i.e. to account for potential influences of auto-correlation.

VIDEOGAMES. It has been shown that playing video games can reduce reaction times (e.g. Dye et al. 2009). The relative frequency of how often a subject engages in playing video games was therefore included as a categorical covariate called VIDEOGAMES.

ITEM. For each item, its orthographic representation was contained as level of item. This covariate was used as a random effect to account for potential differences between individual targets not covered by other covariates.

SUBJECT. SUBJECT ID was included to account for inter-speaker differences in perception.

Closer inspection of the covariates describing subject characteristics, i.e. DOMINANTHAND, MONOMULTILINGUAL, MUSICALINSTRUMENT, and VIDEOGAMES, revealed that for all of these variables there was an uneven distribution of subjects across levels. That is, only 7% (n = 117) of trials had `left` for DOMINANTHAND, and only 19% (n = 302) of trials had `multilingual` for MONOMULTILINGUAL. For MUSICALINSTRUMENT and VIDEOGAMES, which both have five levels, the distribution is even more uneven. The level least represented in MUSICALINSTRUMENT is `very often` with 6% (n = 95), while the most represented level is `never` with 41% (n = 666). A similar picture is found for VIDEOGAMES, where the least represented level is `often` with 8% (n = 126), while the level most represented is `never` with 38% (n = 621). The uneven distribution of data points across variables and their levels also led to some "empty cells" within the possible combinations of levels across covariates. For example, all subjects for which the value of MONOMULTILINGUAL is `multilingual` have `right` as their DOMINANTHAND. There are no `multilingual` subjects who play a musical instrument `often` or `very often` and no `multilingual` subjects who `often` play VIDEOGAMES. Further, all `left`-handed subjects `never` play a MUSICALINSTRUMENT and they `rarely` or `never` play VIDEOGAMES. Due to this issue of sparse data and as for such variables with levels underrepresented in the sample it is unclear whether effects, found or not found, are due to a real effect of the variable or simply an artefact of chance. It was thus decided to drop the following covariates: DOMINANTHAND, MUSICALINSTRUMENT, and VIDEOGAMES. MONOMULTILINGUAL is retained for the analyses as the variable is directly related to language and because the variable has been used in other analyses of this book.

7.2.2 Reaction times

The present reaction time data were analysed using piece-wise additive mixed models (PAMMs; Bender et al. 2018, and as briefly introduced in Section 3.2.2). In the following, I will introduce the basics of PAMMs; the interested reader is referred to Hendrix & Sun (2020) for a more thorough introduction using linguistic data and to Bender et al. (2018) for a more detailed mathematical implementation.

PAMMs are a relatively novel technique of time-to-event analysis, that is they model the time until an event of interest occurs. The event of interest in the present number-decision task is the "one" or "two or more" response to a stimulus. Thus, the dependent variable in a PAMM is the instantaneous probability of a response as it evolves over time, not the reaction time itself (Hendrix & Sun 2020). Using PAMMs allows for an insight into the temporal dynamics of predictor effects. Hence, PAMMs do not only capture effects covering entire trials but also effects that occur only during particular parts of trials or show different effects during different parts of trials. A central function of time-to-event analysis is the probability density function $F(t)$:

$$F(t) = \int_{-\infty}^{t} f(x)dx = P(T \leq t). \tag{7.1}$$

The probability density function describes the probability that the response time T is smaller than or equal to a given time t. Closely related to the probability density function is the survival function:

$$S(t) = 1 - F(t) = P(T > t). \tag{7.2}$$

The survival function describes the probability of the time at which the event of interest occurs, T being greater than at a given time t. For the present experiment, the survival function describes the probability that subjects did not yet respond to a stimulus at time t. However, the mathematical properties of the function are not optimal for modelling purposes (Hendrix & Sun 2020). Thus, PAMMs make use of a closely related function, the hazard function. The hazard function describes the instantaneous probability that the event of interest occurs at time t, given that the event did not occur already. It is defined as

$$\lambda(t) = \lim_{dt \to \infty} \frac{P(t \leq T \leq t \,|\, T \geq t)}{dt} = -\frac{d}{dt} log(S(t)). \tag{7.3}$$

Before one can create PAMMs on reaction time data, though, the data have to be transformed. That is, the modelling of PAMMs requires data in the so-called piece-wise exponential data format (Bender & Scheipl 2018). While standard linear models (e.g. LMERs) or non-linear regression models (e.g. GAMMs) would use reaction times as dependent variable, PAMMs use the information on whether or not a stimulus was responded to at time t as dependent variable. The piece-wise exponential data format splits the time each stimulus is at risk of being responded to into J intervals. The intervals $(k_{j-1}, k_j], j = 1...J$ are defined by the cut points $K_0 < \cdots < K_J$. The choice of cut points is arbitrary (Hendrix & Sun 2020); over-fitting is prevented through penalisation of wiggliness (e.g. Wood 2017). t_j then equals k_j, i.e. t is derived from the defined cut points K. Following

Hendrix & Sun (2020), cut points at the extreme ends of the response time distribution were opted for. That is, cut points prior 770 ms, as only 11 trials (0.68%) were responded to earlier than 770 ms after stimulus onset, and after 1970 ms, as only 12 trials (0.74%) were responded to later than 1970 ms after stimulus onset, were excluded. An example of the transformed data is given in Table 7.4.

Table 7.4: Example of the piece-wise exponential data format for one stimulus instantiating the word *box*.

row	ID	item	tstart	tend	interval	offset	status
1	256	box	0.00	817.00	(0.00,817]	6.706	0
2	256	box	817.00	922.00	(817.00,922.00]	4.654	0
3	256	box	922.00	977.00	(922.00,977.00]	4.007	0
…	…	…	…	…	…	…	…
26	256	box	1345.92	1363.00	(1345.92,1363.00]	2.838	0
27	256	box	1363.00	1379.08	(1363.00,1379.08]	2.079	1

The piece-wise exponential data format contains a separate row for each interval for each stimulus. In each row, the start (tstart) and end point (tend) of the interval are given. The end points are included as predictor in a PAMM to estimate the hazard function over time. The offset variable provides information about the exact response time for each stimulus. For intervals in which no response was recorded (rows 1 to 26), the offset is the log-transformed duration of the interval, while for intervals in which a response was recorded (row 27), the offset is the log-transformed value of the difference of the exact time of response and the start of the interval, i.e. tstart. Please note the notation of the interval information: $(k_{j-1}, k_j]$ indicates that the first value, k_{j-1}, is included in the interval while the second value, k_j, is not. k_j, then, is the starting value of the following interval. The dependent variable in PAMMs, status, is a binary variable, which encodes whether a word was responded to (`1`) or not (`0`) in the pertinent interval.

A PAMM is then defined as follows:

$$\lambda(t|x_i) = \lambda_0(t_j) exp\left(\sum_{k=1}^{p} f_k(x_{i,k}, t_j) + b_{\ell_i}\right), \ \forall t \in (K_{j-1}, K_j] \tag{7.4}$$

with the predictor values x_i for stimulus i defining the hazard function $\lambda(t|x_i)$ at all time points t in the interval $j := (K_{j-1}, K_j]$. $\lambda_0(t_j)$ is the baseline hazard for time interval j, $f_k(x_{i,k}, t_j)$ are smooth functions for predictor $k \in 1, \ldots, p$ for each time point t in the interval j, and b_{ℓ_i} are random intercepts associated with group $l \in 1, \ldots, L$ to which stimulus i belongs (Hendrix & Sun 2020).

7.2.2.1 Overview of the data

An overview of all variables used in the PAMM modelling process and their distribution is given in Table 7.5 and Table 7.6.

Table 7.5: Summary of the dependent variable and the numerical predictors in the final data set.

Dependent variable	Levels			
status	0: 41471		1: 1616	
Numerical predictors	**Mean**	**St. Dev.**	**Min**	**Max**
biphoneProbSum	0.015	0.009	0.002	0.043
age	23.313	5.575	18.000	39.000
neighbourhoodDensity	17.068	10.039	1.000	34.000
trialNumber	36.358	20.678	1.000	72.000
googleFreqLog	9.190	0.787	7.658	10.302
tend	1234.455	212.414	817.000	1960.000

Table 7.6: Summary of the categorical predictors and the explanatory variable in the final data set.

Categorical predictors	Levels	
typeOfS	`nm`: 22159	`pl`: 20928
correct	`no`: 7795	`yes`: 35292
monoMultilingual	`monolingual`: 37578	`multilingual`: 5509
item	24	
subject	39	
Explanatory variable	**Levels**	
condition	`matched`: 21176	`mismatched`: 21911

7.2.2.2 Fitted models

A PAMM was fitted with status as dependent variable and biphoneProbSumBin, age, neighbourhoodDensity, trialNumer, and googleFreqLog as smooth terms. For each smooth, time-varying predictor effects were allowed for by including tensor product interactions between time, i.e. tend, and the predictor itself (see Wood 2017, for further details on tensor product interactions). To ensure interpretable results, the predictor smooths were limited to four basis functions, and time-by predictor interactions were limited to fourth order non-linearities. No limits were set on the smooth for time. The categorical covariates typeOfS, condition, correct, and monoMultilingual were included as parametric effects. The covariates item and subject were included as random smooth terms. Starting from this initial model, the modelling process proceeded as introduced in Section 3.2.2. It was found that the k-index value of the tensor product interaction of tend and trialNumber was 0.006. Recall that k-values well below 0.05 indicate potentially missed patterns in the residuals. Re-modelling with a limit of a sixth instead of a fourth order non-linearity resolved the issue. The prediction error curve (e.g. Mogensen et al. 2012) displaying the Brier score (e.g. Brier 1950; Gerds & Schumacher 2006; Bradley et al. 2008) of the final model is displayed in Figure 7.2. As a reference, the Brier score of the Kaplan-Meier estimate (Kaplan & Meier 1958) is given. The Brier score measures the accuracy of probabilistic predictions. The lower its value for a set of predictions, the better the predictions are calibrated. The range of possible Brier score values is $(0, 1)$. In the present case, the Brier score of the PAMM is considerably better than the Brier score of the Kaplan-Meier estimate. This indicates that the inclusion of the covariates in the PAMM improves the accuracy of the model predictions.

A valid question to ask when using novel statistical methods is whether the extra work is worth the trouble, i.e. whether the novel methods result in, for example, models with a higher fit. To answer this question for the present case, an LMER model with reaction time as dependent variable was fitted. As fixed effects, the parametric effects and smooth terms given in the PAMM formula (i.e. biphoneProbSum, age, neighbourhoodDensity, trialNumber, googleFreqLog, typeOfS, condition, correct and monoMultilingual) were specified. item and subject were included as random intercepts. The modelling process then followed the procedure introduced in Section 3.2.1. The final LMER model and the PAMM model were then compared by their AIC values: The AIC value of the LMER model was $21,811.36$, the AIC value of the PAMM was $15,360.92$. That is, the AIC value of the PAMM was smaller by $6,450.438$ points. Thus, the PAMM shows a significantly better fit. Regarding its model formula, the final LMER model only contained trialNumber and monoMultilingual as fixed effects and subject as random effect. To briefly foreshadow the results presented in the next section, the LMER model did not find a significant effect for age, while the PAMM did find a significant interaction of age and time (tend). This

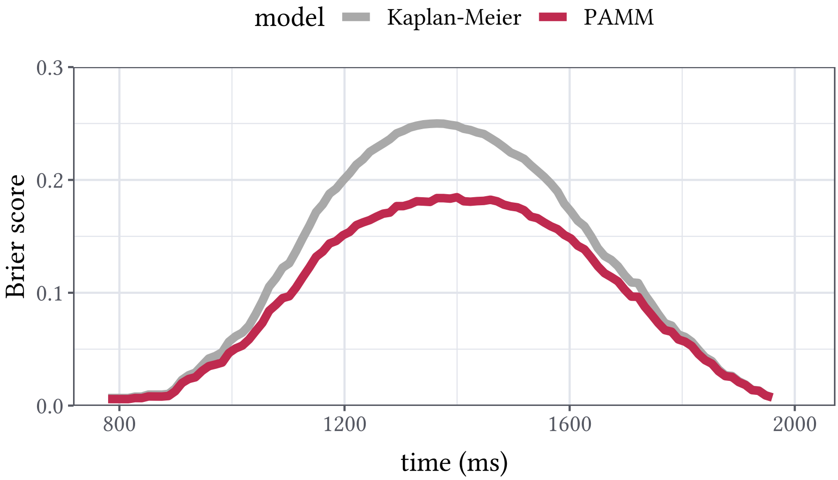


Figure 7.2: Comparison of the Brier scores of the fitted PAMM and its Kaplan-Meier estimate equivalent.

difference then denotes the potentially most prominent advantage of PAMMs: As mentioned in Section 7.2.2, PAMMs allow for an insight into the temporal dynamics of predictor effects, while LMER models fitted to the raw reaction time data do not. Overall, fitting PAMMs instead of LMER models appears to be worthwhile for the present data.

7.2.2.3 Results

Main effects of the following predictors were found: TEND, TRIALNUMBER, and MONOMULTILINGUAL. Additionally, the interactions between TEND and TRIAL and between TEND and AGE reached significance. The results of the PAMM fitted to the reaction time data are given in Table 7.7. For the parametric terms, I provide the β estimates and the corresponding standard errors (SE), *z*-values, and *p*-values. For the smooth terms, the estimated degrees of freedom, the reference degrees of freedom, the χ^2 values, and the *p*-values are given. The R script used for the analyses as well as the data set can be found in the supplementary material given in Chapter 11.

Figure 7.3 shows the distribution of raw reaction times for items in the matched and mismatched CONDITION. On average, matched stimuli are reacted to after 1374 ms, while mismatched stimuli are reacted to after 1388 ms.

Table 7.7: Summary of the PAMM fitted to status with typeOfS, condition, correct, and monoMultilingual as parametric effects, biphoneProbSum, age, NeighbourhoodDensity, trialNumber, and GoogleFreqLog as smooth terms, and item and subject as random smooth terms.

Parametric Terms	Estimate	SE	z value	p-value
(Intercept)	-6.867	0.152	-45.238	0.000
typeOfS`pl`	0.133	0.101	1.321	0.187
monoMultilingual`multilingual`	0.788	0.302	2.614	0.009
condition`mismatched`	-0.044	0.051	-0.858	0.391
correct`yes`	0.061	0.075	0.810	0.418
Smooth Terms	**edf**	**Ref.df**	**Chi.sq**	**p-value**
tend	7.830	8.653	1516.374	0.000
GoogleFreqLog	1.002	1.003	0.468	0.682
biphoneProbSum	1.001	1.002	0.989	0.321
NeighbourhoodDensity	1.300	1.495	1.760	0.391
age	1.002	1.002	3.258	0.071
trialNumber	1.001	1.002	64.063	0.000
Interactions	**edf**	**Ref.df**	**Chi.sq**	**p-value**
tend, GoogleFreqLog	1.352	1.613	0.580	0.753
tend, biphoneProbSum	1.539	1.882	1.238	0.423
tend, NeighbourhoodDensity	1.016	1.031	1.215	0.281
tend, age	5.942	7.321	20.257	0.007
tend, trialNumber	2.982	3.012	50.727	0.000
Random Smooth Terms	**edf**	**Ref.df**	**Chi.sq**	**p-value**
item	3.112	20.000	3.879	0.267
subject	34.184	36.000	505.162	0.000

Taking into account this rather small difference of 14 ms and the overall similarity of shape between the two RT distributions, it is not surprising that condition as a predictor did not reach significance in the PAMM. The significant effects found instead are explained in the following.

Panel A of Figure 7.4 shows the partial effect ($p < 0.001$) of the categorical variable monoMultilingual. It is found that `multilingual` subjects show a higher probability of earlier responses than `monolingual` subjects. This effect is visible in the distribution of the raw RT data (Panel B) as well. However, one should take this effect with caution as the number of `multilingual` subjects' data points (n = 302) is much smaller than the number of `monolingual` subjects' data points (n = 1326).

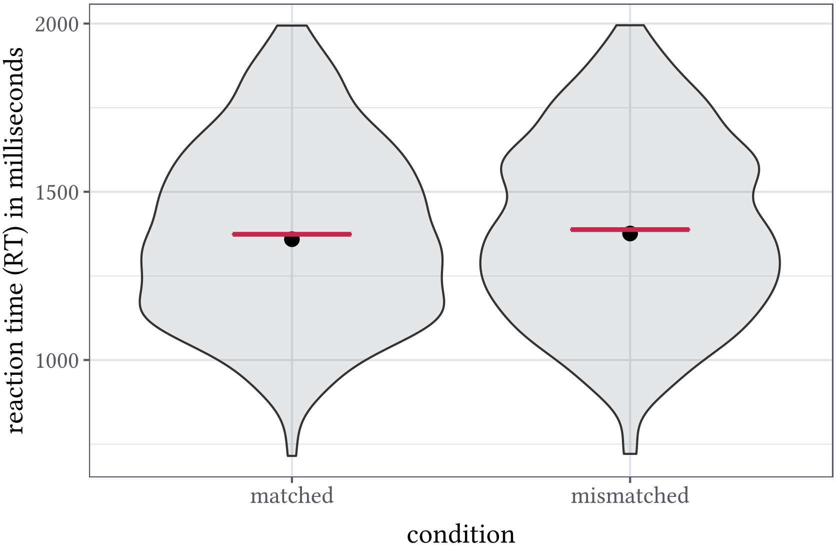


Figure 7.3: Observed reaction times for trials of matched and mismatched items. The dot represents the median, the horizontal line indicates the mean. The violin shapes represent rotated density plots describing the distribution of the data.

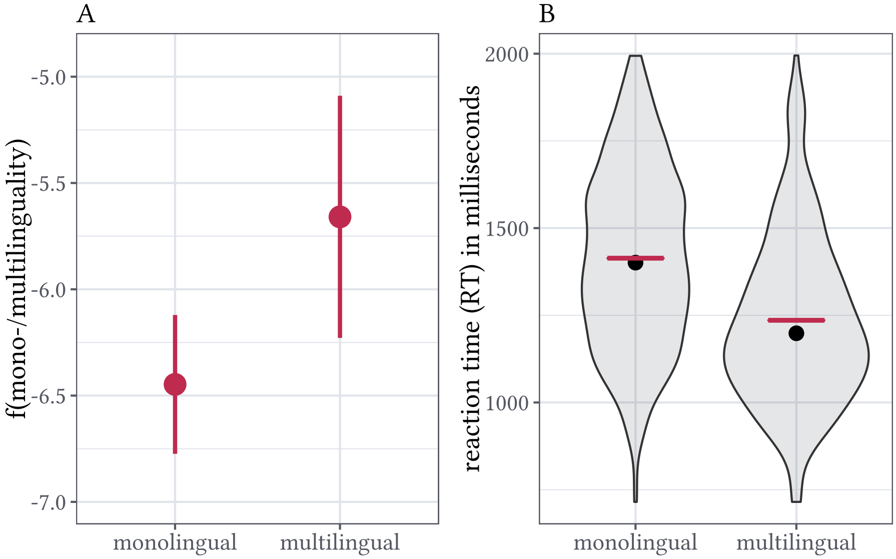


Figure 7.4: Partial main effect of monoMultilingual (A), and observed reaction times for monolingual and multilingual subjects (B).

A significant main effect of TRIALNUMBER (χ^2 = 63.890, $p < 0.001$) was found, as well as a significant interaction between time and TRIALNUMBER (χ^2 = 50.205, $p < 0.001$). The effect of TRIALNUMBER is modulated by the interaction with time as shown in Figure 7.5. Warmer colours indicate higher hazard rates.[1] That is, the interaction between TRIALNUMBER and time indicates that the increase of the instantaneous probability of a response for later trials is especially prominent during the early stages of the response window. Later on, the facilitatory main effect of TRIALNUMBER is offset by an opposite effect of the partial interaction between TRIALNUMBER and time.

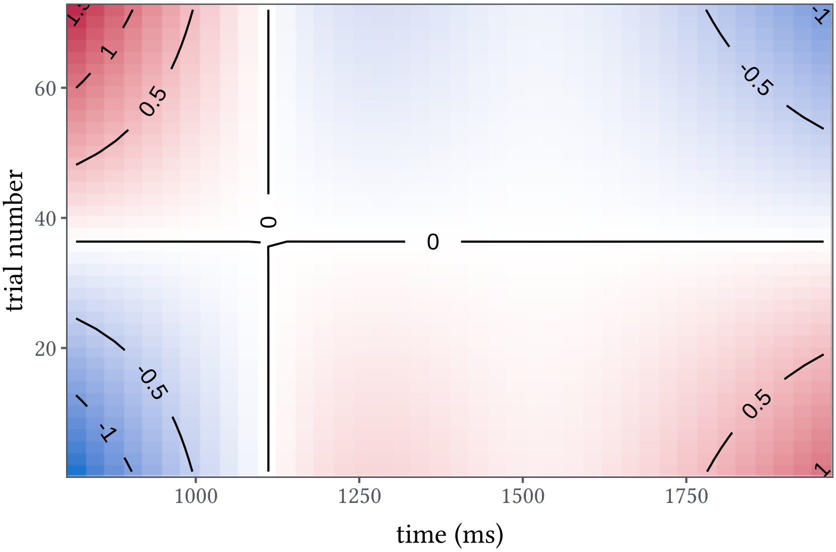


Figure 7.5: The effect of the interaction between TRIALNUMBER and time. Warmer colours indicate higher hazard rates.

Finally, a significant interaction between time and AGE (χ^2 = 20.151, $p < 0.05$) was found. This effect is illustrated in Figure 7.6. Again, warmer colours indicate higher hazard rates. That is, the interaction between AGE and time indicates that the increase of the instantaneous probability of a response for ages between approximately 23 and 28 years is especially prominent during the mid to late stages of the response window, i.e. around 1400 ms to 1750 ms into the trial. The grey area indicates ranges for which no or not enough data were available to the

[1] Note that readers of a black and white version of this book should rely on the numbers on the lines instead. Warmer colours correspond to positive values, while cooler colours correspond to negative values.

model.[2] As only few subjects (n = 4) contribute to the data above the grey area, shown effects should be interpreted with caution.

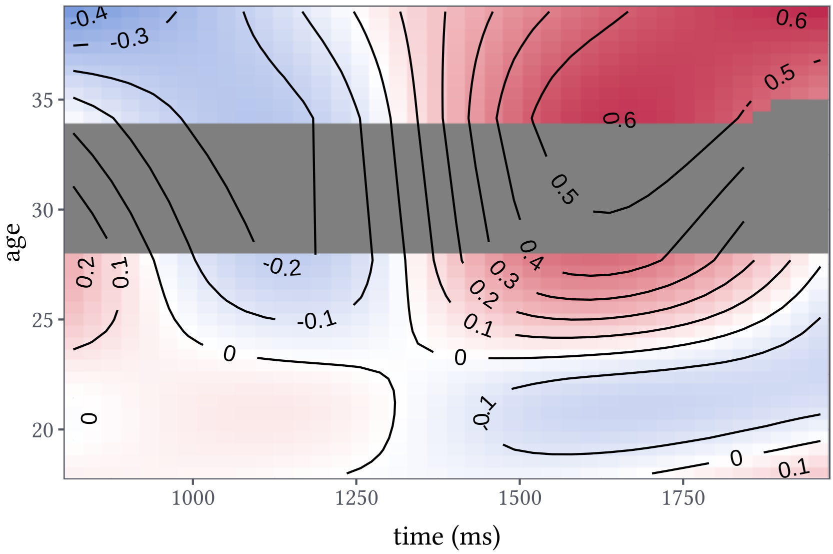


Figure 7.6: The effect of the interaction between AGE and time. Warmer colours indicate higher hazard rates.

7.2.2.4 Interim summary: Reaction times

Overall, CONDITION as a predictor did not reach significance in the PAMM. That is, participants responded with the same speed to matched and mismatched items. Instead, effects of MONOMULTILINGUAL, TRIALNUMBER, and AGE were found.

7.2.3 Mouse-tracks

Mouse-tracking data elicited in OpenSesame using the `mousetrap` plugin (Kieslich & Henninger 2017) were worked with in R using the `mousetrap` package (Kieslich et al. 2019). Following standard procedures, the raw mouse-tracking data were first transformed to the so-called *mousetrap data format* using the `mt_import_mousetrap` function. This function transforms the vectors of X and Y coordinates and their associated timestamps into meaningful row-by-row data for

[2] For readers of a black and white version of this book this area should be visible as dark grey area, which is almost shaped like a perfect rectangle.

further processing. Then, trials without mouse-movement were discarded. During the experiment subjects clicked on the right and left options on screen (see Section 7.1.3). To make mouse-tracks to both sides comparable, those towards the right option were mirrored vertically. Finally, all mouse-tracking data were time-normalised. Time-normalisation is commonly performed if the number of recorded X and Y coordinates varies across trajectories, which typically is the case for trajectories of differing reaction times. After time-normalisation with a constant number of equally sized time steps, all trajectories have the same number of recorded positions, and the positions at different relative time points can be compared across trajectories.

Figure 7.7 shows the mean trajectory of all spatially adjusted and time-normalised mouse-tracks used in the present analysis in the lower left panel. The panel on top displays the overall distribution of all X coordinates, with a clear peak around a value of 0. The panel on the right shows the overall distribution of all Y coordinates, with a clear peak around a value of about 380. The peaks correspond to the position to which the mouse was reset to for each trial.

As all mouse-tracks were spatially transformed, they all move towards the left option in the very end. Thus, one can also derive some further information from Figure 7.7. That is, taking into account the right part of the density plot of the X coordinates, it becomes visible that subjects in some trials must have deviated from a direct path. For example, if the final answer was the left option, at some point during the trial the mouse must have been on the (far) right part of the screen.

Figure 7.8 displays the average mouse-tracks for the variable of interest, CONDITION. Judging from the raw aggregated data alone, a difference between mouse-tracks of matched and mismatched trials is visible. In the following, I will explain how the statistical analysis of the mouse-tracking data investigating the difference between matched and mismatched trials was conducted.

Initially, regular Gaussian generalised additive mixed models were fitted to the X and Y coordinates of the mouse-track data. However, model criticism revealed that the fitted GAMMs showed rather problematic amounts of autocorrelation. Autocorrelation, or more specifically temporal spatial autocorrelation, is the association between data values over time. Depending on the sign of autocorrelation, model estimates can either be over- or underestimated (Charlton 2009). Thus, models with a high degree of autocorrelation are unreliable in their predictions. It was therefore decided to use QGAMs instead of GAMMs. QGAMs, as briefly introduced in Section 3.2.2, are additive quantile regression models. They are a distribution-free method for estimating the predicted values for any given quantile of the response distribution. As QGAMs are a relatively new tool within the

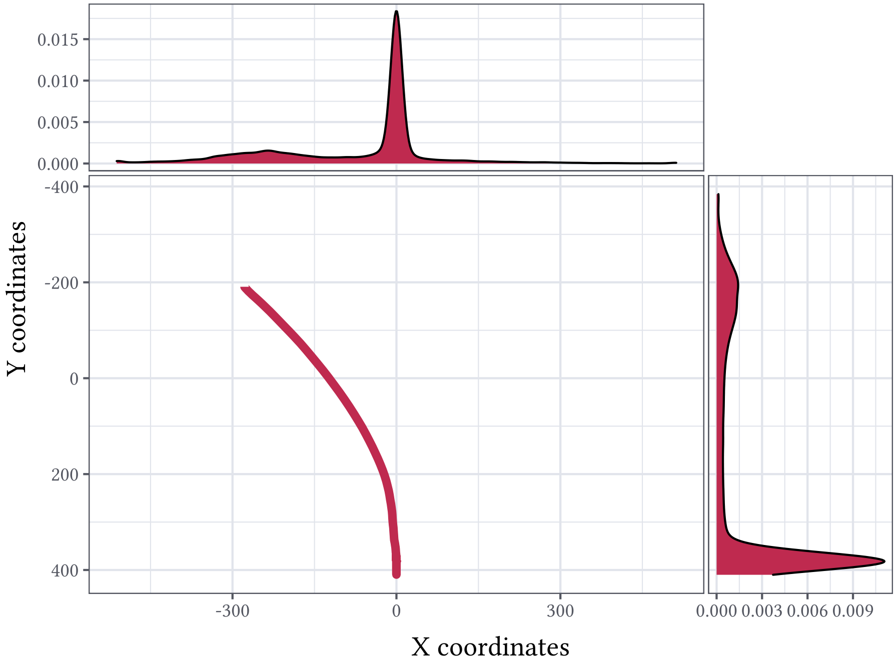


Figure 7.7: Mean trajectory of all spatially adjusted and time-normalised mouse-tracks (lower left), and density distribution of X and Y coordinates (on top and on the right, respectively).

toolbox of GAMs, I will explain the main characteristics of QGAMs as introduced by Fasiolo et al. (2021) in the following. The interested reader is referred to the aforementioned paper for a more thorough mathematical introduction.

Quantile regression, as conducted by QGAMs, aims at modelling the τth quantile of the response, y, conditionally on a p-dimensional vector of covariates, x, with k basis functions, where $\tau \in (0, 1)$. The τth quantile then is

$$\mu = \inf\{y : F(y|x) \geq \tau\}. \tag{7.5}$$

This can also be defined as the minimiser of the expected loss

$$L(\mu|x) = \int \rho_\tau(y - \mu) dF(y|x), \tag{7.6}$$

where the quantity $\rho_\tau(z)$ is the pinball loss (Koenker 2005; Gneiting 2011), which attributes different weights to observations depending on the sign of the residuals z:

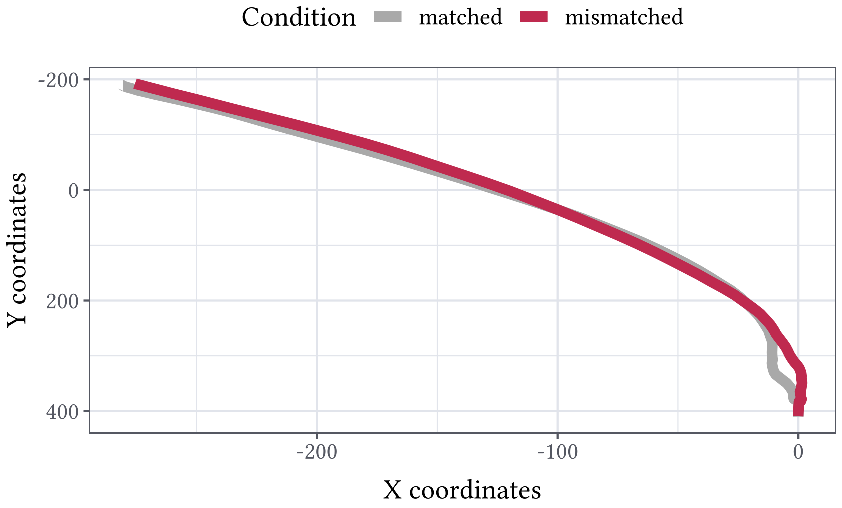


Figure 7.8: Mean trajectories of mouse-tracks for matched and mismatched item trials.

$$\rho_\tau(z) = \begin{cases} (\tau + 1)z & \text{if } z < 0 \\ \tau z & \text{if } z \geq 0 \end{cases}. \tag{7.7}$$

The quantile estimator is thus penalised to prevent overfitting, and the amount of penalisation is determined by the so-called learning rate, which determines the relative weight of the loss and the penalty. A QGAM then is defined as

$$\mu_\tau(i) = \beta_0 + \sum_{k=1}^{\rho} f_k(x_{i,k}) + b_{\ell_i}. \tag{7.8}$$

The term $\sum_{k=1}^{\rho} f_k(x_{i,k})$ can represent either a linear effect or a non-linear effect without a predefined structure. b_{ℓ_i} models random intercepts for group $\ell = 1, ..., L$ to which observation i belongs, and β_0 is the y-intercept.

In less technical terms, QGAMs make use of the general features of GAMs in modelling linear and non-linear effects as well as random effects. Instead of taking into account the whole range of data for their fitting process, each QGAM is restricted to a given conditional quantile of the data. Splitting the data into ten equally sized conditional quantiles, one can infer, for example, the following. Let us assume that the data are ordered from lowest to highest value, as is the case for the density plots of Figure 7.9 Then the so-called 0.1 quantile (Panel A) will consist of the first 10% of data, that is the tenth of the data consisting of the lowest values. The 0.5 quantile (Panel B), then, consists of the first 50% of the data, and the 0.9 quantile (Panel C) contains all data but the highest 10%.

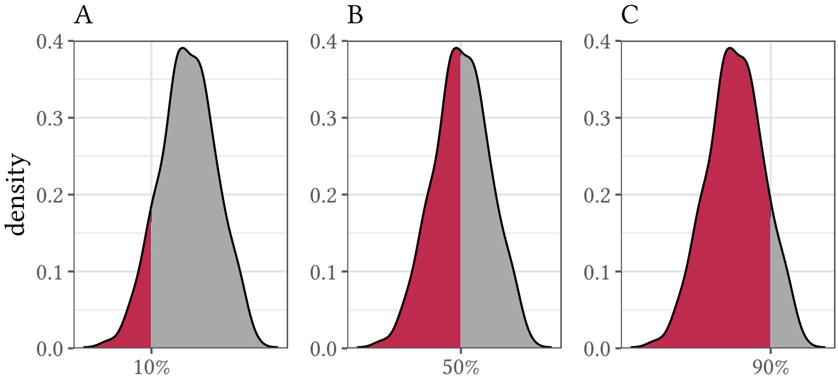


Figure 7.9: Illustration of the three conditional quantiles 0.1 (Panel A), 0.5 (Panel B), and 0.9 (Panel C) in blue.

Fitting QGAMs to several of these quantiles allows for a detailed picture of the effects to be investigated. If an effect is present in the 0.1 quantile but no longer present in the 0.3 quantile, for example, one can conclude that the effect is significant for data of the lowest 10% but loses its significance when taking into account higher valued data points. If the same effect then regains its significance in the 0.5 quantile, one can conclude the opposite. While there is no effect for the lowest 30% of data points, there again is an effect when including the following 20% of data points. QGAMs take into account all covariates specified in their model formula to arrive at their weighted conditional quantile distribution.

Before one can model X and Y coordinates of mouse-tracks as provided by the `mousetrap` plugin (Kieslich & Henninger 2017) for OpenSesame, the data have to be prepared. After the initial preparations mentioned at the beginning of this section, i.e. after time-normalisation and spatial transformation, the `mousetrap` package (Kieslich et al. 2019) provides coordinates and other data in R in a so-called mousetrap object. To extract the pertinent data needed, i.e. the X and Y coordinates, the time stamps corresponding to each coordinate value, as well as a unique identifier per trial, the `extract_x`, `extract_y`, and `extract_t` functions of the `mtqgam` package (Schmitz 2021b) were used.

Taking a closer look at the extracted coordinate data, it is found that the coordinate system as used per default by the `mousetrap` plugin (see, for example, axes in Figure 7.7 and Figure 7.8) is rather unintuitive for the Y dimension: Coordinates higher up on the screen show negative Y coordinate values, while coordinates lower down on the screen show positive Y coordinate values. For X coordinates, the system is more intuitive: Coordinates further to the right have positive X coordinate values, while coordinates further to the left have negative X coordinate values. Using this default coordinate system would result in an obscure order of conditional quantiles. Consider the 0.1 conditional quantile, which consists of the lowest 10% of data points, as an example. As Y shows lowest values for mouse positions high up on the screen and X shows lowest values for mouse positions further to the left, the first quantile would correspond to the end or near-end of the mouse-tracks instead of to their start. Analogously, the 0.9 quantile, then, would correspond to the start or near-start of the mouse-tracks. Thus, modelling the coordinate data with their default sign means doing things from back to front. For this reason, the original coordinate data's sign was reversed.

Merging the coordinate and time stamp data with the data on the set of covariates provided in Section 7.2.1, one can then model QGAMs. As mentioned previously, the coordinates used in the present analysis are time-normalised. For the current implementation, I chose a number of n = 140 time steps for the time-normalisation process. Kieslich et al. (2019) provide no reasoning on why the `mousetrap` package uses a number of n = 101 per default but refer to Spivey et al. (2005) instead. However, in Spivey et al. (2005) the number of time steps remains unmotivated. I arrived at n = 140 by arbitrarily taking the mean RT of all trials $\bar{x} \approx 1400$ ms and dividing it by 10. An example of the prepared data is given in Table 7.8.

Table 7.8: Example of the data format for a matched trial.

order	trialNumber	time	x_coordinate	y_coordinate	condition
1	1	0	0	-380.000	matched
…	…	…	…	…	…
66	1	585.4700	-2.000	-380.000	matched
67	1	594.4748	-2.000	-380.000	matched
68	1	603.4800	-2.348	-382.652	matched
…	…	…	…	…	…
110	1	972.7770	113.944	-124.724	matched
111	1	981.7840	121.141	-103.685	matched
…	…	…	…	…	…
130	1	1152.9210	238.000	216.000	matched
131	1	1233.9860	238.000	216.000	matched

For each trialNumber the prepared data set contains a separate row for each time step. The individual time steps are numbered in the variable order. The point in time of the time stamp is given in TIME. For each time stamp, the X and Y coordinates are contained in X_COORDINATE and Y_COORDINATE, respectively. CONDITION, then, is the previously introduced explanatory variable of interest, and its value is repeated for each row of a trial. The same is true for all other covariates of the data set (not shown in Table 7.8). In the above example, the first row is the very first coordinate pair recorded at time 0. The data of rows 66 to 68 show that even though time passed, the mouse was not moved at all (rows 66 to 67) or moved only slightly (rows 67 to 68). From rows 110 to 111, mouse movement is clearly visible in the X and Y coordinates. Finally, in rows 130 and 131 (and following), the target is reached. Thus, time continues to pass until time stamp number 140 is reached, while X and Y coordinates remain unchanged.

7.2.3.1 Fitted models

The complete set of data (n = 261,240) was split into two separate data sets depending on whether the (pseudo-)base of the target word belonged to a singular or plural noun. This resulted in two smaller data sets, with n = 142,380 for singular pseudo-bases and n = 118,860 for plural bases. This was done because the aim of the present analysis was to investigate whether a mismatch of (pseudo-)base and /s/ duration influenced the mouse-tracks. While this can also be found out with the complete data set, interactions of (pseudo-)base types and further covariates would have been a necessary part of the model formula. It was decided against using such multiple interactions as they make model interpretation more complex while offering basically the same insights as the implementation with split data sets. Moreover, fitting QGAMs is computationally costly, with near-exponentially increasing computation times for bigger data sets and more complex effect structures. Thus, choosing the implementation of several QGAMs for smaller data sets also kept the carbon footprint of the analysis down.

Both data sets were then further reduced by excluding trials which had been responded to incorrectly. While CORRECT was a potential covariate for modelling QGAMs, the difference between correctly and incorrectly answered trials is not the main interest of the present study. This decision led to an overall loss of n = 46,760 data points (17.9%), resulting in n = 102,480 for singular pseudo-bases and n = 112,000 for plural bases. An overview of all variables contained in the two data sets is given in Table 7.9 and Table 7.10.

For both data sets, two sets of QGAMs were fitted. One set of QGAMs was fitted to X coordinates, and one set of QGAMs was fitted to Y coordinates. I aimed at estimating the conditional quantiles corresponding to $\tau = 0.1, 0.3, 0.5, 0.7$ and 0.9. Thus, each set of QGAMs consisted of five QGAMs, one for each of the five

Table 7.9: Summary of the dependent variables and the numerical and categorical predictors in the singular pseudo-base data set.

Dependent variables	Mean	St. Dev.	Min	Max
x_coordinate	74.643	152.998	-511.000	512.000
y_coordinate	-176.898	250.926	-410.000	384.000
Numerical predictors	Mean	St. Dev.	Min	Max
order	70.500	40.414	1.000	140.000
Categorical predictors	Levels			
item	12			
subject	39			
Explanatory variable	Levels			
condition	`matched`: 50120		`mismatched`: 52360	

Table 7.10: Summary of the dependent variables and the numerical and categorical predictors in the plural base data set.

Dependent variables	Mean	St. Dev.	Min	Max
x_coordinate	75.413	150.252	-512.000	511.000
y_coordinate	-192.441	243.169	-410.000	384.000
Numerical predictors	Mean	St. Dev.	Min	Max
order	70.500	40.414	1.000	140.000
Categorical predictors	Levels			
item	12			
subject	39			
Explanatory variable	Levels			
condition	`matched`: 50120		`mismatched`: 52360	

quantiles. Taking into account more extreme quantiles, i.e. 0.1 and 0.9, as well as the median quantile 0.5 and the quantiles between the median and the extreme quantiles, i.e. 0.3 and 0.7, one obtains a detailed picture of how predictors affect the coordinate data. In total, ten QGAMs for each of the two data sets were fitted, that is five for X coordinates and five for Y coordinates. This resulted in a total number of twenty QGAMs.

The model formula for all QGAMs was similar. The dependent variable was either X_COORDINATE or Y_COORDINATE. CONDITION was introduced as parametric term. ORDER was given as smooth term with the default *k*-value of 9, and ITEM and SUBJECT were included as random smooth terms. The model formula was kept simple due to the extensive computational times of QGAMs.

All models were then checked according to the process introduced in Section 3.2.2. It was found that the *k*-index value of the ORDER smooth term was well below 0.05 for all QGAMs, thus indicating potentially missed patterns. Remodelling the set of QGAMs for X coordinates of plural bases as a test case with higher *k*-values ($k = 18, 30, 60$, and 120) revealed that no matter what the *k*-value, the general effect of all covariates remained unchanged. Following Wood (2017), it was therefore concluded that the *k*-value was large enough so that re-fitting all twenty computationally costly QGAMs was not necessary. The final data sets, as well as the analysis and results discussed in the following sections, can be found in the supplementary material given in Chapter 11.

7.2.3.2 Results

Across all twenty models, an effect of CONDITION was found 12 times. An effect of the ORDER smooth was found in all models. Similarly, the random smooths of ITEM and SUBJECT reached significance in all QGAMs. The overall model fit is high with a mean deviance explained of $\overline{D} = 70.74\%$. For both data sets, QGAMs fitted to Y coordinates show overall higher rates of deviance explained ($\overline{D} = 79.67\%$) than their X coordinate counterparts ($\overline{D} = 61.81\%$). For all four sets of QGAMs, the QGAM fitted to the 0.5 quantile shows the lowest rate of deviance explained ($\overline{D}_{0.5} = 61.15\%$), while the QGAMs fitted to the more extreme quantiles show the highest rates of deviance explained ($\overline{D}_{0.1} = 82.86\%$ and $\overline{D}_{0.9} = 79.63\%$).

The effects found in the QGAMs fitted to X and Y coordinates of the monomorphemic pseudo-base data set are displayed in Table 7.11. The model estimates of these and all following QGAMs are part of the supplementary material given in Chapter 11. Note that here and in the following, I will refrain from discussing the effects of the smooth terms as they are not the main interest of investigation.

There are significant effects of condition in four QGAMs fitted to the X coordinate data and in four QGAMs fitted to Y coordinate data. For $\tau = 0.3, 0.5, 0.7$ and 0.9 condition shows a significant effect for X and Y coordinates. The effects are illustrated in Figure 7.10. Where a significant effect is found, X coordinates are further to the right and Y coordinates are further down in the mismatched condition. Recall that QGAMs rely on conditional quantiles. Thus, the estimates shown in Figure 7.10 and similar plots illustrate the nature of an effect taking into account a certain quantity of the overall data (e.g. the lowest 10% of dependent variable data values in $\tau = 0.1$). The estimates do not illustrate the positions of mouse-tracks at certain points of their trajectory.

Table 7.11: Summary of the effects found in the QGAMs fitted to the X and Y coordinates of the monomorphemic pseudo-base data set. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	X coordinates					Y coordinates				
quantiles:	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Parametric Terms										
(Intercept)	***	***	***	***	***	***	***	***	***	**
conditionmismatched	n.s.	*	***	***	***	n.s.	***	***	***	***
Smooth Terms										
order	***	***	***	***	***	***	***	***	***	***
Random Smooth Terms										
item	***	***	***	***	***	***	***	***	***	***
subject	***	***	***	***	***	***	***	***	***	***

For the QGAMs fitted to the plural base data set, the effects found for X and Y coordinates are given in Table 7.12. condition reaches significance in four models fitted to the X coordinate data: $\tau = 0.1, 0.3, 0.5$ and 0.7. The effect is illustrated in Figure 7.11. Where condition shows a significant effect for X coordinates, coordinates of mismatched trials are further to the right. For Y coordinates, condition misses significance across all models.

7.2.3.3 Interim summary: Mouse-tracks

Across all QGAMs, a significant effect of condition emerged 12 times. Especially X coordinates are affected by condition, as 8 of the 12 significant effects are found in QGAMs fitted to X coordinate data. For Y coordinates, significant effects of condition are only found in the monomorphemic pseudo-base data set.

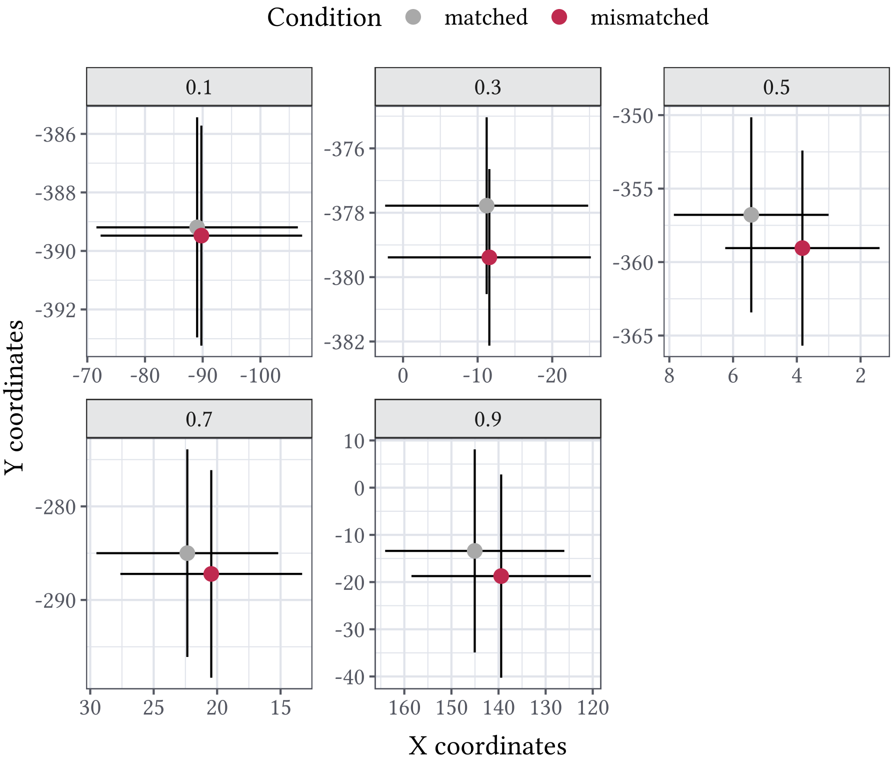


Figure 7.10: Effects of CONDITION as found in the QGAMs modelled to the X and Y coordinates of the monomorphemic pseudo-base data set. The lines indicate the confidence intervals of the estimated X and Y coordinate values.

Where a significant effect is found, coordinates of mismatched trials are further to the right and lower down.

7.3 Discussion

The present number-decision study set out to investigate if listeners make use of subphonemic durational differences in the comprehension of non-morphemic and plural word-final /s/. This question was analysed following H COMP, the *Mismatch Hypothesis*: If subphonemic durational differences are made use of, then a mismatch of subphonemic detail and intended meaning leads to a) slowed down comprehension processes and b) deviated mouse trajectories.

Table 7.12: Summary of the effects found in the QGAMs fitted to the X and Y coordinates of the plural base data set. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	X coordinates					Y coordinates				
quantiles:	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Parametric Terms										
(Intercept)	**	***	***	***	***	***	***	***	***	***
CONDITIONmismatched	**	***	***	**	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Smooth Terms										
ORDER	***	***	***	***	***	***	***	***	***	***
Random Smooth Terms										
ITEM	***	***	***	***	***	***	***	***	***	***
SUBJECT	***	***	***	***	***	***	***	***	***	***

Part a) of the hypothesis was tested by modelling the reaction time data in a PAMM. It was found that reaction times are not significantly influenced by the mismatch of durational information. As such, the first part of the hypothesis is rejected. That is, reaction times are similar for trials with matched and mismatched durational information. Thus, comprehension processes apparently are not slowed down.

Part b) of the hypothesis was investigated by fitting QGAMs to the X and Y coordinate data of the mouse-tracks recorded during the experiment. QGAMs were fitted separately for singular pseudo-bases and plural bases to achieve a way of direct comparisons between matched and mismatched /s/ trials within one type of (pseudo-)base. The results of the QGAMs show an overall significant effect of matched versus mismatched durational information on X coordinates. That is, X coordinates of trials with durationally mismatched items are significantly further to the right. For Y coordinates, significant effects were only found in the singular pseudo-base data: Mismatched trials come with Y coordinates which are further down.

How do these findings relate to the second part of the hypothesis? Looking at the results for X coordinates, which are further to the right in mismatched trials, one can interpret the findings as a confirmation of the hypothesis. Recall that mouse-tracks were mirrored where applicable, that is all tracks move towards the upper left corner of the coordinate system. Thus, an ideal non-deviated trajectory would be a straight line between the mouse cursor starting position and one of the answer options. As this non-deviated straight line moves linearly to-

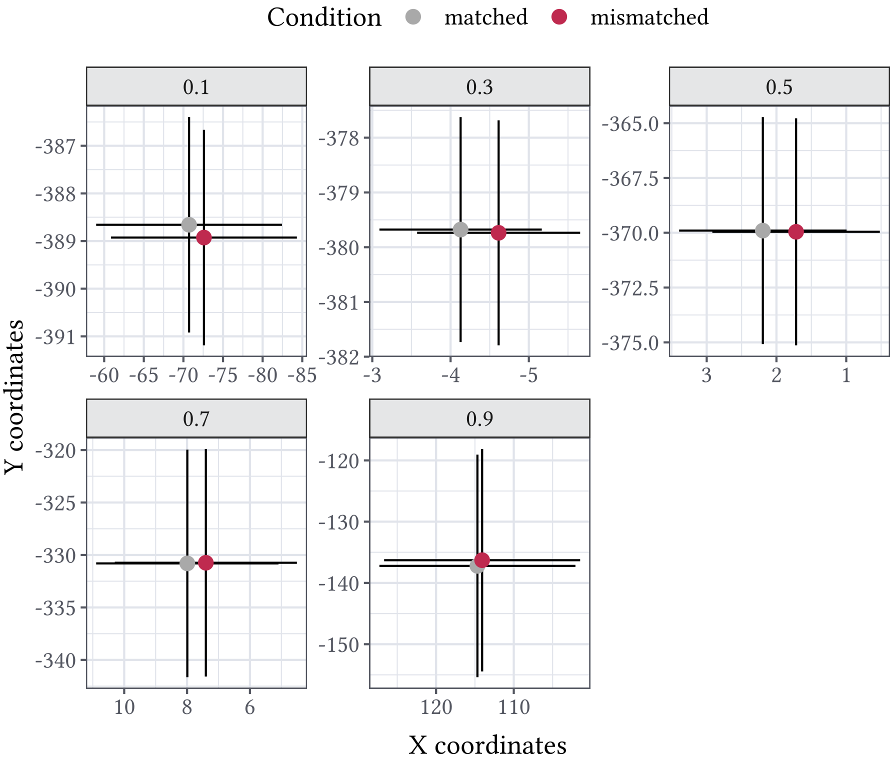


Figure 7.11: Effects of CONDITION as found in the QGAMs modelled to the X and Y coordinates of the plural base data set. The lines indicate the confidence intervals of the estimated X and Y coordinate values.

wards the upper left corner, X coordinates which are further to the right can be understood as deviation from that direct path and as a detour towards the other answer. Taking into account that the X coordinates of mismatched trials are overall significantly further to the right, then, one can conclude that mismatched durational information led to overall higher deviations from the direct path. While this effect on X coordinates was found for both data sets, the effect on Y coordinates was only found for singular pseudo-bases. Analogously, the lower Y coordinates for mismatched trials can be interpreted as a deviation from the direct path. Thus, the results of the mouse-tracking analysis confirm the second part of the hypothesis: Mouse-tracks of trials with mismatched items are deviated.

However, there are two major points that need to be addressed. First, the analysis did not consider whether the word-final /s/ of a particular stimulus has been heard already. While this, in theory, should not pose a problem, because QGAMs were fitted to quantiles across the distribution of coordinates and thus included ranges of coordinates for which the word-final /s/ has been heard, I nonetheless checked whether a difference in results is found. For this, I created two binary variables: s_onset and s_offset. s_onset encodes the time of the onset of the word-final /s/ and s_offset encodes the time of the offset of the word-final /s/. Based on these variables, I created data sets which contained only data for which either the onset of the /s/ was audible or for which the offset of the /s/ was audible already. Using these data sets, I refitted the QGAMs presented in the main analysis of this chapter. The overall results of these new QGAMs are similar to those reported here (see the supplementary material given in Chapter 11). Thus, considering only data for which the onset or offset of the word-final /s/ was audible does not change the general results.

Second, the analysis presented in this chapter excluded time as a relevant factor. Recall that all mouse-tracks were time-normalised during the pre-processing of the data. While this made mouse-tracks more easily comparable for my purposes, time is nonetheless a factor one might consider in other types of analyses. Following, for example, Blazej & Cohen-Goldberg (2015), one can analyse the raw, non-normalised mouse trajectories. For this, I divided all X and Y coordinate data into 200 increments of 10 ms. The average for all of these increments was then calculated. The result of this procedure is illustrated in Figure 7.12. As indicated by the dashed line, the /s/ onset was on average at 389 ms, while the /s/ offset, as indicated by the dotted line, was at 689 ms after the stimulus onset. Comparing Panels A and B, higher deviations between coordinates of matched and mismatched stimuli are again found for X coordinates as compared to Y coordinates. For both types of coordinates, differences between the trajectories become visible between the onset and offset of the word-final /s/. This finding indicates that listeners make use of the durational information. As time unfolds, the duration of the pertinent /s/ either corresponds to its expected duration (match) or it over-/undershoots its expected duration (mismatch). In the latter case, then, this mismatch of expected and perceived duration leads to a deviation of the mouse-track towards the other option. Considering time as a factor thus confirms the main findings of this chapter and provides further insight into the found effects.

Let us now turn to the theoretical implications of the present results. As participants of the present study showed an influence of subphonemic durational differences on their comprehension in terms of mouse-tracks, theories which exclude such information from the result of the perception process cannot account for these findings (e.g. Klatt 1979; McClelland & Elman 1986; Norris 1994; Norris & McQueen 2008). If perception as such is not sensitive to subphonemic

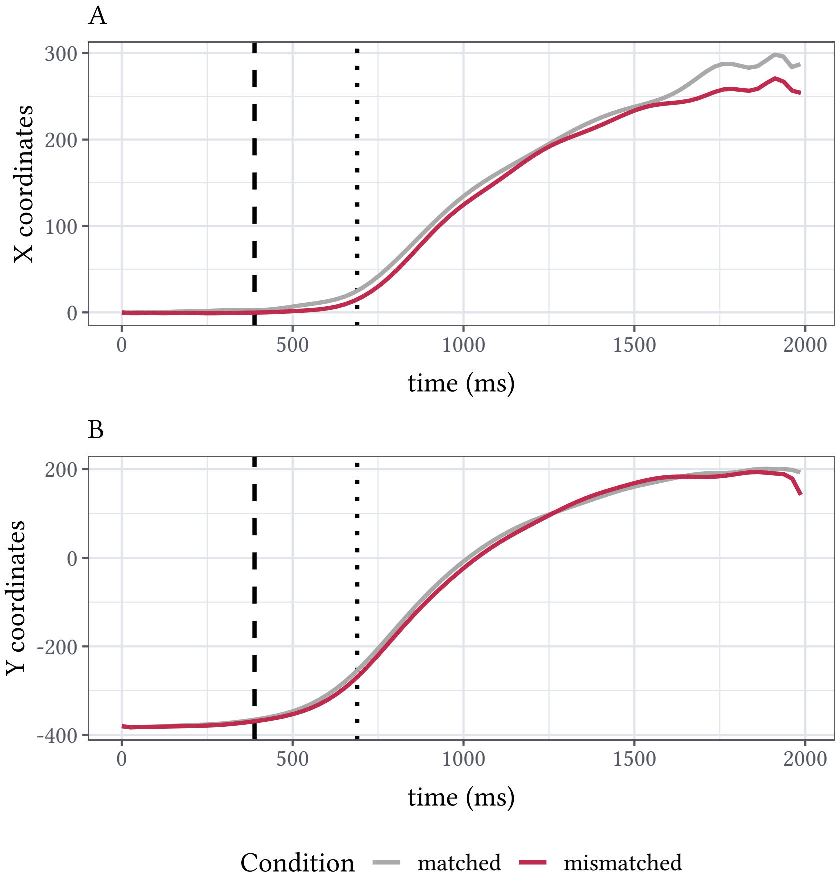


Figure 7.12: Averaged mouse position on the x-axis (Panel A) and on the y-axis (Panel B) for matched and mismatched trials as a function of time. The dashed horizontal lines indicate the average /s/ onset time; the dotted horizontal lines indicate the average /s/ offset time.

durational differences, and as a result, no such information is forwarded to the comprehension process, the comprehension process cannot make use of such durational detail. Thus, no difference between matched and mismatched trials should have been found. However, such theories can account for the null results of the reaction time analysis.

Exemplar and hybrid models (e.g. Goldinger 1996; Hawkins & Smith 2001; Pierrehumbert 2002; Hanique, Aalders, et al. 2013) as well as computational models such as DIANA (ten Bosch et al. 2015; ten Bosch & Boves 2021) and LDL (Baayen, Chuang, Shafaei-Bajestan, et al. 2019) can potentially account for the findings of the mouse-track analysis. As such approaches assume the storage of fine-phonetic detail, such detail can be perceived and made use of in comprehension. However, it remains unclear why an effect of subphonemic durational information is found in mouse-tracks but not in reaction times.

Overall, it seems that no theoretical account can straightforwardly explain the findings of the present number-decision task. While reaction times are not influenced by durational mismatches in word-final /s/, mouse-tracks are. One might argue that reaction times on the one hand and mouse-tracks on the other hand represent different parts of the comprehension process, deliver different amounts of detail on the comprehension process, or show different levels of sensitivity towards mismatched information. Reaction times provide a single data point per trial and allow for little insight into the time window between the start and end of a trial. Even when analysed with novel sophisticated statistical methods such as PAMMs, they provide much less detail on what happens during a particular trial as compared to the continuously measured mouse-tracks. Thus, reaction times between matched and mismatched trials may very well be similar as is the case in the present study, while what happens before a response is recorded is not. These potential differences in the time window between the start of the trial and the response are captured by mouse-tracks. In the present case, this more detailed account of the comprehension process showed a significant influence of mismatched durational information. The present findings, then, can be understood as non-contradictory, as their underlying measures, reaction times and mouse-tracks, capture different aspects of the comprehension process: speed versus decision-making.

However, a detailed account of such potential differences is a subject for future research. Similarly, it has been briefly shown that time should not be disregarded for the analysis of mouse-tracking data. Thus, further analyses considering time as a factor, for example, an analysis of saccades, should be the aim of future research. Finally, one question remains: Are the results presented in this chapter confounded by lexical effects of the real word items used as stimuli? To investigate this question, a second comprehension task in which pseudowords were used is presented in the following chapter.

8 Comprehension of plural and clitic /s/

As explained in detail in Section 2.2, two comprehension studies are part of this book. This chapter presents the second of these studies on the comprehension of subphonemic differences in word-final /s/. The two studies differ in two main regards. First, the comprehension study described in Chapter 7 made use of real words in isolation, the comprehension study presented in this chapter uses pseudowords embedded within sentences as stimuli. Second, the first comprehension study used non-morphemic and plural word-final /s/, while this second study uses plural, *is*-, and *has*-clitic word-final /s/. As in the previous comprehension study, effects on comprehension were tested using a number-decision task in a mouse-tracking paradigm. Considering extant models and approaches of speech perception and comprehension, H comp, the *Mismatch Hypothesis*, again is explored. However, taking into account the findings of the first comprehension study, reaction times are not investigated in the present study. That is, only the second part of the hypothesis is considered: If listeners make use of subphonemic durational differences in the comprehension of different types of word-final /s/, then a mismatch of subphonemic detail and intended meaning is expected to lead to deviated mouse trajectories.

8.1 Methdology

8.1.1 Participants

Forty-two native speakers of New Zealand English took part in the experiment. Their mean age was 22.5 years, ranging from 18 to 54. Eight participants identified as multilingual. The experiment took place at the University of Canterbury, Christchurch, New Zealand, from December 2020 to March 2021.

8.1.2 Materials

The speech materials consisted of pseudowords embedded within sentences. The pseudowords used are those forty-eight described in Section 3.1.2. I repeat all pseudowords in Table 8.1 for convenience.

Table 8.1: Orthographic (*orth.*) and phonological (*phon.*) representations of all pseudowords used in the number-decision task.

	/glɪ/	/prʌ/	/pli:/	/clu:/	/blaʊ/	/gleɪ/
orth.	*glips*	*prups*	*pleeps*	*cloops*	*bloups*	*glaips*
phon.	/glɪps/	/prʌps/	/pli:ps/	/klu:ps/	/blaʊps/	/gleɪps/
orth.	*glits*	*pruts*	*pleets*	*cloots*	*blouts*	*glaits*
phon.	/glɪts/	/prʌts/	/pli:ts/	/klu:ts/	/blaʊts/	/gleɪts/
orth.	*gliks*	*pruks*	*pleeks*	*clooks*	*blouks*	*glaiks*
phon.	/glɪks/	/prʌks/	/pli:ks/	/klu:ks/	/blaʊks/	/gleɪks/
orth.	*glifs*	*prufs*	*pleefs*	*cloofs*	*bloufs*	*glaifs*
phon.	/glɪfs/	/prʌfs/	/pli:fs/	/klu:fs/	/blaʊfs/	/gleɪfs/

All pseudowords were embedded into short context sentences of either simple past, present progressive, or present perfect tense. Additionally, the remaining context disambiguated between plural and non-plural contexts. In sentences with simple past tense, the agents were two aliens of the same kind (see (1) & (2)) doing something together or to each other. This ensured a plural reading of the context. In sentences with present progressive tense, agents were a single alien doing something to or with another alien in object position (see (3) & (4)). In sentences with present perfect tense, agents were single aliens who had done something to or with another alien in object position (see (5) & (6)). That is, for the *is-* and *has*-clitic, the following verb ensured the pertinent clitic reading of the context. Almost exclusively irregular verbs were used to create the context sentences to ensure no ambiguities between them. Twenty-four contexts per type of /s/ were created, resulting in a total number of seventy-two context sentences. See the supplementary material given in Chapter 11 for a list of all verbs and contexts.

(1) The *glips* ate their lunch together.

(2) The *glips* blew a kiss to each other.

(3) The *glip's* eating cake with the bloup.

(4) The *glip's* blowing a kiss to the bloup.

(5) The *glip's* eaten the bloup's lunch.

(6) The *glip's* blown kisses to the bloup every day of their marriage.

The context sentences were made into a reading list, which was then read and recorded three times by a trained native speaker of New Zealand English. Recordings took place at the soundproof booth of the Department of Linguistics at the University of Tübingen. The recordings were sampled at 44.1 kHz, 16 bit.

For each sentence the best of the three recordings was chosen by manual inspection. First, all recordings were analysed using Praat following the segmentation conventions laid out in Section 4.1.4. Recordings with production errors, e.g. laughter, stutter or vocal fry, or segmentation difficulties, e.g. the absence of a stop release, were dismissed. Second, the speaking rate of sentences was measured using a Praat script (de Jong & Wempe 2008) and then analysed in R. As the contexts used in the current experiment differ in length, i.e. in number of syllables, speaking rate appeared to be a more appropriate measurement of similarity across utterances as compared to duration itself as used in the previous two experiments. Speaking rate was computed as number of syllables divided by utterance duration. The resulting mean speaking rate was 3.024 with a standard deviation of 0.551. Lastly, for each sentence, the iteration closest to the mean speaking rate was chosen for further use in the experiment resulting in a final mean speaking rate of 3.021 with a standard deviation of 0.380.

Then, the final /s/ duration of all items was manipulated in such a way that it corresponded to the mean /s/ duration for plural, *is*-, and *has*-clitic /s/ found in the reference study by Plag et al. (2017). That is, in the case of a plural context such as (1) the duration of the final /s/ was changed to 283 ms, while in the case of *is*-clitic contexts such as (3) the duration of the final /s/ was changed to 261 ms, and in the case of *has*-clitic contexts such as (5) the duration of the final /s/ was changed to 253 ms. These versions are manipulated so that their /s/ durations match those of previous findings. Thus, these items are referred to as *matched* items.

For *mismatched* items, /s/ durations were changed as follows. For each plural context, two new versions were created. One contained the typical duration of an *is*-clitic /s/, while the other one contained the typical duration of a *has*-clitic /s/. For each *is*-clitic and *has*-clitic context, a new version was created with the duration of a typical plural /s/. The final number of contexts and their /s/ durations are given in Table 8.2. Each participant took part in 192 trials, i.e. 2 × 24 matched plural /s/ items + 2 × 24 mismatched plural /s/ items + 24 matched *is*-clitic /s/ items + 24 mismatched *is*-clitic /s/ items + 24 matched *has*-clitic /s/ items + 24 mismatched *has*-clitic /s/ items.

8.1.3 Procedure

Similar to the experiment in Chapter 7, the number-decision task was conducted in OpenSesame using the `mousetrap` plugin for mouse-tracking (Kieslich & Hen-

Table 8.2: Number of versions per type of word-final /s/ and their /s/ durations. Mean values are taken from Plag et al. (2017).

	Version 1: matched	Version 2: mismatched	Version 3: mismatched
is-clitic context	mean of *is*-clitic /s/ 261 ms	mean of plural /s/ 283 ms	
has-clitic context	mean of *has*-clitic /s/ 253 ms	mean of plural /s/ 283 ms	
plural context	mean of plural /s/ 283 ms	mean of *is*-clitic /s/ 261 ms	mean of *has*-clitic /s/ 253 ms

ninger 2017). First, participants were introduced to the task at hand. They were told that in the following experiment they had to decide whether a sentence is about the action of two identical aliens, i.e. aliens of the same species and name, or about the action of one alien. They were told to mouse-click on the matching "one" or "two or more" button, respectively, in the top left and top right corner of the screen. Figure 8.1 illustrates what participants saw on screen for each trial. The participants were told that if they did not decide on either option within a certain amount of time, the next trial would start automatically. Each participant started with five practice trials.

Each trial was preceded by a stretch of silence of 450 ms accompanied by a white screen. Then, one of the recordings was played, with reaction time and

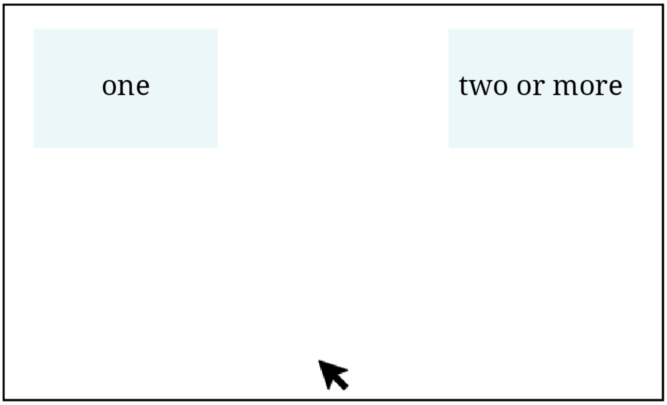


Figure 8.1: Option display during the comprehension experiment. The mouse cursor indicates the position the mouse was reset to in each trial.

mouse-tracking measurement starting at the onset of the recording. Participants were given a window of 4500 ms starting after the onset of the recording to react, after that a time-out was recorded. The next trial then started automatically 5000 ms after the onset of the previous recording, starting with the next inter-trial white screen. Mouse-tracks were recorded with a frequency of 100 Hz.

8.2 Analysis

The analysis of the present data is similar to the analysis of the mouse-tracking data in the comprehension study on non-morphemic and plural word-final /s/ in Chapter 7. First, mouse-tracking data were extracted, spatially transformed, and time-normalised with n = 140 time steps using the `mousetrap` package (Kieslich et al. 2019) in R. Figure 8.2 shows the aggregated mean trajectory of the spatially transformed and time-normalised mouse-tracks in the lower left panel. The panel on top gives the overall distribution of all X coordinates, with a peak visible around a value of 0. The panel on the right displays the overall distribution of all Y coordinates, with a peak around a value of 380. As in Chapter 7, the positions of the peaks corresponds to the position at which the mouse cursor started for each trial.

Then, X and Y coordinates were extracted using the `mtqgam` package (Schmitz 2021b). As in Section 7.2.3, the sign of the coordinate data were reversed to allow for a straightforward interpretation. An example of the resulting data structure is given in Table 8.3.

Finally, the prepared data set was analysed using additive quantile regression models (QGAMs; Fasiolo et al. 2021).

Table 8.3: Example of the data format for a matched trial.

order	trialNumber	time	x_coordinate	y_coordinate	condition
1	1	0	0	-380.000	matched
…	…	…	…	…	…
9	1	201.9568	2.000	-380.000	matched
10	1	227.2014	3.440	-379.560	matched
11	1	252.4460	6.734	-375.266	matched
…	…	…	…	…	…
110	1	2751.6619	72.487	90.160	matched
111	1	2776.9065	139.647	120.597	matched
…	…	…	…	…	…
130	1	3256.5540	274.000	199.000	matched
131	1	3281.7986	274.000	199.000	matched

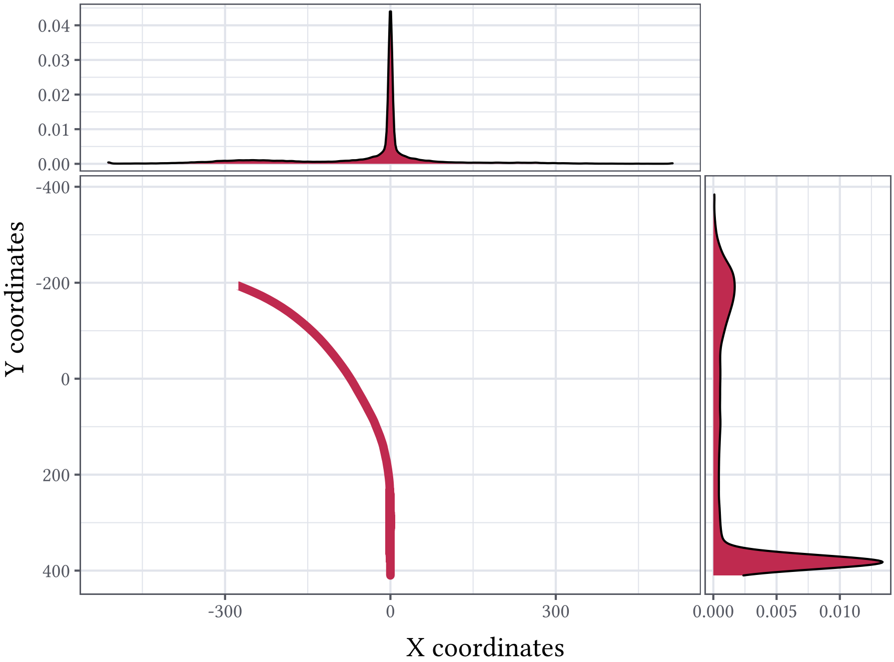


Figure 8.2: Mean trajectory of all spatially adjusted and time-normalised mouse-tracks (lower left), and density distribution of X and Y coordinates (on top and on the right, respectively).

8.2.1 Fitted models

The complete set of coordinate data (n = 1,017,800) was split into four separate data sets. Recall that there were three types of word-final /s/ involved in this study, i.e. plural, *is*-, and *has*-clitic /s/. Targets in plural context sentences were once manipulated to bear the mismatched /s/ duration of an *is*-clitic, and once to bear the mismatched /s/ duration of a *has*-clitic. An overview of the four subsets and the contexts they contain is given in Table 8.4.

SUBSET$_{\text{IP}}$ thus contained results on *is*-clitic contexts with either *is*-clitic /s/ or plural /s/ durations (n = 260,400). SUBSET$_{\text{HP}}$ contained results on *has*-clitic contexts with either *has*-clitic /s/ or plural /s/ durations (n = 229,600). SUBSET$_{\text{PI}}$ contained results on plural contexts with either plural /s/ or *is*-clitic /s/ durations (n = 263,900). SUBSET$_{\text{PH}}$, finally, contained results on plural contexts with either plural /s/ or *has*-clitic /s/ durations (n = 263,900). Similar to the analysis of mouse-tracks in the comprehension study on non-morphemic and plural /s/ presented in Section 7.2.3, the individual subsets were created in order to determine

Table 8.4: Overview of the four subsets used in the QGAM modelling process. Each subset contains mouse-track coordinate data of durationally matched and mismatched stimuli. Within each subset, the type of context is kept constant, while the manipulation of the pertinent word-final /s/ either corresponds to a match or mismatch in duration. Subset names contain information on the type of context (first subscript letter) and the /s/ duration that constitutes a mismatch (second subscript letter).

Subset name	Condition	Context	/s/ duration
SUBSET$_{\mathrm{IP}}$	matched mismatched	*is*-clitic *is*-clitic	*is*-clitic plural
SUBSET$_{\mathrm{HP}}$	matched mismatched	*has*-clitic *has*-clitic	*has*-clitic plural
SUBSET$_{\mathrm{PI}}$	matched mismatched	plural plural	plural *is*-clitic
SUBSET$_{\mathrm{PH}}$	matched mismatched	plural plural	plural *has*-clitic

whether a mismatch of context and word-final /s/ influenced mouse-tracks to a significant extent. While this is also possible with the specification of interaction terms in the QGAM formula, it was again decided against this method due to the high computational costs as well as due to the increased complexity of model interpretation.

Two sets of QGAMs were fitted to each of the four subsets. One set of QGAMs was fitted to X coordinates, one set of QGAMs was fitted to Y coordinates. I aimed at estimating the conditional quantiles corresponding to $\tau = 0.1, 0.3, 0.5, 0.7$ and 0.9. Thus, each set of QGAMs consisted of five individual QGAMs, one for each of the five quantiles. In total, ten QGAMs for each of the four subsets were fitted, that is five for X coordinates and five for Y coordinates. This resulted in a total number of forty QGAMs.

Taking into account the low number of incorrectly answered trials in the data of Chapter 7, I checked the amount of data points for which the wrong answer was given in the present data. Again, only few data points for wrong answers, i.e. about 9% (n = 89,740), were found. It was decided to exclude CORRECT as a variable for the QGAM model formula, and to only use data on correctly answered trials instead. This led to slightly smaller data sets, i.e. n = 243,600 for SUBSET$_{\mathrm{IP}}$; n = 193,340 for SUBSET$_{\mathrm{HP}}$; n = 246,820 for SUBSET$_{\mathrm{PI}}$; and n = 244,300 for SUBSET$_{\mathrm{PH}}$. An overview of all variables contained in the four subsets is given in Table 8.5. See Section 7.2.1 for the definitions of all covariates.

Table 8.5: Summary of the dependent variables and the numerical and categorical predictors in the four subsets.

Subset(s)	Mean	St. Dev.	Min	Max
subset$_{IP}$	50.326	135.957	-511.000	512.000
subset$_{HP}$	48.251	142.696	-511.000	512.000
subset$_{PI}$	32.086	142.656	-512.000	511.000
subset$_{PH}$	32.507	141.780	-512.000	511.000
subset$_{IP}$	-173.983	250.326	-410.000	384.000
subset$_{HP}$	-160.199	253.547	-410.000	384.000
subset$_{PI}$	-181.272	245.044	-410.000	384.000
subset$_{PH}$	-177.168	248.281	-410.000	384.000
Subset(s)	Mean	St. Dev.	Min	Max
all	70.500	40.414	1.000	140.000
Subset(s)	Levels			
all	24			
all	42			
Subset(s)	Levels			
subset$_{IP}$	`matched`: 121940		`mismatched`: 121660	
subset$_{HP}$	`matched`: 97020		`mismatched`: 96320	
subset$_{PI}$	`matched`: 123760		`mismatched`: 123060	
subset$_{PH}$	`matched`: 121800		`mismatched`: 122500	

All QGAMs used the same model formula. The dependent variable was either x_coordinate or y_coordinate. condition was introduced as parametric term. order was given as smooth term with the default *k*-value of 9, and item and subject were included as random smooth terms. Checks revealed that the *k*-value of the order smooth term was too low. However, as in Section 7.2.2.2, it was found that no matter the *k*-value, the effect of all covariates remained unchanged. Following Wood (2017), it was therefore, again, decided to not re-fit the computationally costly QGAMs. The final data set as well as the analysis and results discussed in the following sections can be found in the supplementary material given in Chapter 11. In the following, the results of the modelling process will be presented.

8.2.2 Results

A significant effect of condition was found 24 times across all forty models. The smooth term of order as well as the random smooth terms of item and subject reached significance in all models. The overall model fit is rather high with a mean deviance explained of $\overline{D} = 58.75\%$. Across all four data sets, QGAMs fitted to Y coordinates showed overall higher rates of deviance explained ($\overline{D} = 68.29\%$) than their X coordinate counterparts ($\overline{D} = 49.21\%$). For all four sets of QGAMs, the QGAM fitted to the 0.5 quantile showed the lowest rate of deviance explained ($\overline{D}_{0.5} = 44.69\%$), while the QGAMs fitted to the more extreme quantiles showed the highest rates of deviance explained ($\overline{D}_{0.1} = 75.41\%$ and $\overline{D}_{0.9} = 72.06\%$).

8.2.2.1 SUBSET$_{\text{IP}}$

The effects found in the QGAMs fitted to the X and Y coordinates of SUBSET$_{\text{IP}}$ are given in Table 8.6. The model estimates of these and all subsequent QGAMs are part of the supplementary material given in Chapter 11. Note that here and in the following, I will refrain from discussing the effects of the smooth terms as they are not the main interest of investigation.

There are significant effects of CONDITION in two QGAMs fitted to X coordinate data and in two QGAMs fitted to Y coordinate data. For both types of coordinates, these effects are found for $\tau = 0.7$ and $\tau = 0.9$. The effects are illustrated by Figure 8.3. Where a significant effect is found for X coordinates, tracks of mismatched trials are further to the left as compared to tracks of matched trials. For Y coordinates, the effect of condition leads to coordinates further up for mismatched trials. Recall that due to the use of conditional quantiles in QGAMs, the estimates shown in Figure 8.3 and similar plots illustrate the nature of an effect taking into account a certain quantity of the overall data. Such plots do not illustrate the positions of mouse-tracks at certain points of their trajectory.

8.2.2.2 SUBSET$_{\text{HP}}$

For SUBSET$_{\text{HP}}$, the found effects are given in Table 8.7. The effect of CONDITION reaches significance in three QGAMs fitted to X coordinates, i.e. in the $\tau = 0.1, 0.7$ and 0.9 quantiles. For Y coordinates, significant effects are found in all QGAMs but the QGAM fitted to the $\tau = 0.9$ quantile. The effects are illustrated in Figure 8.4. For X coordinates in $\tau = 0.1$, the effect of CONDITION leads to coordinates further to the right. Taking into account more data, the effect is reversed in $\tau =$ 0.7 and 0.9, i.e. coordinates of mismatched trials are further left. The effect of CONDITION on Y values is similar across all quantiles. That is, mismatched trials show further up Y coordinates.

Table 8.6: Summary of the effects found in the QGAMs fitted to the X and Y coordinates of SUBSET$_{\text{IP}}$. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	X coordinates					Y coordinates				
quantiles:	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Parametric Terms										
(Intercept)	***	***	***	***	***	***	***	***	***	n.s.
CONDITION`mismatched`	n.s.	n.s.	n.s.	***	***	n.s.	n.s.	n.s.	**	***
Smooth Terms										
ORDER	***	***	***	***	***	***	***	***	***	***
Random Smooth Terms										
ITEM	***	***	***	***	***	***	***	***	***	***
SUBJECT	***	***	***	***	***	***	***	***	***	***

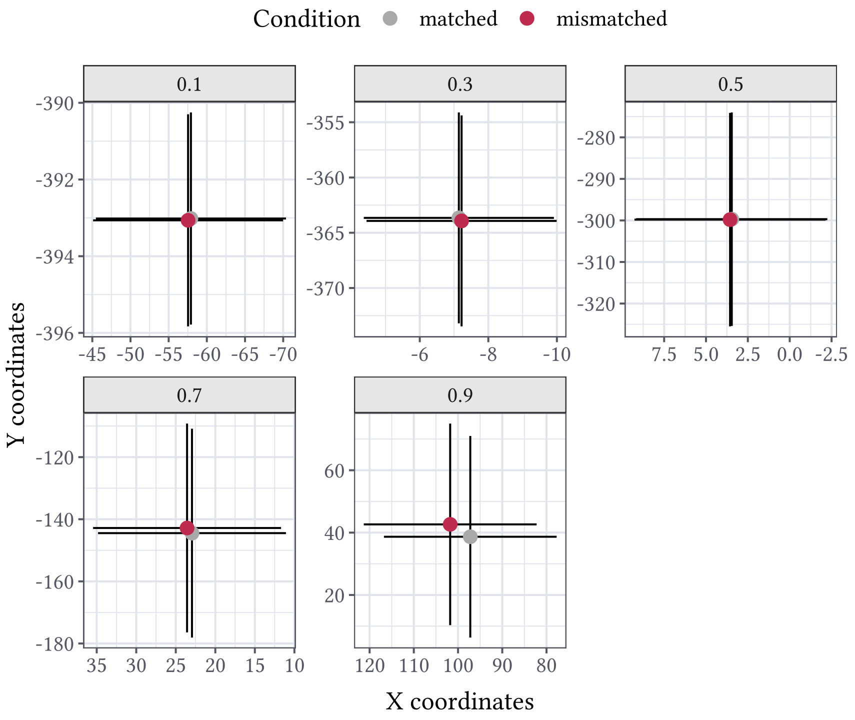


Figure 8.3: Effect of CONDITION as found in the QGAMs modelled to the X and Y coordinates of SUBSET$_{\text{IP}}$. The lines indicate the confidence intervals of the estimated X and Y coordinate values.

Table 8.7: Summary of the effects found in the QGAMs fitted to the X and Y coordinates of SUBSET$_{HP}$. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	X coordinates					Y coordinates				
quantiles:	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Parametric Terms										
(Intercept)	***	***	***	***	***	***	***	***	***	**
CONDITION`mismatched`	***	n.s.	n.s.	**	***	*	***	***	***	n.s.
Smooth Terms										
ORDER	***	***	***	***	***	***	***	***	***	***
Random Smooth Terms										
ITEM	***	***	***	***	***	***	***	***	***	***
SUBJECT	***	***	***	***	***	***	***	***	***	***

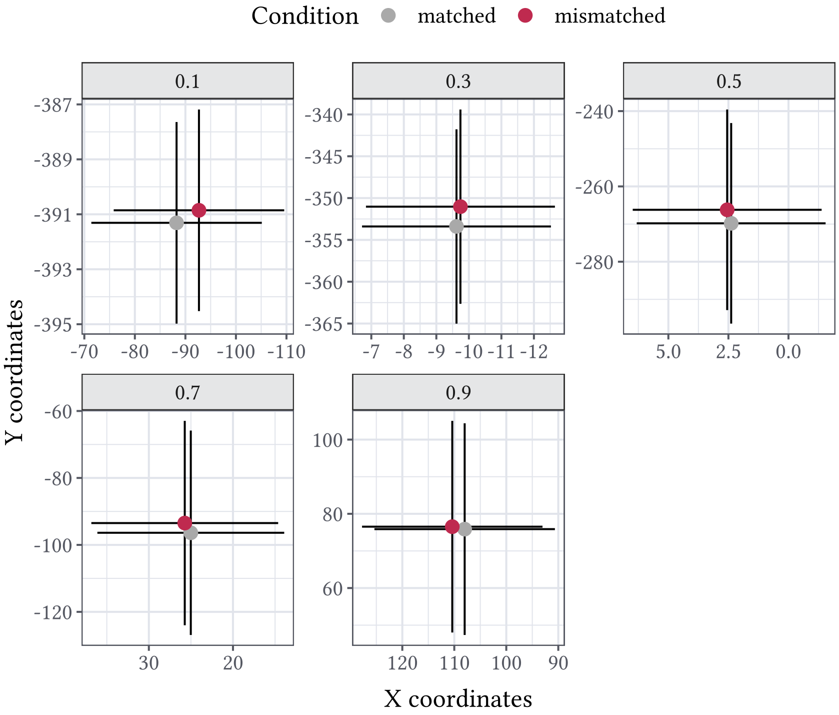


Figure 8.4: Effect of CONDITION as found in the QGAMs modelled to the X and Y coordinates of SUBSET$_{HP}$. The lines indicate the confidence intervals of the estimated X and Y coordinate values.

8.2.2.3 SUBSET$_{\mathrm{PI}}$

Table 8.8 presents the effects found for QGAMs fitted to the SUBSET$_{\mathrm{PI}}$ data. A significant effect of CONDITION is found in quantiles $\tau = 0.1, 0.3$ and 0.5 for X coordinates, and in all quantiles but $\tau = 0.1$ for Y coordinates. The effects are displayed in Figure 8.5. Where the effect of CONDITION is significant for X values, coordinates of mismatched trials are further right. For Y coordinates, the effect of CONDITION comes with coordinates further down for mismatched trial coordinates.

8.2.2.4 SUBSET$_{\mathrm{PH}}$

Finally, the effects found in the QGAMs fitted to the X and Y coordinates of the SUBSET$_{\mathrm{PH}}$ data are given in Table 8.9. CONDITION shows a significant effect on X coordinates in quantiles $\tau = 0.5, 0.7$ and 0.9. For Y coordinates, a significant effect is found in all quantiles but $\tau = 0.9$. The effects are illustrated in Figure 8.5. For X coordinates, the effect of CONDITION comes with coordinates further left for mismatched trials. For Y coordinates, the effect of CONDITION leads to coordinates lower down for mismatched trials.

8.2.2.5 Overall results

An overview of significant deviations found for the coordinates of mismatched stimuli trials across all quantiles and subsets is given in Table 8.10. Considering the overall influence of CONDITION, one can make two general observations. First, Y coordinates of mismatched stimuli trials are higher and, with the exception of $\tau = 0.1$ for SUBSET$_{\mathrm{HP}}$, further to the left if a mismatch is caused by a plural /s/ duration, as is the case in SUBSET$_{\mathrm{IP}}$ and SUBSET$_{\mathrm{HP}}$. Second, Y coordinates of mismatched stimuli trials are lower if a mismatch is caused by an *is*- or *has*-clitic /s/ duration, as is the case in SUBSET$_{\mathrm{PI}}$ and SUBSET$_{\mathrm{PH}}$, while for X coordinates no clear pattern is visible.

8.3 Discussion

The number-decision task presented in this chapter investigated whether listeners make use of subphonemic durational differences in comprehension. It is different to the comprehension study presented in Chapter 7 by several aspects. First, pseudowords instead of real words were used as items. Thus, the effects found in the present study cannot be confounded by effects of lexical storage,

Table 8.8: Summary of the effects found in the QGAMs fitted to the X and Y coordinates of SUBSET$_{\text{PI}}$. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	X coordinates					Y coordinates				
quantiles:	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Parametric Terms										
(Intercept)	***	n.s.	***	***	***	***	***	***	***	n.s.
CONDITIONmismatched	n.s.	n.s.	n.s.	***	***	***	***	***	***	n.s.
Smooth Terms										
ORDER	***	***	***	***	***	***	***	***	***	***
Random Smooth Terms										
ITEM	***	***	***	***	***	***	***	***	***	***
SUBJECT	***	***	***	***	***	***	***	***	***	***

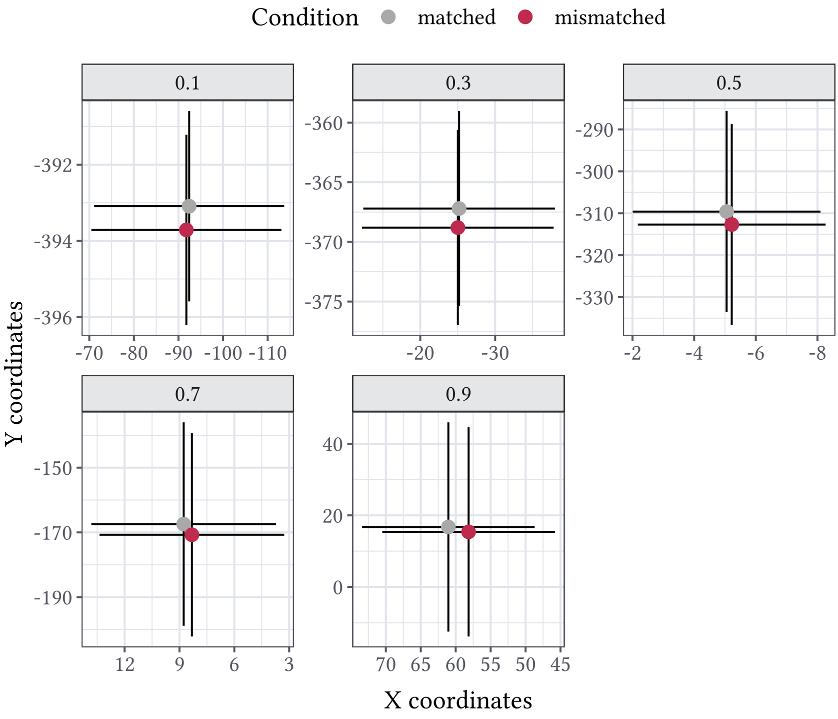


Figure 8.5: Effect of CONDITION as found in the QGAMs modelled to the X and Y coordinates of SUBSET$_{\text{PI}}$. The lines indicate the confidence intervals of the estimated X and Y coordinate values.

Table 8.9: Summary of the effects found in the QGAMs fitted to the X and Y coordinates of subset$_{\text{PH}}$. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	X coordinates					Y coordinates				
quantiles:	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Parametric Terms										
(Intercept)	***	n.s.	***	***	***	***	***	***	***	n.s.
conditionmismatched	**	***	***	n.s.	n.s.	n.s.	*	***	***	***
Smooth Terms										
order	***	***	***	***	***	***	***	***	***	***
Random Smooth Terms										
item	***	***	***	***	***	***	***	***	***	***
subject	***	***	***	***	***	***	***	***	***	***

Table 8.10: Overview of the direction of significant deviations found for coordinates in mismatched stimuli trials across all quantiles and subsets. Where no direction is given, no significant effect of condition was found.

	subset$_{\text{IP}}$		subset$_{\text{HP}}$		subset$_{\text{PI}}$		subset$_{\text{PH}}$	
τ	X	Y	X	Y	X	Y	X	Y
0.1			right	higher		lower	left	
0.3				higher		lower	left	lower
0.5				higher		lower	left	lower
0.7	left	higher	left	higher	right	lower		lower
0.9	left	higher	left		right			lower

frequency, or relatedness which are commonly associated with real words and their representations in the mental lexicon (see Section 3.1.1). Second, items were presented in carrier sentences and not in isolation. This was necessary to disambiguate between different types of /s/ in the long run, as the number-decision process for pseudowords cannot rely on lexical knowledge. Third, plural, *is*-, and *has*-clitic word-final /s/ were part of the items, while the previous comprehension study investigated non-morphemic and plural word-final /s/. By investigating different types of /s/ across studies, one obtains a more detailed picture of potential effects. Despite these differences, both comprehension studies shared the same hypothesis. Building on extant models of speech perception and comprehension, H comp, the *Mismatch Hypothesis*, was explored: If subphonemic

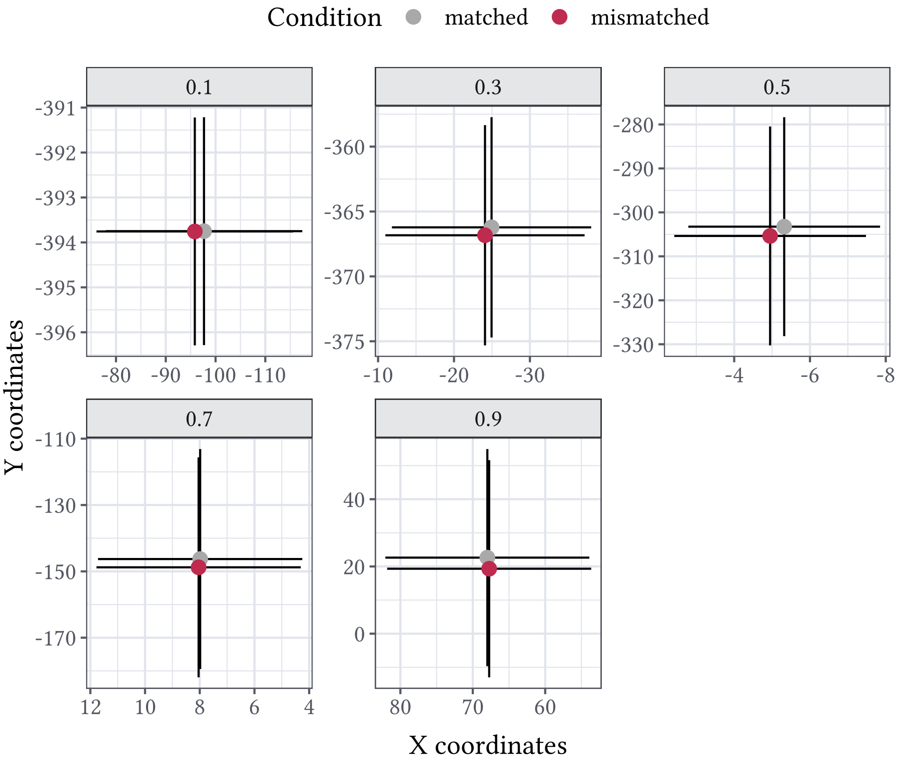


Figure 8.6: Effect of CONDITION as found in the QGAMs modelled to the X and Y coordinates of SUBSET$_{PH}$. The lines indicate the confidence intervals of the estimated X and Y coordinate values.

durational differences are made use of, then a mismatch of subphonemic detail and intended meaning leads to a) slowed down comprehension processes, and b) deviated mouse trajectories. Part a) of the hypothesis was not investigated in the present study, as null results were found in the comprehension study of Chapter 7. Thus, one question remained: Did a mismatch of subphonemic durational information lead to deviated mouse trajectories?

This question was investigated using QGAMs fitted to the X and Y coordinate data of the mouse-tracks recorded in the number-decision task. QGAMs were fitted separately for four subsets of data: 1) SUBSET$_{IP}$, 2) SUBSET$_{HP}$, 3) SUBSET$_{PI}$, and 4) SUBSET$_{PH}$, where the first subscript letter indicates the context and the second subscript letter indicates the mismatched type of /s/ (I = *is*-clitic; H = *has*-clitic; P = plural). In each of the subsets, trials of items with matched and mismatched /s/

duration were compared, while the items' bases as well as the sentence the items were embedded within were kept constant. The overall results of the QGAMs show an effect of matched versus mismatched durational information across all four subsets. Taking a closer look at the nature of the found effects, one finds effects going into different directions between subsets (see Table 8.10). For X coordinates, coordinate values of mismatched stimuli trials are further to the left for SUBSET$_{\mathrm{IP}}$, SUBSET$_{\mathrm{HP}}$, and SUBSET$_{\mathrm{PH}}$. For SUBSET$_{\mathrm{PI}}$, however, X coordinate values of mismatched trials are further to the right. For Y coordinates, one finds a difference between both clitic contexts and both plural contexts: Y coordinates are higher up when a mismatch is caused by a plural /s/ duration, but they are further down when the mismatch is caused by a clitic /s/ duration. Similar effects were found for QGAMs fitted post-hoc to the data for which the onset of the word-final /s/ has been heard already (see Section 7.3 for a discussion and the supplementary material given in Chapter 11 for model overviews).

What do these findings mean in regard to the notion of deviated mouse-tracks due to mismatched subphonemic durational information? Recall that the mouse-tracks were mirrored where applicable, that is all tracks move towards the upper left corner of the coordinate system. Thus, an ideal non-deviated trajectory would be a straight line between the mouse cursor starting position and one of the answer options. As this non-deviated straight line moves linearly towards the upper left corner, X coordinates which are further to the left or right and Y coordinates which are further up or down can be understood as deviation from that direct path. If mismatched subphonemic durational information was to cause deviation from that ideal path, one would expect X coordinates to be further to the right and Y coordinates to be lower down, as such a deviation would express the expected effect of a mismatch: As context and /s/ duration do not match up, comprehension is influenced, and the mouse-track is deviated towards the incorrect response for the pertinent trial. Taking a trial with a clitic context as an example, a mismatch is caused by the plural /s/ duration of the target word's word-final /s/. If comprehension is influenced by this durational mismatch, one would predict mouse movement towards the plural response due to the word-final /s/ duration. Once the entire context is processed, a correct answer is given. Moving the mouse away from the incorrect towards the correct response then results in an overall more deviated mouse-track.

How do the present findings relate to this prediction of a deviated path? For X coordinates, a deviation to the right was found for $\tau = 0.9$ of SUBSET$_{\mathrm{HP}}$ and across SUBSET$_{\mathrm{PI}}$. In all other significant cases, mismatched trials showed X coordinates further to the left instead. For Y coordinates, the expected lower coordinate values were found for SUBSET$_{\mathrm{PI}}$ and SUBSET$_{\mathrm{PH}}$, while SUBSET$_{\mathrm{IP}}$ and SUBSET$_{\mathrm{HP}}$

showed higher Y coordinate values instead. That is, only the results for subset$_{\text{PI}}$ fully meet the expected directions of deviations. subset$_{\text{PH}}$ meets the directions for Y coordinates, but not for X coordinates. The two subsets in which a mismatch is caused by a plural /s/ duration, subset$_{\text{IP}}$ and subset$_{\text{HP}}$, show deviations of the opposite directions instead: X coordinates are mainly further to the left and Y coordinates are higher up. Nonetheless, the *Mismatch Hypothesis* is confirmed by the overall findings: Comparing mouse-tracks of matched and mismatched stimuli trials, one finds that they significantly deviate from each other across more than half of all QGAMs.

However, how can one explain the opposing findings between subset$_{\text{IP}}$ and subset$_{\text{HP}}$ on the one, and subset$_{\text{PI}}$ and subset$_{\text{HP}}$ on the other hand? Noticeably, the effects found within the two clitic context subsets, as well as the effects found within the two plural context subsets are mostly similar. General differences in effect directions for Y coordinates are only found between these two groups. One potential explanation that comes to mind is the overall frequency of plural and clitic /s/ in the language. In the British National Corpus (Davies 2004), *is*-clitic <'s> is attested 311,146 times and *has*-clitic <'s> is attested 22,816 times. For plural <s>, the most frequent entry alone, *things*, has a frequency of 40,453 which is almost double the frequency of the *has*-clitic. Considering just plural /s/, one finds a frequency of about 140,000 when taking into account the top ten most frequent /s/ plural forms alone. That is, plural /s/ is overall far more frequent than clitic /s/. If a pseudoword contains the duration of a plural /s/, mouse-tracks deviate differently than predicted, i.e. further to the left and further up, as plural /s/ duration is the expected duration. The plural /s/ duration is expected as it is more frequent across the language. If a pseudoword contains the duration of a clitic /s/, mouse-tracks are deviated as predicted, i.e. further down, as this is a less expected duration due to the relatively low frequency of the clitic /s/ duration. Note that this is but an idea which requires further investigation.

Overall, the present results confirm that comprehension is significantly influenced by a mismatch of subphonemic durational information in word-final /s/. This finding is in line with the results of the mouse-track analysis presented for the comprehension study in Chapter 7 of this book. The nature of the found deviations, however, remains unaccounted for and requires further research.

Let us now turn to the theoretical implications of the present findings. Abstractionist theories which exclude subphonemic durational information from the perception and comprehension process cannot account for the present findings (e.g. Klatt 1979; McClelland & Elman 1986; Norris 1994; Norris & McQueen 2008). If such durational differences are not perceived, they cannot be used in comprehension. As significant differences between mouse-tracks of trials with matched versus mismatched durational information were found, such abstractionist approaches cannot explain the present results.

Exemplar and hybrid models (e.g. Goldinger 1996; Hawkins & Smith 2001; Pierrehumbert 2002; Hanique, Aalders, et al. 2013) and computational models (DIANA, ten Bosch et al. 2015; ten Bosch & Boves 2021; LDL, Baayen, Chuang, Shafaei-Bajestan, et al. 2019) could in principle account for the present results. These approaches assume the storage of subphonemic detail. Such detail can be perceived and made use of in comprehension. However, it remains unclear how exemplar and hybrid models would account for the reverse effects found for clitic versus plural contexts. Computational models, however, might be able to shed further light on this issue. Taking LDL as a starting point, one could use the phonological and semantic measures derived from an implementation such as the one given in Chapter 5 as predictors to model coordinate data. Considering that one of these measures apparently reflects the distinction between non-morphemic and plural /s/ (see Section 5.4), it might very well be the case that another measure can capture the effect of durational matches and mismatches. However, for such an implementation additional steps are required. First, audio data instead of phonological triphones has to be used as input to provide information on subphonemic durational differences. Second, one has to find a way to include clitic /s/, because clitic /s/ has not been incorporated in LDL implementations yet.

In sum, no theoretical account can straightforwardly explain the findings of the present number-decision task. Mouse-tracks are influenced by mismatched subphonemic durational information in pseudowords. However, the nature of this influence is unaccounted for: Opposing directions of effects are found when comparing mismatched trials of clitic contexts and plural contexts. An explanation for these reversed effects should be motivation for future research. Such research might benefit from new LDL implementations and derived measures. Overall, the present study showed that subphonemic durational information is used in comprehension, and that such results are found independently of effects of lexical storage, frequency, and relatedness.

9 General discussion

In this book, I set out to establish substantiated knowledge on subphonemic detail and its role in production, perception, and comprehension. To achieve this goal, I used real words and pseudowords as items where applicable and I conducted thorough statistical analyses using novel statistical techniques where appropriate. To investigate the production, perception, and comprehension of subphonemic detail I made use of word-final /s/ in English as it is not only found as non-morphemic segment, but also has numerous morphological functions: plural, genitive, genitive plural, third-person singular, as well as the clitics of *is*, *has*, and *us* (as in *let's*). Using a subset of these different types of /s/ – non-morphemic, plural, *is*-, and *has*-clitic /s/ – I conducted five studies. The aims of these studies were to determine whether such different types of /s/ show differences in their acoustic duration in production (Chapter 4), to gain further insight into how such durational differences come to be (Chapter 5), to learn whether durational differences in word-final /s/ are perceptible (Chapter 6), and to examine if durational differences in word-final /s/ are made use of in comprehension (Chapters 7 and 8). All investigations were of an explorative nature, addressing hypotheses derived from relevant theories to provide elaborate discussions of the pertinent findings. In the following, the respective hypotheses are repeated and then discussed based on the findings of the individual studies. Finally, all results are brought together to draw an overall picture of the production, perception, and comprehension of subphonemic detail in word-final /s/.

The production study presented in Chapter 4 of this book investigated whether there are durational differences in the acoustics of non-morphemic, plural, *is*-, and *has*-clitic word-final /s/. Using pseudowords as items in a highly controlled production task, it was made sure that effects of lexical frequency, predictability, and storage did not confound the results. It was shown that non-morphemic /s/ was longest, plural /s/ was shorter, and clitic /s/ was shortest. While these differences were found to be significant, the difference between the *is*- and the *has*-clitic was not. The following hypotheses were investigated:

H PROD$_1$: *Feed-Forward Hypothesis*

There is no durational difference between word-final non-morphemic /s/, plural /s/, and auxiliary clitic /s/.

H PROD$_2$: *Prosodic Hypothesis*
There are durational differences between different types of word-final /s/: non-morphemic /s/ is shorter than plural /s/, plural /s/ is shorter than auxiliary clitic /s/.

H PROD$_3$: *Emergence Hypothesis*
There are durational differences between different types of word-final /s/ (non-morphemic, plural, and auxiliary clitic).

H PROD$_1$, the *Feed-Forward Hypothesis*, is rejected as it predicted no durational differences between different types of word-final /s/. If standard feed-forward models of speech production underlying this hypothesis were refined in such a way that post-lexical processes can arise from certain kinds of lexical information, only then the present findings could be accounted for. H PROD$_2$, the *Prosodic Hypothesis*, is rejected as it predicted the opposite direction for durational differences, with non-morphemic /s/ being shortest and clitic /s/ being longest in duration. This pattern is clearly not compatible with the present results. The theories underlying H PROD$_3$, the *Emergence Hypothesis*, can potentially account for the present findings. The fact that durational differences were found indicates that such differences might emerge through the mechanisms introduced by the theories underlying this hypothesis. However, claiming that the hypothesis is therefore confirmed would be a fallacy: Only an implementation of one of such underlying theories can show whether the particular theory and its mechanisms can account for the durational differences found in the present production study.

Hence, an implementation of one of the underlying theories, linear discriminative learning, was used to further investigate the hypothesis. This LDL implementation and its analysis were presented in Chapter 5 of this book. Using the non-morphemic and plural /s/ durational data elicited in the production study, the analyses of the LDL implementation resulted in three main findings. First, measures derived from an LDL network trained on real words and pseudowords are predictive of word-final /s/ duration in pseudowords. Such measures are indeed just as predictive of /s/ durations as are more traditional variables. Second, even though such LDL measures show about the same level of predictivity, the effect of the type of /s/ as a variable is not fully captured by them. That is, the type of the word-final /s/ remained a significant predictor when introduced among measures derived from the LDL network. This indicates that there is more to the type of /s/ than the variables used in the present implementation. Third, even though the type of /s/ is not fully captured by the LDL measures, especially one of these measures, the correlation with the semantic nearest neighbour, showed a high correlation with the type of /s/. Hence, intricate semantic properties of the types of /s/ under investigation are indeed captured by measures derived from the LDL network. Coming back to the hypothesis at hand, H PROD$_3$, the *Emergence Hypothesis*, it is found that it can be confirmed in regard to one of its underlying theories, linear discriminative learning.

Taking the results on the production of word-final /s/ as a starting point, the perception study presented in Chapter 6 asked whether such durational differences are perceptible. For this, a same-different task with real words and pseudowords as items was conducted. For each item, a version with the pertinent prototypical duration of non-morphemic or plural /s/ was created. Then, four further versions were constructed with their word-final /s/ either being incrementally shortened (mono-morphemic items) or lengthened (plural items) by 10 ms, 20 ms, 35 ms, and 75 ms. The results indicate that, on average, listener sensitivity is rather low for durational differences of 10 ms and 20 ms, and slightly but significantly higher for a durational difference of 35 ms. For the 75 ms durational difference, a significantly improved sensitivity was found. The following hypotheses were investigated:

H PERC$_1$: *Abstractionist Hypothesis*
Listeners are not sensitive to subphonemic durational differences between different types of word-final /s/.

H PERC$_2$: *Phonetic Detail Hypothesis*
Listeners are sensitive to subphonemic durational differences between different types of word-final /s/.

H PERC$_1$, the *Abstractionist Hypothesis*, is rejected. The hypothesis was built on theories which assume that subphonemic durational differences are not perceptible. Due to the strictly phonological nature of perception found in such theories, these and the present findings are fully incompatible. H PERC$_2$, the *Phonetic Detail Hypothesis*, can be confirmed under two premises. First, only an implementation of the models underlying the hypothesis can sufficiently confirm whether a particular model's mechanisms can account for the present findings. Second, listeners showed sensitivity to subphonemic durational differences. However, major increases in sensitivity and overall high levels of sensitivity were only found for the biggest durational difference of 75 ms – a difference that is not found in studies on the durational differences between different types of /s/. Thus, according to the present findings, not all durational differences between different types of /s/ found in studies on their acoustic duration are assumed to be well perceptible.

Importantly, there most likely is an issue of methodology at hand here. Same-different tasks such as the one used in the present perception study are metalinguistic tasks. Hence, certain properties of language are the main focus for participants of such tasks instead of language or language use itself. Thus, participants encountered a task they are not familiar with and that extends beyond their day-to-day usage of language: differentiating isolated words by the duration of their word-final /s/. It might thus very well be the case that a same-different task is not the most appropriate experimental setup to investigate the perceptibility of subphonemic durational differences.

A type of task that focuses more narrowly on language use itself was used in the two comprehension tasks presented in Chapters 7 and 8. In number-decision tasks, participants were asked to decide whether an isolated word (Chapter 7) or the agent in a sentence (Chapter 8) was singular or plural. In the case of isolated words, words with non-morphemic and plural /s/ were used as target items. In the case of agents in a sentence, pseudowords with plural, *is*-, and *has*-clitic /s/ were used as target items. In both experiments, /s/ durations were either matched with their context, e.g. a plural word had an /s/ with a typical plural /s/ duration, or /s/ durations were mismatched with their context, e.g. a plural word had an /s/ with a typical non-morphemic or clitic /s/ duration. It was found that reaction time was not influenced by the durational mismatch of word-final /s/ and (pseudo-)base. Mouse-tracks, however, showed a significant effect of mismatched durations in that they followed significantly different paths as compared to the mouse-tracks of matched items. In both comprehension studies, the following hypothesis was investigated:

H COMP: *Mismatch Hypothesis*

> If listeners make use of subphonemic durational differences in the comprehension of different types of word-final /s/, then a mismatch of subphonemic detail and intended meaning leads to
> a) slowed down comprehension processes.
> b) deviated mouse trajectories.

As no differences in reaction times were found in Chapter 7, part a) of H COMP cannot be confirmed. That is, the overall time to react to an audio stimulus with a mismatched /s/ duration is just as long as the time to react to an audio stimulus with a matched /s/ duration. Part b) of H COMP, however, is confirmed by the findings, as mouse-tracks of both conditions, matched and mismatched, significantly differed, i.e. the mouse-tracks of the mismatched stimuli trials deviated from the mouse-tracks of the matched stimuli trials. While the patterning of deviation as such is not straightforwardly explainable, especially taking into account the results of Chapter 8, an influence of mismatched subphonemic durational differences was found, nonetheless.

How do the findings of the individual studies relate to the overarching goal of this book to draw a more detailed, intricate, and exhaustive picture of the production, perception, and comprehension of subphonemic detail? For production, it was found that different types of word-final /s/ are indeed different in terms of their acoustic duration. The nature of these differences is in line with previous corpus studies, but not with previous experimental studies. Analysing the durational differences not only by means of traditional variables but also by using measures derived from an LDL implementation, it was shown that such measures are predictive of word-final /s/ durations. Thus, the origin of durational differences in word-final /s/ can most likely be explained by the resonance of words with the lexicon. Taking into account the highly controlled methodology of the production study, its results, the measures derived from the LDL implementation, and the analyses of these measures, the first general aim of this book can be addressed: Subphonemic durational differences between different types of word-final /s/ exist. A potential explanation for the contradictory nature of previous results lies within the applied methodology and statistical analyses used in previous experimental studies. While these previous studies used homophonous real words as target items, I used pseudowords instead, avoiding the potential issues and uncertainties regarding the representation of homophones within the mental lexicon. Making use of an LDL implementation, pseudowords were shown not to be semantically empty, but to resonate with the lexicon. The measures derived from the LDL implementation, then, allowed for further insight into the origin of such durational differences. That is, higher degrees of semantic activation diversity and higher levels of phonological certainty come with shorter /s/ durations.

For perception, it was found that listeners showed higher sensitivity for durational differences of 35 ms and 75 ms as compared to the smaller differences of 10 ms and 20 ms. The results indicate that durational differences of 35 ms are somewhat perceptible, while durational differences of 75 ms show a further increased level of perceptibility. These results are more or less in line with the findings by Klatt & Cooper (1975) in that these authors claimed 25 ms to be the just-noticeable durational difference to a segment. As Klatt & Cooper (1975) also noted that durational differences in word-final fricatives are less well perceptible, the increase in sensitivity and thus perceptibility found for 35 ms is close to their 25 ms, but the threshold is most likely higher due to /s/ being word-final and a fricative. Regarding the second general aim of this book, then, one can conclude that the durational difference to a single segment to be perceptible should be at least of 35 ms if it is a fricative in word-final position. Note, however, the aforementioned issue of the metalinguistic nature of the same-different task on why the overall sensitivity was found to be rather low.

For comprehension, it was found that subphonemic durational differences indeed influence comprehension. Using target items with matched and mismatched durational /s/ information, it was found that mouse-trajectories for matched versus mismatched items were significantly different across all types of /s/ under investigation. This finding suggests that durational differences are used in comprehension, and should thus also be perceptible even though the overall low sensitivity values obtained in the perception task might suggest otherwise. Reaction times, on the other hand, did not significantly differ between matched versus mismatched item trials. However, reaction times only represent a single data point per trial while mouse-tracks give insight into the decision-making process during comprehension. Regarding the third general aim of this book, then, one may conclude that comprehension is influenced by subphonemic durational differences. More precisely, while the time between perception and the outcome of comprehension is not significantly influenced, the comprehension process between the input of an audio stimulus and the outcome of comprehension appears to be significantly affected.

So what does the overall picture of the production, perception, and comprehension of subphonemic detail look like? Subphonemic detail is influenced by morphological make-up as different types of word-final /s/ show differences in their acoustic durations and is perceptible if durational differences are above a certain threshold. Subphonemic detail influences and is made use of in the process of comprehension. As was demonstrated, these overall results ultimately call for revisions of models of speech production, perception, and comprehension which do not incorporate subphonemic detail in their pertinent representations and processes.

10 Conclusion

This book set out to investigate the production, perception, and comprehension of subphonemic detail. To operationalise the investigation, word-final /s/ in English was used in real word and pseudoword target items for a production task, an implementation of linear discriminative learning, a same-different task, and two number-decision tasks.

The first general aim of the present book was to examine whether durational differences in morphologically different types of word-final /s/ – non-morphemic, plural, *is*-, and *has*-clitic /s/ – can be found and how such differences can be accounted for. While previous studies reported such differences, the nature of these differences deviated between previous corpus studies and previous experimental studies. The results obtained in the production task presented in this book are in line with the findings of previous corpus studies. That is, non-morphemic /s/ is longest in duration, clitic /s/ is shortest in duration, and plural /s/ duration is in between non-morphemic /s/ and clitic /s/ durations. The two clitics under investigation were found not to be significantly different in terms of their durations. Turning to the results of the LDL implementation, it seems that the durational differences are connected to a word's resonance with the lexicon in that its semantic activation diversity and its phonological certainty are predictors of its word-final /s/ duration.

The second general aim of this book was to investigate how small a durational difference in word-final /s/ is perceptible. Using a same-different task, it was found that listeners showed a higher sensitivity for a durational difference of 35 ms as compared to smaller durational differences. This finding is more or less in line with previous work in that the just-noticeable durational difference should be at about 25 ms, but higher for word-final fricatives as is the case for word-final /s/.

The third general aim of this book was to find out whether subphonemic durational differences significantly influence comprehension. To investigate this issue, two number-decision tasks in a mouse-tracking paradigm were used. One task made use of isolated real words with either durationally matched or mismatched non-morphemic and plural /s/ duration, while the other task used pseudowords embedded within sentences with either durationally matched or mismatched plural, *is*-, and *has*-clitic /s/ duration as target items. It was found that reaction times are not influenced by the mismatch of durational information. However, both comprehension studies found a significant difference between mouse-tracks of trials of matched versus trials of mismatched durational information. Thus, the process of comprehension itself apparently is influenced by subphonemic detail, while the duration of the process of comprehension is not.

The investigation of the general aims revealed that a discernible number of extant models of speech production, perception, and comprehension cannot account for the present findings. Subphonemic durational differences are not predicted at all, or their directions are either unpredicted or said to be the opposite of what was found. The perception of subphonemic durational detail is ruled out completely, and an influence on comprehension is thus not considered. In light of the findings presented in this book, then, such models need to be revised. Yet, some promising, especially computational, approaches already exist. Future implementations of such accounts will show whether and how such approaches can be used to explain the intricacy of language structure. The complexities of speech production, perception, and comprehension remain enormous. The present book may have shed light on only a few of many issues: the production, perception, and comprehension of subphonemic detail. It was demonstrated by the findings of this book that various theoretical approaches to the production, perception, and comprehension of language and its fine-grained phonetic detail are in need of revision.

11 Supplementary material

The supplementary material for this book consists of additional tables sorted by chapters, scripts and data for all analyses, and a markdown documentation of the LDL implementation - all originally created for the dissertation this book is based on.

The supplementary material is available at: https://osf.io/rc7xj/

References

Afshartous, David & Richard A. Preston. 2011. Key results of interaction models with centering. *Journal of Statistics Education* 19(3). 1–35. DOI: 10.1080/10691898.2011.11889620.

Albright, Adam. 2002. Islands of reliability for regular morphology: Evidence from Italian. *Language* 78(4). 684–709.

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2). 119–161. DOI: 10.1016/S0010-0277(03)00146-X.

Anshen, Frank & Mark Aronoff. 1988. Producing morphologically complex words. *Linguistics* 26(4). 641–656. DOI: 10.1515/ling.1988.26.4.641.

Anvari, Sima H., Laurel J. Trainor, Jennifer Woodside & Betty Ann Levy. 2002. Relations among musical skills, phonological processing, and early reading ability in preschool children. *Journal of Experimental Child Psychology* 83(2). 111–130. DOI: 10.1016/S0022-0965(02)00124-8.

Arnold, Denis, Fabian Tomaschek, Konstantin Sering, Florence Lopez & R. Harald Baayen. 2017. Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLoS ONE* 12(4). e0174623. DOI: 10.1371/journal.pone.0174623.

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56. DOI: 10.1177/00238309040470010201.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511801686.

Baayen, R. Harald, Yu-Ying Chuang & Maria Heitmeier. 2019. *WpmWithLdl: Implementation of word and paradigm morphology with linear discriminative learning*. http://www.sfs.uni-tuebingen.de/~hbaayen/publications/WpmWithLdl_1.0.tar.gz.

Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan & James P. Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity* 2019. 4895891. DOI: 10.1155/2019/4895891.

Baayen, R. Harald, Laurie Beth Feldman & Robert Schreuder. 2006. Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language* 55(2). 290–313. DOI: 10.1016/j.jml.2006.03.008.

Baayen, R. Harald & Maja Linke. 2020. Generalized additive mixed models. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 563–591. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-46216-1_23.

Baayen, R. Harald & Petar Milin. 2010. Analyzing reaction times. *International Journal of Psychological Research* 3(2). 12–28. DOI: 10.21500/20112084.807.

Baayen, R. Harald, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118(3). 438–481. DOI: 10.1037/a0023851.

Baayen, R. Harald, Richard Piepenbrock & Leon Gulikers. 1995. *The CELEX lexical database (CD-ROM).* Philadelphia: University of Philadelphia.

Baayen, R. Harald, Cyrus Shaoul, Jon Willits & Michael Ramscar. 2016. Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience* 31(1). 106–128. DOI: 10.1080/23273798.2015.1065336.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278. DOI: 10.1016/j.jml.2012.11.001.

Barton, Kamil. 2020. *MuMIn: Multi-model inference.* https://cran.r-project.org/package=MuMIn.

Bates, Douglas, Martin Mächler, Benjamin M. Bolker & Steven C. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI: 10.18637/jss.v067.i01.

Beach, Sara D., Ola Ozernov-Palchik, Sidney C. May, Tracy M. Centanni, John D. E. Gabrieli & Dimitrios Pantazis. 2021. Neural decoding reveals concurrent phonemic and subphonemic representations of speech across tasks. *Neurobiology of Language* 2(2). 254–279. DOI: 10.1162/nol_a_00034.

Belke, Eva & Antje S. Meyer. 2002. Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *European Journal of Cognitive Psychology* 14(2). 237–266. DOI: 10.1080/09541440143000050.

Bell, Alan, Jason M. Brenier, Michelle Gregory, Cynthia Girand & Daniel Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1). 92–111. DOI: 10.1016/j.jml.2008.06.003.

Bell, Melanie J., Sonia Ben Hedia & Ingo Plag. 2021. How morphological structure affects phonetic realisation in English compound nouns. *Morphology* 31(2). 87–120. DOI: 10.1007/s11525-020-09346-6.

Ben Hedia, Sonia. 2019. *Gemination and degemination in English affixation: Investigating the interplay between morphology, phonology and phonetics*. Berlin: Language Science Press. DOI: 10.5281/zenodo.3232849.

Ben Hedia, Sonia & Ingo Plag. 2017. Gemination and degemination in English prefixation: Phonetic evidence for morphological organization. *Journal of Phonetics* 62. 34–49. DOI: 10.1016/j.wocn.2017.02.002.

Bender, Andreas, Andreas Groll & Fabian Scheipl. 2018. A generalized additive model approach to time-to-event analysis. *Statistical Modelling* 18(3-4). 299–321. DOI: 10.1177/1471082X17748083.

Bender, Andreas & Fabian Scheipl. 2018. Pammtools: Piece-wise exponential additive mixed modeling tools. https://arxiv.org/abs/1806.01042v1.

Bergen, Benjamin K. 2004. The psychological reality of phonaesthemes. *Language* 80(2). 290–311. DOI: 10.1353/lan.2004.0056.

Berko Gleason, Jean. 1958. The child's learning of English morphology. *WORD* 14(2-3). 150–177. DOI: 10.1080/00437956.1958.11659661.

Blazej, Laura J. & Ariel M. Cohen-Goldberg. 2015. Can we hear morphological complexity before words are complex? *Journal of Experimental Psychology: Human Perception and Performance* 41(1). 50–68. DOI: 10.1037/a0038509.

Boersma, Paul & David Weenink. 2019. *Praat: Doing phonetics by computer*. http://www.praat.org/.

Bolinger, Dwight L. 1950. Rime, assonance, and morpheme analysis. *WORD* 6(2). 117–136. DOI: 10.1080/00437956.1950.11659374.

Booij, Geert E. 1983. Principles and parameters in prosodic phonology. *Linguistics* 21(1). 249–280. DOI: 10.1515/ling.1983.21.1.249.

Bradley, A. Allen, Stuart S. Schwartz & Tempei Hashino. 2008. Sampling uncertainty and confidence intervals for the Brier score and Brier skill score. *Weather and Forecasting* 23(5). 992–1006. DOI: 10.1175/2007WAF2007049.1.

Brewer, Jordan B. 2008. *Phonetic reflexes of orthographic characteristics in lexical representation*. University of Arizona. (Doctoral dissertation). https://repository.arizona.edu/handle/10150/195213.

Brier, Glenn W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78(1). 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brysbaert, Marc, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte & Andrea Böhl. 2011. The word frequency effect. *Experimental Psychology* 58(5). 412–424. DOI: 10.1027/1618-3169/a000123.

Burani, Cristina, Francesca M. Dovetto, Alberto Spuntarelli & Anna M. Thornton. 1999. Morpholexical access and naming: The semantic interpretability of new root-suffix combinations. *Brain and Language* 68(1-2). 333–339. DOI: 10.1006/brln.1999.2073.

Burnage, Gavin. 1988. *CELEX, a guide for users*. Nijmegen: Centre for Lexical Information.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511612886.

Caselli, Naomi K., Michael K. Caselli & Ariel M. Cohen-Goldberg. 2016. Inflected words in production: Evidence for a morphologically rich lexicon. *Quarterly Journal of Experimental Psychology* 69(3). 432–454. DOI: 10.1080/17470218.2015.1054847.

Charlton, Martin. 2009. Quantitative data. In Rob Kitchin & Nigel Thrift (eds.), *International encyclopedia of human geography*, 19–26. Oxford: Elsevier. DOI: https://doi.org/10.1016/B978-008044910-4.00502-2.

Chavent, Marie, Vanessa Kuentz, Amaury Labenne, Benoit Liquet & Jerome Saracco. 2017. *PCAmixdata: Multivariate analysis of mixed data*. https://cran.r-project.org/package=PCAmixdata.

Cho, Taehong. 2001. Effects of morpheme boundaries on intergestural timing: Evidence from Korean. *Phonetica* 58(3). 129–162. DOI: 10.1159/000056196.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York: Harper & Row.

Chuang, Yu-Ying, Kaidi Lõo, James P. Blevins & R. Harald Baayen. 2020. Estonian case inflection made simple: A case study in word and paradigm morphology with linear discriminative learning. In Lívia Körtvélyessy & Pavol Štekauer (eds.), *Complex words: Advances in morphology*, 119–141. Cambridge: Cambridge University Press.

Chuang, Yu-Ying, Marie Lenka Vollmer, Elnaz Shafaei-Bajestan, Susanne Gahl, Peter Hendrix & R. Harald Baayen. 2021. The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods* 53(3). 945–976. DOI: 10.3758/s13428-020-01356-w.

Coetzee, Andries W. 2005. The obligatory contour principle in the perception of English. In Sónia Frota, Marina Vigário & Maria João Freitas (eds.), *Prosodies*, 223–245. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110197587.2.223.

Coetzee, Andries W. 2009. Grammar is both categorical and gradient. In Stephen Parker (ed.), *Phonological argumentation: Essays on evidence and motivation*, 9–42. London: Equinox.

Cohen, Clara. 2014. Probabilistic reduction and probabilistic enhancement: Contextual and paradigmatic effects on morpheme pronunciation. *Morphology* 24(4). 291–323. DOI: 10.1007/s11525-014-9243-y.

Cohen Priva, Uriel. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6(2). 243–278. DOI: 10.1515/lp-2015-0008.

Cooper, William E. & Martha Danly. 1981. Segmental and temporal aspects of utterance-final lengthening. *Phonetica* 38. 106–115. DOI: 10.1159/000260017.

Cribari-Neto, Francisco & Achim Zeileis. 2010. Beta regression in R. *Journal of Statistical Software* 34(2). 1–24. DOI: 10.18637/jss.v034.i02.

Ćwiek, Aleksandra, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, Gary Lupyan, Grace E. Oh, Jing Paul, Caterina Petrone, Rachid Ridouane, Sabine Reiter, Nathalie Schümchen, Ádám Szalontai, Özlem Ünal-Logacev, Jochen Zeller, Marcus Perlman & Bodo Winter. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences* 377(1841). DOI: 10.1098/rstb.2020.0390.

Dabrowska, Ewa. 2008. The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language* 58(4). 931–951. DOI: 10.1016/j.jml.2007.11.005.

Davies, Mark. 2004. *British national corpus (from Oxford University Press)*. https://www.english-corpora.org/bnc/.

de Jong, Nivja & Ton Wempe. 2008. *Praat script syllable nuclei v2 [praat script]*. https://sites.google.com/site/speechrate/speech-rate-praat-script-that-detects-syllable-nuclei/praat-script-syllable-nuclei-v2.

Dollaghan, Chilis. 1985. Child meets word: "fast mapping" in preschool children. *Journal of Speech and Hearing Research* 28. 449–454.

Drager, Katie K. 2011. Sociophonetic variation and the lemma. *Journal of Phonetics* 39(4). 694–707. DOI: 10.1016/j.wocn.2011.08.005.

Dye, Matthew W. G., C. Shawn Green & Daphne Bavelier. 2009. Increasing speed of processing with action video games. *Current Directions in Psychological Science* 18(6). 321–326. DOI: 10.1111/j.1467-8721.2009.01660.x.

Eddington, David. 2000. Analogy and the dual-route model of morphology. *Lingua* 110(4). 281–298. DOI: 10.1016/s0024-3841(99)00043-1.

Elsen, Hilke. 2008. *Phantastische Namen: Die Namen in Science Fiction und Fantasy zwischen Arbitrarität und Wortbildung*. Tübingen: Narr Francke Attempto.

Engemann, Marie & Ingo Plag. 2021. Phonetic reduction and paradigm uniformity effects in spontaneous speech. *The Mental Lexicon* 16(1). 165–198. DOI: 10.1075/ml.20023.eng.

Fasiolo, Matteo, Simon N. Wood, Margaux Zaffran, Raphaël Nedellec & Yannig Goude. 2021. Fast calibrated additive quantile regression. *Journal of the American Statistical Association* 116(535). 1402–1412. DOI: 10.1080/01621459.2020.1725521.

Ferrari, Silvia & Francisco Cribari-Neto. 2004. Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 31(7). 799–815. DOI: 10.1080/0266476042000214501.

Fort, Mathilde, Alexander Martin & Sharon Peperkamp. 2015. Consonants are more important than vowels in the bouba-kiki effect. *Language and Speech* 58(2). 247–266. DOI: 10.1177/0023830914534951.

Fox, John & Sanford Weisberg. 2019. *An R companion to applied regression*. Thousand Oaks: SAGE Publications, Ltd.

Fozard, James L., Max Vercruyssen, Sara L. Reynolds, P. A. Hancock & Reginald E. Quilter. 1994. Age differences and changes in reaction time: The Baltimore longitudinal study of aging. *Journal of Gerontology* 49(4). 179–189. DOI: 10.1093/geronj/49.4.P179.

Friedrich, Manuela & Angela D. Friederici. 2005. Phonotactic knowledge and lexical-semantic processing in one-year-olds: Brain responses to words and nonsense words in picture contexts. *Journal of Cognitive Neuroscience* 17(11). 1785–1802. DOI: 10.1162/089892905774589172.

Frisch, Stefan A., Nathan R. Large & David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42(4). 481–496. DOI: 10.1006/jmla.1999.2692.

Gahl, Susanne. 2008. *Time* and *thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84(3). 474–496. DOI: 10.1353/lan.0.0035.

Gahl, Susanne, Yao Yao & Keith Johnson. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66(4). 789–806. DOI: 10.1016/j.jml.2011.11.006.

Gahl, Susanne & Alan C. L. Yu. 2006. Introduction to the special issue on exemplar-based models in linguistics. *Linguistic Review* 23(3). 213–216. DOI: 10.1515/TLR.2006.007.

Gerds, Thomas A. & Martin Schumacher. 2006. Consistent estimation of the expected Brier score in general survival models with right-censored event rimes. *Biometrical Journal* 48(6). 1029–1040. DOI: 10.1002/bimj.200610301.

Gignac, Gilles E. & Philip A. Vernon. 2004. Reaction time and the dominant and non-dominant hands: An extension of Hick's Law. *Personality and Individual Differences* 36(3). 733–739. DOI: 10.1016/S0191-8869(03)00133-8.

Gneiting, Tilmann. 2011. Quantiles as optimal point forecasts. *International Journal of Forecasting* 27(2). 197–207. DOI: 10.1016/j.ijforecast.2009.12.015.

Goad, Heather. 1998. Plurals in SLI: Prosodic deficit or morphological deficit? *Language Acquisition* 7(2-4). 247–284. DOI: 10.1207/s15327817la0702-4_6.

Goad, Heather. 2002. Markedness in right-edge syllabification: Parallels across populations. *Canadian Journal of Linguistics* 47. 151–186.

Goad, Heather. 2003. Defective syntax or L1-constrained prosodic representations? *Canadian Journal of Linguistics* 48(3-4). 243–263.

Goad, Heather & Lydia White. 2019. Prosodic effects on L2 grammars. *Linguistic Approaches to Bilingualism* 9(6). 769–808. DOI: 10.1075/lab.19043.goa.

Goldinger, Stephen 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.

Goldinger, Stephen D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105(2). 251–279. DOI: 10.1037/0033-295X.105.2.251.

Goldstone, Robert L. & Andrew T. Hendrickson. 2010. Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science* 1(1). 69–78. DOI: 10.1002/WCS.26.

Gontijo, Possidonia F. D., Ivar Gontijo & Richard Shillcock. 2003. Grapheme-phoneme probabilities in British English. *Behavior Research Methods, Instruments, and Computers* 35(1). 136–157. DOI: 10.3758/BF03195506.

Gouskova, Maria & Michael Becker. 2013. Nonce words show that Russian *yer* alternations are governed by the grammar. *Natural Language & Linguistic Theory* 31(3). 735–765. DOI: 10.1007/s11049-013-9197-5.

Green, Bert F. & John W. Tukey. 1960. Complex analyses of variance: General problems. *Psychometrika* 25(2). 127–152. DOI: 10.1007/BF02288577.

Gries, Stefan Th. 2015. The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1). 95–125. DOI: 10.3366/cor.2015.0068.

Günther, Hartmut. 1983. The role of meaning and linearity in reading. In Florian Coulmas & Ehrlich Konrad (eds.), *Writing in focus*, 355–370. Berlin: Walter de Gruyter.

Hanique, Iris, Ellen Aalders & Mirjam Ernestus. 2013. How robust are exemplar effects in word comprehension? *The Mental Lexicon* 8(3). 269–294. DOI: 10.1075/ml.8.3.01han.

Hanique, Iris, Mirjam Ernestus & Barbara Schuppler. 2013. Informal speech processes can be categorical in nature, even if they affect many different words. *The Journal of the Acoustical Society of America* 133(3). 1644–1655. DOI: 10.1121/1.4790352.

Hawkins, Sarah & Rachel Smith. 2001. Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics* 13(1). 99–189.

Hendrix, Peter & Ching Chu Sun. 2020. A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning Memory and Cognition* 47(1). 157–183. DOI: 10.1037/xlm0000819.

Hsieh, Li, Laurence B. Leonard & L. Lori Swanson. 1999. Some differences between English plural noun inflections and third singular verb inflections in the input: The contributions of frequency, sentence position, and duration. *Journal of Child Language* 26(3). 531–543. DOI: 10.1017/S030500099900392X.

Hulme, Charles, Steven Roodenrys, Gordon Brown & Robin Mercer. 1995. The role of long-term memory mechanisms in memory span. *British Journal of Psychology* 86(4). 527–536. DOI: 10.1111/j.2044-8295.1995.tb02570.x.

Ivens, Stephen H. & Bertram L. Koslin. 1991. *Demands for reading literacy require new accountability methods.* Brewster: Touchstone Applied Science Associates.

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62. DOI: 10.1016/j.cogpsych.2010.02.002.

Jones, Michael N. & Douglas J. K. Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review* 114(1). 1–37. DOI: 10.1037/0033-295X.114.1.1.

Jurafsky, Daniel, Alan Bell & Cynthia Girand. 2002. The role of the lemma in form variation. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory phonology 7*, 3–34. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110197105.

Kaplan, Edward L. & Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282). 457–481. DOI: 10.1080/01621459.1958.10501452.

Kawahara, Shigeto. 2012. Lyman's Law is active in loanwords and nonce words: Evidence from naturalness judgment studies. *Lingua* 122(11). 1193–1206. DOI: 10.1016/j.lingua.2012.05.008.

Kawahara, Shigeto, Atsushi Noto & Gakuji Kumagai. 2018. Sound symbolic patterns in Pokémon names. *Phonetica* 75(3). 219–244. DOI: 10.1159/000484938.

Keating, Patricia A. 2006. Phonetic encoding of prosodic structure. In Jonathan Harrington & Marija Tabain (eds.), *Speech production: Models, phonetic processes, and techniques*, 167–186. New York: Psychology Press.

Kemps, Rachèl J. J. K., Mirjam Ernestus, Robert Schreuder & R. Harald Baayen. 2005. Prosodic cues for morphological complexity: The case of Dutch plural nouns. *Memory & Cognition* 33(3). 430–446. DOI: 10.3758/BF03193061.

Kemps, Rachèl J. J. K., Lee H. Wurm, Mirjam Ernestus, Robert Schreuder & R. Harald Baayen. 2005. Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes* 20(1-2). 43–73. DOI: 10.1080/01690960444000223.

Keuleers, Emmanuel, Kevin Diependaele & Marc Brysbaert. 2010. Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology* 1. DOI: 10.3389/fpsyg.2010.00174.

Kieslich, Pascal J. & Felix Henninger. 2017. Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods* 49(5). 1652–1667. DOI: 10.3758/s13428-017-0900-z.

Kieslich, Pascal J., Felix Henninger, Dirk U. Wulff, Jonas M. B. Haslbeck & Michael Schulte-Mecklenbeck. 2019. Mouse-tracking: A practical guide to implementation and analysis. In Michael Schulte-Mecklenbeck, Anton Kühberger & Joseph G. Johnson (eds.), *A handbook of process tracing methods*, 111–130. New York: Routledge. DOI: 10.31234/osf.io/zuvqa.

Kiparsky, Paul. 1982. Lexical morphology and phonology. In In-Seok Yang (ed.), *Linguistics in the morning calm: Selected papers from SICOL1*, 3–91. Seoul: Hanshin.

Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech and Language* 45. 326–347. DOI: 10.1016/j.csl.2017.01.005.

Klatt, Dennis H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America* 59(5). 1208. DOI: 10.1121/1.380986.

Klatt, Dennis H. 1979. Speech perception: A model of acoustic–phonetic analysis and lexical access. *Journal of Phonetics* 7(3). 279–312. DOI: 10.1016/S0095-4470(19)31059-9.

Klatt, Dennis H. & William E. Cooper. 1975. Perception of segment duration in sentence contexts. In Antonie Cohen & Sibout G. Nooteboom (eds.), *Structure and process in speech perception*, 69–89. Berlin: Springer. DOI: 10.1007/978-3-642-81000-8_5.

Koenker, Roger. 2005. *Quantile regression*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511754098.

Köhler, Wolfgang. 1929. *Gestalt psychology*. New York: Liveright.

Krämer, Martin. 2009. Main stress in Italian nonce nouns. In Danièle Torck & W. Leo Wetzels (eds.), *Romance languages and linguistic theory*, 127–142. Amsterdam: John Benjamins. DOI: 10.1075/cilt.303.08kra.

Kreft, Ita & Jan de Leeuw. 1998. *Introducing multilevel modeling*. London: SAGE Publications, Ltd. DOI: 10.4135/9781849209366.

Krivokapić, Jelena. 2007. Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics* 35(2). 162–179. DOI: 10.1016/j.wocn.2006.04.001.

Kuperman, Victor, Mark Pluymaekers, Mirjam Ernestus & R. Harald Baayen. 2007. Morphological predictability and acoustic duration of interfixes in Dutch compounds. *The Journal of the Acoustical Society of America* 121(4). 2261–2271. DOI: 10.1121/1.2537393.

Kuznetsova, Alexandra, Per B. Brockhoff & Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. DOI: 10.18637/jss.v082.i13.

Kwon, Nahyun & Erich R. Round. 2015. Phonaesthemes in morphological theory. *Morphology* 25(1). 1–27. DOI: 10.1007/s11525-014-9250-z.

Ladefoged, Peter. 2003. *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Malden: Wiley.

Lahiri, Aditi & William D. Marslen-Wilson. 1991. The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition* 38(3). 245–294. DOI: 10.1016/0010-0277(91)90008-R.

Landauer, Thomas K. & Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2). 211–240. DOI: 10.1037/0033-295X.104.2.211.

Lavoie, Lisa. 2002. Some influences on the realization of *for* and *four* in American English. *Journal of the International Phonetic Association* 32(2). 175–202. DOI: 10.1017/S0025100302001032.

Lee, Kyu Yup. 2013. Pathophysiology of age-related hearing loss (peripheral and central). *Korean Journal of Audiology* 17(2). 45–49. DOI: 10.7874/KJA.2013.17.2.45.

Lee, Sangho & Yung Hwan Oh. 1999. Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems. *Speech Communication* 28(4). 283–300. DOI: 10.1016/S0167-6393(99)00014-X.

Lee, Sue A. & Gregory K. Iverson. 2012. Stop consonant productions of Korean–English bilingual children. *Bilingualism: Language and Cognition* 15(2). 275–287. DOI: 10.1017/S1366728911000083.

Lee, Yoonjeong, Elsi Kaiser & Louis Goldstein. 2020. *I scream for ice cream*: Resolving lexical ambiguity with sub-phonemic information. *Language and Speech* 63(3). 526–549. DOI: 10.1177/0023830919866870.

Levelt, Willem J. M., Ardi Roelofs & Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22(01). 1–75. DOI: 10.1017/S0140525X99001776.

Linke, Maja, Franziska Bröker, Michael Ramscar & R. Harald Baayen. 2017. Are baboons learning "orthographic" representations? Probably not. *PLoS ONE* 12(8). e0183876. DOI: 10.1371/journal.pone.0183876.

Lohmann, Arne. 2018. *Time* and *thyme* are not homophones: A closer look at Gahl's work on the lemma-frequency effect, including a reanalysis. *Language* 94(2). e180–e190. DOI: 10.1353/lan.2018.0032.

Lupker, Stephen J., Mariko Nakayama & Masahiro Yoshihara. 2018. Phonologically-based priming in the same-different task with L1 readers. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 44(8). 1317–1324. DOI: 10.1037/xlm0000515.

Mack, Molly. 1982. Voicing-dependent vowel duration in English and French: Monolingual and bilingual production. *Journal of the Acoustical Society of America* 71(1). 173–178. DOI: 10.1121/1.387344.

Macmillan, Neil A. 1993. Signal Detection Theory as data analysis method and psychological decision model. In Gideon Keren & Charles Lewis (eds.), *A handbook for data analysis in the behaviorial sciences*, 21–58. New York: Taylor & Francis.

Macmillan, Neil A. & C. Douglas Creelman. 2005. *Signal Detection Theory: A user's guide.* 2nd edn. Mahwah: Lawrence Erlbaum Associates.

Makowski, Dominique. 2018. The psycho package: An efficient and publishing-oriented workflow for psychological science. *The Journal of Open Source Software* 3(22). 470. DOI: 10.21105/joss.00470.

Marian, Viorica, James Bartolotti, Sarah Chabal & Anthony Shook. 2012. Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS ONE* 7(8). e43230. DOI: 10.1371/journal.pone.0043230.

Marslen-Wilson, William D. 1984. Function and process in spoken word recognition: A tutorial review. In H. Bouma & D. Bouwhuis (eds.), *Attention & Performance X*. Hillsdale: Lawrence Erlbaum Associates.

Massaro, Dominic W. & Jeffry A. Simpson. 1987. *Speech perception by ear and eye.* New York: Psychology Press. DOI: 10.4324/9781315808253.

Mathôt, Sebastiaan, Daniel Schreij & Jan Theeuwes. 2012. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44(2). 314–324. DOI: 10.3758/s13428-011-0168-7.

Maurer, Daphne, Thanujeni Pathman & Catherine J. Mondloch. 2006. The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental Science* 9(3). 316–322. DOI: 10.1111/j.1467-7687.2006.00495.x.

McClelland, James L. & Jeffrey L. Elman. 1986. The trace model of speech perception. *Cognitive Psychology* 18(1). 1–86. DOI: 10.1016/0010-0285(86)90015-0.

McElreath, Richard. 2015. *Statistical rethinking: A Bayesian course with examples in R and stan.* Boca Raton: CRC Press.

McKay, Adam, Chris Davis, Greg Savage & Anne Castles. 2008. Semantic involvement in reading aloud: Evidence from a nonword training study. *Journal of Experimental Psychology: Learning Memory and Cognition* 34(6). 1495–1517. DOI: 10.1037/a0013357.

Meunier, Fanny & Catherine Marie Longtin. 2007. Morphological decomposition and semantic integration in word processing. *Journal of Memory and Language* 56(4). 457–471. DOI: 10.1016/j.jml.2006.11.005.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13) - Volume 2.*

Milin, Petar, Dagmar Divjak & R. Harald Baayen. 2017. A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43(11). 1730–1751. DOI: 10.1037/xlm0000410.

Milin, Petar, Laurie Beth Feldman, Michael Ramscar, Peter Hendrix & R. Harald Baayen. 2017. Discrimination in lexical decision. *PLoS ONE* 12(2). e0171935. DOI: 10.1371/journal.pone.0171935.

Milovanov, Riia, Minna Huotilainen, Paulo A. A. Esquef, Paavo Alku, Vesa Välimäki & Mari Tervaniemi. 2009. The role of musical aptitude and language skills in preattentive duration processing in school-aged children. *Neuroscience Letters* 460(2). 161–165. DOI: 10.1016/j.neulet.2009.05.063.

Mogensen, Ulla B., Hemant Ishwaran & Thomas A. Gerds. 2012. Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software* 50(11). 1–23. DOI: 10.18637/jss.v050.i11.

Moore, E. Hastings. 1920. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society* 26. 394–395.

Nakagawa, Shinichi, Paul C. D. Johnson & Holger Schielzeth. 2017. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface* 14(134). 20170213. DOI: 10.1098/rsif.2017.0213.

Nespor, Marina & Irene Vogel. 2007. *Prosodic phonology: With a new foreword.* Berlin: De Gruyter. DOI: 10.1515/9783110977790.

Norris, Dennis. 1994. Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52(3). 189–234. DOI: 10.1016/0010-0277(94)90043-4.

Norris, Dennis & Sachiko Kinoshita. 2008. Perception as evidence accumulation and Bayesian inference: Insights from masked priming. *Journal of Experimental Psychology: General* 137(3). 434–455. DOI: 10.1037/a0012799.

Norris, Dennis & James M. McQueen. 2008. Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115(2). 357–395. DOI: 10.1037/0033-295X.115.2.357.

O'Rourke, Norm, Larry Hatcher & Edward J. Stepanski. 2005. *Using SAS for univariete & multivariate statistics.* Cary: SAS Institute Inc.

Oh, Grace E. & Melissa A. Redford. 2012. The production and phonetic representation of fake geminates in English. *Journal of Phonetics* 40(1). 82–91. DOI: 10.1016/j.wocn.2011.08.003.

Ozubko, Jason D. & Steve Joordens. 2011. The similarities (and familiarities) of pseudowords and extremely high-frequency words: Examining a familiarity-based explanation of the pseudoword effect. *Journal of Experimental Psychology: Learning Memory and Cognition* 37(1). 123–139. DOI: 10.1037/a0021099.

Penrose, Roger. 1955. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society* 51(3). 406–413. DOI: 10.1017/S0305004100030401.

Pfeifer, Jasmin & Silke Hamann. 2018. The nature and nurture of congenital amusia: A twin case study. *Frontiers in Behavioral Neuroscience* 12(June). 1–11. DOI: 10.3389/fnbeh.2018.00120.

Pierrehumbert, Janet B. 2001. Exemplar dynamics. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency effects and the emergence of linguistics structure*. Amsterdam: John Benjamins. DOI: 10.1075/tsl.45.08pie.

Pierrehumbert, Janet B. 2002. Word-specific phonetics. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory Phonology 7*, 101–140. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110197105.1.101.

Pierrehumbert, Janet B. 2006. The statistical basis of an unnatural alternation. In Louis Goldstein, D. H. Whalen & Catherine T. Best (eds.), *Laboratory Phonology 8*. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110197211.1.81.

Pitt, Mark A., Leslie Dilley, Keith Johnson, Scott Kiesling, William D. Raymond, Elizabeth Hume & Eric Fosler-Lussier. 2007. *Buckeye corpus of conversational speech (2nd release)*. Columbus: Department of Psychology, Ohio State University.

Plag, Ingo, Julia Homann & Gero Kunter. 2017. Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics* 53(1). 181–216. DOI: 10.1017/S0022226715000183.

Plag, Ingo, Arne Lohmann, Sonia Ben Hedia & Julia Zimmermann. 2020. An *s* is an *s*, or is it? Plural and genitive-plural are not homophonous. In Lívia Körtvélyessy & Pavol Štekauer (eds.), *Complex words: Advances in morphology*. Cambridge: Cambridge University Press.

Pluymaekers, Mark, Mirjam Ernestus & R. Harald Baayen. 2005a. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62(2-4). 146–159. DOI: 10.1159/000090095.

Pluymaekers, Mark, Mirjam Ernestus & R. Harald Baayen. 2005b. Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America* 118(4). 2561–2569. DOI: 10.1121/1.2011150.

Pollack, Irwin & Donald A. Norman. 1964. A non-parametric analysis of recognition experiments. *Psychonomic Science* 1(1-12). 125–126. DOI: 10.3758/BF03342823.

Prasada, Sandeep & Steven Pinker. 1993. Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes* 8(1). 1–56. DOI: 10.1080/01690969308406948.

Pratha, Nimish K., Natalie Avunjian & Neil Cohn. 2016. Pow, punch, pika, and chu: The structure of sound effects in genres of American comics and Japanese manga. *Multimodal Communication* 5(2). 93–109. DOI: 10.1515/mc-2016-0017.

R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna, Austria. https://www.r-project.org/.

Ramsay, Timothy O., Richard T. Burnett & Daniel Krewski. 2003. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology* 14(1). 18–23. DOI: 10.1097/00001648-200301000-00009.

Ramscar, Michael & Daniel Yarlett. 2007. Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science* 31(6). 927–960. DOI: 10.1080/03640210701703576.

Ramscar, Michael, Daniel Yarlett, Melody Dye, Katie Denny & Kirsten Thorpe. 2010. The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science* 34(6). 909–957. DOI: 10.1111/j.1551-6709.2009.01092.x.

Rescorla, Robert A. 1988. Pavlovian conditioning: It's not what you think it is. *American Psychologist* 43(3). 151–160. DOI: 10.1037/0003-066X.43.3.151.

Rescorla, Robert A. & Allan R. Wagner. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Abraham H. Black & William F. Prokasy (eds.), *Classical conditioning II: Current research and theory*, 64–99. New York: Appleton-Century-Crofts.

Ridouane, Rachid & Pierre Hallé. 2017. Word-initial geminates: From production to perception. In Haruo Kubozono (ed.), *The phonetics and phonology of geminate consonants*, 66–84. Oxford: Oxford University Press.

Robinson, Cecil & Randall Schumacker. 2009. Interaction effects: Centering, variance inflation factor, and interpretation issues. *Multiple Linear Regression Viewpoints* 35(1). 6–11.

Roelofs, Ardi & Victor S. Ferreira. 2019. The architecture of speaking. In Peter Hagoort (ed.), *Human language: From genes and brains to behavior*, 35–50. Cambridge: MIT Press.

RStudio Team. 2020. *RStudio: Integrated development for R*. Boston, MA. http://www.rstudio.com/.

Rubenstein, Herbert, Lonnie Garfield & Jane A. Millikan. 1970. Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior* 9. 487–494.

Schiel, Florian. 1999. Automatic phonetic transcription of nonprompted speech. *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS 1999)*. 607–610.

Schmid, Helmut. 1999. Improvements in part-of-speech tagging with an application to German. In Susan Armstrong, Kennneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann & David Yarowsky (eds.), *Natural language processing using very large corpora*, 13–25. Dordrecht: Springer. DOI: 10.1007/978-94-017-2390-9_2.

Schmitz, Dominic. 2021a. *LDLConvFunctions: Functions for measure computation, extraction, and other handy stuff.* https://github.com/dosc91/LDLConvFunctions.

Schmitz, Dominic. 2021b. *Mtqgam: Mouse-tracking data in QGAMs.* https://github.com/dosc91/mtqgam.

Schmitz, Dominic, Dinah Baer-Henney & Ingo Plag. 2021. The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords. *Phonetica* 78(5-6). 571–616. DOI: 10.1515/phon-2021-2013.

Schmitz, Dominic, Hae-Eun Cho & Henrik Niemann. 2018. Vowel shortening in German as a function of syllable structure. In *Proceedings 13. Phonetik und Phonologie Tagung (P&P13)*, 181–184. Berlin.

Schmitz, Dominic & Janina Esser. 2021. *SfL: Statistics for Linguistics.* https://github.com/dosc91/SfL.

Schmitz, Dominic, Ingo Plag, Dinah Baer-Henney & Simon David Stein. 2021. Durational differences of word-final /s/ emerge from the lexicon: Modelling morpho-phonetic effects in pseudowords with linear discriminative learning. *Frontiers in Psychology* 12. DOI: 10.3389/fpsyg.2021.680889.

Schriefers, Herbert, Angela D. Friederici & U. Rose. 1998. Context effects in visual word recognition: Lexical relatedness and syntactic context. *Memory and Cognition* 26(6). 1292–1303. DOI: 10.3758/BF03201201.

Searle, Shayle R., George Casella & Charles E. McCulloch. 2009. *Variance components.* Hoboken: John Wiley & Sons.

Selkirk, Elisabeth. 1996. The prosodic structure of function words. In James Morgan & Katherine Demuth (eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, 187–213. New York: Lawrence Erlbaum Associates.

Sering, Konstantin, Petar Milin & R. Harald Baayen. 2018. Language comprehension as a multi-label classification problem. *Statistica Neerlandica* 72(3). 339–353. DOI: 10.1111/stan.12134.

Seyfarth, Scott. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133(1). 140–155. DOI: 10.1016/j.cognition.2014.06.013.

Seyfarth, Scott, Marc Garellek, Gwendolyn Gillingham, Farrell Ackerman & Robert Malouf. 2017. Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience* 33(1). 32–49. DOI: 10.1080/23273798.2017.1359634.

Shaoul, Cyrus & Chris Westbury. 2010. Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods* 42(2). 393–413. DOI: 10.3758/BRM.42.2.393.

Shatzman, Keren B. & James M. McQueen. 2006a. Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science* 17(5). 372–377. DOI: 10.1111/j.1467-9280.2006.01714.x.

Shatzman, Keren B. & James M. McQueen. 2006b. Segment duration as a cue to word boundaries in spoken-word recognition. *Perception and Psychophysics* 68(1). 1–16. DOI: 10.3758/BF03193651.

Shih, Stephanie S. & Deniz Rudin. 2020. On sound symbolism in baseball player names. *Names*. 1–18. DOI: 10.1080/00277738.2020.1759353.

Singson, Maria, Diana Mahony & Virginia Mann. 2000. The relation between reading ability and morphological skills: Evidence from derivational suffixes. *Reading and Writing* 12(3). 219–252. DOI: 10.1023/a:1008196330239.

Smith, Rachel, Rachel Baker & Sarah Hawkins. 2012. Phonetic detail that distinguishes prefixed from pseudo-prefixed words. *Journal of Phonetics* 40(5). 689–705. DOI: 10.1016/j.wocn.2012.04.002.

Smithson, Michael & Jay Verkuilen. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* 11(1). 54–71. DOI: 10.1037/1082-989X.11.1.54.

Snijders, Tom A. B. & Roel J. Bosker. 2011. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: SAGE Publications Ltd.

Spivey, Michael J., Marc Grosjean & Günther Knoblich. 2005. From the cover: Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences* 102(29). 10393–10398. DOI: 10.1073/pnas.0503903102.

Stanislaw, Harold & Natasha Todorov. 1999. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers* 31(1). 137–149. DOI: 10.3758/BF03207704.

Stein, Simon David & Ingo Plag. 2021. Morpho-phonetic effects in speech production: Modeling the acoustic duration of English derived words with linear discriminative learning. *Frontiers in Psychology* 12. DOI: 10.3389/fpsyg.2021.678712.

Sugahara, Mariko & Alice Turk. 2004. Phonetic reflexes of morphological boundaries at a normal speech rate. In Bernard Bel & Isabelle Marlien (eds.), *Proceedings of the International Conference on Speech Prosody 2004*, 353–356.

Sugahara, Mariko & Alice Turk. 2009. Durational correlates of English sub-lexical constituent structure. *Phonology* 26(3). 477–524. DOI: 10.1017/S0952675709990248.

Swanson, Lori A. & Laurence B. Leonard. 1994. Duration of function-word vowels in mothers' speech to young children. *Journal of Speech and Hearing Research* 37(6). 1394–1405. DOI: 10.1044/jshr.3706.1394.

ten Bosch, Louis & Lou Boves. 2021. Word competition: An entropy-based approach in the DIANA model of human word comprehension. In *Proceedings of Interspeech 2021*. DOI: 10.21437/Interspeech.2021-1394.

ten Bosch, Louis, Lou Boves & Mirjam Ernestus. 2015. DIANA, and end-to-end computational model of human word comprehension. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*.

Tomaschek, Fabian, Peter Hendrix & R. Harald Baayen. 2018. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics* 71. 249–267. DOI: 10.1016/j.wocn.2018.09.004.

Tomaschek, Fabian, Ingo Plag, Mirjam Ernestus & R. Harald Baayen. 2019. Phonetic effects of morphology and context: Modeling the duration of word-final s in English with naïve discriminative learning. *Journal of Linguistics* 57(2019). 123–161. DOI: 10.1017/S0022226719000203.

Tomaschek, Fabian, Benjamin V. Tucker, Matteo Fasiolo & R. Harald Baayen. 2018. Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard* 4(s2). DOI: 10.1515/lingvan-2017-0018.

Tomaschek, Fabian, Benjamin V. Tucker, Michael Ramscar & R. Harald Baayen. 2021. Paradigmatic enhancement of stem vowels in regular English inflected verb forms. *Morphology* 31(2). 171–199. DOI: 10.1007/s11525-021-09374-w.

Torreira, Francisco & Mirjam Ernestus. 2009. Probabilistic effects on French [t] duration. *Proceedings of Interspeech 2009* (September 2014). 448–451.

Tremblay, Antoine & Johannes Ransijn. 2020. *LMERConvenienceFunctions: Model selection and post-hoc analysis for (G)LMER models*. https://cran.r-project.org/package=LMERConvenienceFunctions.

Tucker, Benjamin V., Daniel Brenner, D. Kyle Danielson, Matthew C. Kelley, Filip Nenadić & Michelle Sims. 2019. The massive auditory lexical decision (MALD) database. *Behavior Research Methods* 51(3). 1187–1204. DOI: 10.3758/s13428-018-1056-1.

Tucker, Benjamin V., Michelle Sims & R. Harald Baayen. 2019. Opposing forces on acoustic duration. *PsyArXiv*. 1–38. DOI: 10.31234/osf.io/jc97w.

Turcsan, Gabor & Sophie Herment. 2015. Making sense of nonce word stress in English. In Jose A. Mompean & Jonás Fouz-González (eds.), *Investigating English pronunciation: Trends and directions*, 23–46. London: Palgrave MacMillan.

Turk, Alice & Stefanie Shattuck-Hufnagel. 2020. *Speech timing*. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198795421.001.0001.

Umeda, Noriko. 1977. Consonant duration in American English. *Journal of the Acoustical Society of America* 61(3). 846–858. DOI: 10.1121/1.381374.

van de Vijver, Ruben & Dinah Baer-Henney. 2014. Developing biases. *Frontiers in Psychology* 5. DOI: 10.3389/fpsyg.2014.00634.

Venables, William N. & Brian D. Ripley. 2002. *Modern applied statistics with S.* New York: Springer. DOI: 10.1007/978-0-387-21706-2.

Vitevitch, Michael S. & Paul A. Luce. 1998. When words compete: Levels of processing in perception of spoken words. *Psychological Science* 9(4). 325–329. DOI: 10.1111/1467-9280.00064.

Vitevitch, Michael S. & Paul A. Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers* 36(3). 481–487. DOI: 10.3758/BF03195594.

Wagner, Allan R. & Robert A. Rescorla. 1972. Inhibition in Pavlovian conditioning: Application of a theory. In R. A. Boakes & M. S. Halliday (eds.), *Inhibition and learning*, 301–334. London: Academic Press Inc.

Walsh, Liam, Jen Hay, Derek Bent, Liz Grant, Jeanette King, Paul Millar, Viktoria Papp & Kevin Watson. 2013. The UC QuakeBox project: Creation of a community-focused research archive. *New Zealand English Journal* 27. 20–32.

Walsh, Thomas & Frank Parker. 1983. The duration of morphemic and nonmorphemic /s/ in English. *Journal of Phonetics* 11(2). 201–206. DOI: 10.1016/s0095-4470(19)30816-2.

Warner, Natasha, Allard Jongman, Joan Sereno & Rachèl J. J. K. Kemps. 2004. Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics* 32(2). 251–276. DOI: 10.1016/S0095-4470(03)00032-9.

Wieling, Martijn, John Nerbonne & R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE* 6(9). e23613. DOI: 10.1371/journal.pone.0023613.

Wightman, Colin W., Stefanie Shattuck-Hufnagel, Mari Ostendorf & Patti J. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91(3). 1707–1717. DOI: 10.1121/1.402450.

Winter, Bodo. 2019. *Statistics for linguists: An introduction using R.* New York: Routledge. DOI: 10.4324/9781315165547.

Winter, Bodo & Martine Grice. 2021. Independence and generalizability in linguistics. *Linguistics* 59(5). 1251–1277. DOI: 10.1515/ling-2019-0049.

Winter, Bodo & Marcus Perlman. 2021. Size sound symbolism in the English lexicon. *Glossa: a journal of general linguistics* 6(1). DOI: 10.5334/gjgl.1646.

Winter, Bodo, Márton Sóskuthy, Marcus Perlman & Mark Dingemanse. 2022. Trilled /r/ is associated with roughness, linking sound and touch across spoken languages. *Scientific Reports* 12(1). 1035. DOI: 10.1038/s41598-021-04311-7.

Wood, Simon N. 2017. *Generalized additive models: An introduction with R.* New York: CRC Press. 1–476. DOI: 10.1201/9781315370279.

Yao, Yao. 2007. Closure duration and VOT of word-initial voiceless plosives in English in spontaneous connected speech. *UC Berkeley Phonology Lab Annual Report* 8. 183–225.

Zee, Tim, Louis ten Bosch, Ingo Plag & Mirjam Ernestus. 2021. Paradigmatic relations interact during the production of complex words: Evidence from variable plurals in Dutch. *Frontiers in Psychology* 12. DOI: 10.3389/fpsyg.2021.720017.

Zimmermann, Julia. 2016. Morphological status and acoustic realization: Findings from New Zealand English. In Christopher Carignan & Michael D. Tyler (eds.), *Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology (SST-2016)*, 201–204. Canberra: ASSTA.

Zuur, Alain F., Elena N. Ieno & Chris S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* 1(1). 3–14. DOI: 10.1111/j.2041-210X.2009.00001.x.

Zvonik, Elena & Fred Cummins. 2003. The effect of surrounding phrase lengths on pause duration. In *Proceedings of Eurospeech 2003*, 777–780. Geneva.

Name index

Production, perception, and comprehension of subphonemic detail

The complexities of speech production, perception, and comprehension are enormous. Theoretical approaches of these complexities most recently face the challenge of accounting for findings on subphonemic differences. The aim of the present dissertation is to establish a robust foundation of findings on such subphonemic differences.

One rather popular case for differences in subphonemic detail is word-final /s/ and /z/ in English (henceforth S) as it constitutes a number of morphological functions. Using word-final S, three general issues are investigated. First, are there subphonemic durational differences between different types of word-final S? If there are such differences, how can they be accounted for? Second, can such subphonemic durational differences be perceived? Third, do such subphonemic durational differences influence the comprehension of S?

These questions are investigated by five highly controlled studies: a production task, an implementation of Linear Discriminative Learning, a same-different task, and two number-decision tasks. Using not only real words but also pseudowords as target items, potentially confounding effects of lexical storage are controlled for.

Concerning the first issue, the results show that there are indeed durational differences between different types of word-final S. Non-morphemic S is longest in duration, clitic S is shortest in duration, and plural S duration is in-between non-morphemic S and clitic S durations. It appears that the durational differences are connected to a word's semantic activation diversity and its phonological certainty. Regarding the second issue, subphonemic durational differences in word-final S can be perceived, with higher levels of perceptibility for differences of 35 ms and higher. In regard to the third issue, subphonemic durational differences are found not to influence the speed of comprehension, but show a significant effect on the process of comprehension. The overall results give raise to a revision of various extant models of speech production, perception, and comprehension.