Internal and external factors in a Bavarian sound change: Agent-based simulations and measurements in apparent time

Markus Jochim

München 2022

Internal and external factors in a Bavarian sound change: Agent-based simulations and measurements in apparent time

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
der Ludwig-Maximilians-Universität München

vorgelegt von
Markus Jochim

aus
München

2022

Referentin:	Dr. Felicitas Kleber
Korreferent:	Prof. Dr. Jonathan Harrington
Tag der mündlichen Prüfung:	02.07.2020

Acknowledgments

So ... working on this thesis has indeed come to an end. After many years of ups and downs, I am grateful not only for the very fact that all journeys do end at some point, but even more so to the many, many people who have supported me in my academic life. First and foremost to Felicitas, who not only has been a tremendously helpful and constructive PhD advisor, but also taught me phonetics "von der Pike auf"; that is, starting from my very first year as an undergrad student. We have discussed countless ideas and issues, we have certainly had our share of disagreements, but at the end of the day I have always enjoyed working with her and learning a lesson or two. Or three. A sizable amount, anyway.

Then, in no particular order, I would also like to thank Ulrich and Jessica, with whom I had the honour of sharing an office when I started my PhD and who answered all my questions about statistics, research, English and really anything. Later, those same office seats were taken by two of my closest friends and "doctor sisters," Kati and Miri. I have been really lucky that we have shared so much of our academic and personal lives. The same goes for Niki; I am not sure we ever actually went dancing in the rain, but still we had no shortage of Vida Loca or at least not of Ricky Martin. I would also like to thank Raphy, who, amongst his friends, is considered beautiful; who inhabited the office next door and sparked many interesting conversations on software development, with many hairs turned gray over computers that sucked; and who still hasn't quit IT for gardening.

There have been a great number of people at the IPS over the years (might I say a fine selection of people?) who have made the institute a great place to research and simply to be; I am afraid that to name all of them here would make this section longer than the actual thesis and I have a feeling that that might be inappropriate. Still, I would like to thank Jonathan for directing the institute in a very productive and encouraging way.

Last but not least, I would like to thank my parents and sisters (the real sisters, not the doctor ones) for always supporting and encouraging me to find my way. Even when more than ten years ago I told them about this program that no-one had ever heard of before and that bore the strange name "phonetics and speech processing."

Overview

Contents

List of Figures

List of Tables

1 Introduction

> A world of unchanging linguistic excellence, based on the brilliance of earlier literary forms, exists only in fantasy.
>
> *David Crystal (1987, p. 328)*

Languages change over time. Many discussions in the field of linguistics have revolved around what exactly triggers these changes and how they spread through a language community. Between the moment an innovation like a new word, a new phoneme or a new pronunciation of an existing phoneme is used for the first time and the moment this innovation is considered standard usage in a variety, a period of several decades usually passes, spanning a number of generations of language users. Neither of these moments can be determined exactly. In fact, there is no such thing as a particular moment in which an innovation turns from non-standard to standard usage. All of this makes empirical work on the subject difficult: naturally-occurring language changes are hard to observe because researchers cannot predict when they start taking place; and once researchers realize a change is in progress, it may be too late to collect data about the initiation of the change, the language's old surface forms (phonetic, morphosyntactical, etc.) and the (neuro-)physiological processes that governed the language before the change. Creating a language change in the laboratory is inherently hard to implement and observe because the process takes decades. Still, some empirical methods of investigating language change do exist. The present thesis focuses on sound change, exploring and applying two of these methods: observations in apparent time and observations in agent-based simulations of sound change. The individual chapters will be concerned to varying degrees with theoretical questions, empirical research that contributes to these questions, and methodological questions of how to conduct the empirical research. Thus, depending on the reader's focus, the chapters need not be read strictly in order. In the following, I will lay out how the various perspectives are addressed and show how the chapters interconnect. In doing so, I will employ a dichotomy between theory and methodology; some issues that readers might reasonably expect to be labelled empirical rather than theoretical fall on the theory side of this dichotomy and are labelled accordingly.

On the theoretical side, the main focus of this dissertation is on a sound change in Central Bavarian – a German dialect – and its implications for sound change theory. This focus is discussed in Chapters 4 and 5. The change affects a phonotactical rule in Central Bavarian's grammar that determines the distribution of quantity in vowels and consonants; that is, the role of sound durations in the grammar of the language. The affected rule is a major feature of Central Bavarian called Pfalz's law (Pfalz, 1911).

The rule is detailed in Section 1.1 and Chapter 4. We aim to leverage our investigation of this sound change to contribute to broader questions, too; in particular, the role of language-internal and external/social trigger mechanisms of sound change. A short overview of the issue of internal vs. external trigger mechanisms is given in Section 1.2. Another contribution to the theoretical side of this dissertation is in Chapter 3, where we broaden the scope of investigated languages. The main scope for this dissertation spans Standard German and Central Bavarian German, but in Chapter 3, we include Finnish – a Finno-Ugric language, genetically very far away from the Germanic family. Finnish is known to make strong use of quantity (e. g. Lehtonen, 1970). In Chapter 3, we establish empirically that this strong use also holds across generations. This allows us and other researchers to have a firm laboratory-phonological point of reference for apparent-time analyses of quantity in other languages – like the analysis in Chapter 4. In order to understand the relation between the languages, Section 1.1 gives a short overview of the relevant parts of the phonology of Standard German, Central Bavarian and Finnish.

On the methodological side, the main focus is on the application of agent-based modeling (ABM, see Cioffi-Revilla, 2017a; Manzo, 2014) to sound change research. Some epistemological concerns about this are addressed in the later sections of Chapter 1. Chapter 5 comprises a study that tests the validity of this relatively novel research method (see Harrington, Kleber, Reubold, Schiel, et al. (2019) for a historical overview). At the same time, however, Chapter 5 will also apply ABM – in a manner that is already known to be valid – to inform future theoretical research on sound change. Specifically, it will contribute to the above-mentioned discussion of language-internal vs. social trigger mechanisms – so it is at this point, in the last study of this dissertation, where methodological and theoretical questions are brought back together.

This dissertation also draws heavily on the apparent time paradigm, a method of comparing the language of different generations (see Section 1.3). Apparent time has been used in many linguistic studies and can therefore be considered a *method* that is stable (Bailey et al., 1991; Labov, 1963; see also Section 1.3). Like Finnish – a *language* we consider stable –, it is suitable as a reference, and we use it as the point of reference to test ABM against (Chapter 5). Chapters 3 and 4 use apparent time not in a methodological way, but plainly as a method – to test the stability of parts of the phonology of two languages.

A further methodological focus is on a part of the *EMU Speech Database Management System* or *EMU-SDMS* (Winkelmann et al., 2017): the *emuDB Manager* described in Chapter 2. It was developed as a part of this dissertation to address the problems of creating speech corpora (Draxler, 2008; Harrington, 2010) across many institutions. The empirical data that Chapters 3, 4 and 5 are based on are subsets of corpora created during a trinationally funded research project[1], making heavy use of both the emuDB

[1]Typology of vowel and consonant quantities in Southern German varieties, funded by the DFG (Germany), FWF (Austria) and SNF (Switzerland), directed by Felicitas Kleber, Sylvia Moosmüller (†),

Manager and other parts of the EMU-SDMS. The emuDB Manager allowed the research team to cleanly organize and analyze the corpora while having multiple labellers work on them at the same time. The data collection and analysis would have been much more difficult without specifically developing this tool.

The present Chapter 1 serves to briefly outline the languages investigated in the later chapters (Section 1.1) as well as the dichotomies of origin and spread of sound change and internal vs. external factors contributing to them (Section 1.2). It then proceeds to discuss some methodological questions: The field of phonetics has seen a wide variety of "real," that is, non-simulated measurement techniques and experimental setups, among them acoustic recordings, palatography, electromagnetic articulography, careful listening, spectral analyses, forced-choice classification tasks, eye tracking, and event-related potentials. All of them have their strengths and weaknesses, and the explanatory power of many of them has been thoroughly discussed, e. g. in Ladefoged's (2003) seminal book *Phonetic Data Analysis*, or in such volumes as *Coarticulation: Theory, Data and Techniques* (Hardcastle & Hewlett, 1999) and *The Handbook of Phonetic Sciences* (Hardcastle et al., 2010). It is one of the aims of this thesis to contribute to a similar discussion about simulation techniques in phonetics, which is useful in its own right and also, for the purposes of this thesis, particularly necessary in order to fully appreciate the simulation study in Chapter 5. Another important contribution to this discussion is found in de Boer (2006).

Since this discussion has received less attention in the phonetic literature than I believe it merits so far, I will use most of the following sections of this introduction to delve into some epistemological concerns behind simulations in general and agent-based modeling in particular. After looking at the apparent and real time paradigms to have a baseline to compare against (Section 1.3), I will describe some properties (Section 1.4.1), key terminology (1.4.2), and examples (1.4.3) of simulations. This will culminate in Section 1.4.4, a list of four principal kinds of insights that simulation studies have the potential to yield for the advancement of linguistic theory. I will then move on to a wrap-up in Section 1.5 and bring together the apparent time paradigm, the real time paradigm (i. e. longitudinal studies), and agent-based simulation.

1.1 Languages under investigation

This dissertation is concerned with three languages: German, Central Bavarian (a dialect of German), and Finnish. This section is to show why they were chosen and it will give a very brief overview of these languages; specifically, due to the nature of the investigated sound change, an overview of how their grammar allows to combine short/long vowels with short/long consonants.

Standard German and Central Bavarian are members of the Germanic language family, which in turn is a part of the Indo-European language family. German is spoken

Michael Pucher and Stephan Schmid.

throughout a number of countries in central Europe, especially in Germany, Austria and Switzerland. The Central Bavarian dialect is spoken in parts of Austria and in parts of Bavaria, which is a federal state of Germany. Exactly how many people speak German and/or the dialect is a research question in itself and is of no particular concern here. The number of inhabitants of the respective regions can serve as a rough indicator of the magnitude, though: Germany has around 83 million inhabitants, with 13 million of them living in the state of Bavaria. Austria has around 9 million inhabitants.

Finnish is the official language of Finland in Northern Europe, and it is spoken by around 5 million people. It is a member of the Finno-Ugric language family, which is a part of the Uralic, rather than the Indo-European family. The genetic distance between Finnish on the one hand and German and Bavarian on the other hand is therefore quite large.

Due to the nature of the Central Bavarian sound change we are investigating, the phonological feature of concern in these languages is quantity in both vowels and consonants. Bannert (1976) shows a typology that classifies languages along these features. The typology allows languages to have phonological quantity contrasts in (a) vowels only, or (b) consonants only, or (c) both vowels and consonants, independently of each other, or (d) interdependently in both vowels and consonants. In his typology, Central Bavarian falls in category d and Finnish in category c; Standard German is not mentioned, but it contains features of types a and c (see below). Our investigation is concerned with whether Central Bavarian is in the process of changing towards type c.

In order to have a solid empirical baseline of a type c language, we chose to include Finnish in our investigations. Chapter 3 is dedicated to the analysis of Finnish. In that chapter, we test how stable quantity in Finnish is across generations, by means of the apparent-time paradigm. Standard German, on the other hand, serves as a control language in Chapter 4 to see whether younger as opposed to older dialect speakers approximate standard speakers on a given phonetic dimension as a result of dialect levelling (Hinskens, 1998). With regard to the typology, Standard German can either be assigned to type a or c. It has a long–short contrast in vowels, but almost all vowel pairs also exhibit a major quality contrast (Wiese, 1996) It also has a two-way contrast in consonants, but that is most often termed fortis–lenis and not short–long. However, the fortis–lenis contrast is cued, among others, by duration (Kohler, 1979; Wiese, 1996). Both Finnish and Standard German, therefore, allow a four-way contrast of word types, where a short or long vowel is followed by a short/lenis or long/fortis consonant; however, in Standard German, quantity is only one of several cues while Finnish relies much more on quantity. Table 1.1 gives examples of each word type.

What does interdependence mean in type d, specifically in Central Bavarian? Phonetically, the language has two measurable vowel lengths and two measurable consonant lengths. Phonologically, however, they cannot be freely combined: Long vowels can only be followed by short consonants and short vowels can only be followed by long consonants. Many authors such as Hinderling (1980) claim that the vowel length is allophonic. If one were to follow this interpretation, Central Bavarian would be type b rather than

c. Bannert (1976), however, suggests that vowel and consonant should be regarded as a phonotactic unit and the phonological contrast is not between long and short consonant, but rather between the long plus short and short plus long combinations in a vowel-consonant sequence. Table 1.1 also illustrates which word types are considered illegal in the Central Bavarian grammar.

Language	**V:C**	**V:C:**	**VC**	**VC:**
Standard German	wieder	Bieter	Widder	bitter
Finnish	koota	tuutti	kota	tutti
Central Bavarian	legal	illegal	illegal	legal

Table 1.1: Illustration of word types in Finnish, Standard German and Central Bavarian. V denotes a short vowel, V: a long vowel, C a lenis or short consonant and C: a fortis or long consonant.

1.2 Internal and external factors in the origin and spread of sound change

Sound change researchers often consider two questions – although almost always independently of one another (see Harrington, Kleber, Reubold, Schiel, et al., 2019): First, the origin of sound change – how does a linguistic innovation start? (see e. g. Beddor, 2009; Ohala, 1981, 1993b; Solé and Recasens, 2012) And second, the spread of sound change – how does that innovation eventually turn from being an innovation used by few to a linguistic norm used by many? (see e. g. Baker et al., 2011) In trying to categorize factors contributing to these processes, linguists sometimes employ a dichotomy of *internal* vs. *external* factors (see e. g. Milroy, 2003; Torgersen and Kerswill, 2004). Internal, in this dichotomy, refers to linguistic mechanisms/factors that can be found inside the language(s) undergoing change. One hypothetical example would be a language with a large number of acoustically similar – and therefore easy-to-mix-up – back vowels and a low number of front vowels. Such an imbalance can be hypothesized to trigger a shift of some vowel phonemes to be articulated farther to the front in the vocal tract. External, on the other hand, refers to mechanisms/factors that are outside the linguistic system itself. These are sometimes also referred to as *social* factors. Examples of this include language policy; that is, when authorities or private groups/institutions create guidelines on language use, e. g. on whether and how to use gender-neutral speech. Another example of external/social factors is contact between language communities. Such contact can lead to mutual influence between two or more languages. Note, however, that this mutualness does not imply that the languages influence each other to the same degree; the mutualness can be lopsided.

Within this dissertation, Chapter 4 aims to deal with the origin of sound change. It tests and discusses the possibility of a particular internal factor – namely, speech-rate-induced hypoarticulation – triggering the investigated Central Bavarian sound change. Chapter 5, on the other hand, focusses more on the spread of sound change. A guiding question in that chapter is, would the theoretical linguistic ideas behind a particular computer simulation indeed predict that a certain innovation will spread through a language community and eventually become a norm? While it discusses both external and internal factors, the stronger focus of Chapter 5 is on the external factor language contact between Standard German and the German Dialect Western Central Bavarian.

1.3 Methods of sound change research

Sound change is a process that takes decades and is therefore very hard to observe directly in its entirety, let alone manipulate experimentally. Hence, it is necessary to consider a variety of research paradigms for investigating sound change. Established paradigms include the *apparent time* approach (Bailey et al., 1991; Labov, 1963), where different generations of a language community are tested at the same time. Two different approaches are sometimes called *real time* in the field of language change research: longitudinal studies, where the same members of a language community are tested repeatedly over many years (Harrington, 2007; Reubold et al., 2010; Sankoff & Blondeau, 2007); and an approach where different participants are tested at different points in time, but all participants are the same age at the respective time of testing (e. g. Rathcke & Stuart-Smith, 2016). Established paradigms also include text rather than audio-based language comparisons in the framework of historical linguistics established by the neogrammarians (the comparative method, see e. g. Kümmel, 2007). Computer simulation is, in comparison, a new addition to the toolbox.

The most direct among these paradigms are longitudinal studies. However, since the process "natural sound change" takes so long, a longitudinal study needs several decades' worth of data. This can be done with speech data that exist independently of the study; that is, data that were generated for very different reasons and usually by other people than the investigators (e. g. with broadcast data, see Reubold et al., 2010). However, if no such data exist for a particular research question, it is usually impossible for an investigator to conduct data collection for such a long time span (but see examples for such endeavors in Sankoff and Blondeau, 2007). This is especially true when testing hypotheses – rather than exploring reality –, which typically requires very specific data.

Due to the impracticability of letting several generations' worth of real time pass, the paradigm of apparent time was devised. In this paradigm, all observations of a language community are made at the same time, but more than one age group is observed. If, say, one group is, on average, 20 years old and another 70 years, then 50 years pass in this study – not really, but apparently, hence the name. This has two tremendous advantages. Obviously, the amount of time required to conduct the study is now decoupled from

the amount of time the investigator wants to observe (a chosen number of generations). Additionally, it is now possible to do actual experimental manipulation; that is, variables of interest can now be controlled for, such as gender, speech rate, or the elicited speech material.[2]

On the downside, apparent time studies measure an abstraction of the change process, while longitudinal studies can potentially measure the actual process. It is therefore necessary to test whether observations in apparent time in fact reflect observations to be made in real time. One major concern is that an age group difference observed in apparent time might reflect age-conditioned language change that repeats in every generation instead of language change that separates all former generations from all later generations.[3] Such tests have been conducted by comparing the results of apparent time studies with available real time data. Bailey et al. (1991) conclude from their comparison of a number of phonological variables in Texas that apparent time is a robust analytical method. Labov (1994) gives an overview of four such validation studies (Cedergren, 1973; Fowler, 1986; Hermann, 1929; Trudgill, 1988). He, too, finds that real-time data did support apparent-time observations. He also points out that in many comparisons, both generational and age-conditioned changes have been discovered; however, Labov puts forward that it may be misleading to regard the two as an opposition and that age-conditioned changes may well be a factor contributing to the mechanism of generational language change. Overall, these comparisons show that the apparent time paradigm is, generally speaking, a valid abstraction of observations made in real time.

With this validation in place, apparent time can be considered a very powerful research method. In this thesis, we want to explore computer simulations of sound change as a method, since they are increasingly used in phonetic and phonological studies of sound change research (e. g. Harrington and Schiel, 2017; Kirby, 2014; Todd et al., 2019). Simulations are inherently, to some degree, decoupled from reality. When evaluating them, we therefore need some way to check the credibility of simulation results. One good way to do that is to compare simulation results with results from the apparent time paradigm, to see if they match. We will now turn our focus to the basic ideas of simulation.

1.4 Simulation

1.4.1 Concept and goals of (agent-based and other) simulations

In terms of its broad goals, *simulation* can be regarded as a method of testing hypotheses in areas of scientific interest where conducting experiments with human participants or

[2]This is also possible in real time studies if they actually plan their own data collection, but not in those real time studies that rely on pre-existing data.

[3]Note that this difference must also be dealt with in in longitudinal studies (see e. g. Reubold and Harrington, 2015).

real-world objects is infeasible for practical reasons (e. g. financial, manpower or time limitations, see Cioffi-Revilla, 2017a, p. 378). Obviously, however, conclusions drawn from a simulation model cannot necessarily be generalized to the real world (note that the same is true of laboratory experiments, where generalizability depends on the quality of the specific design). We will discuss this limitation in the following sections.

Simulation can also be regarded, by its very definition, as a process by which a real-world phenomenon is implemented in computer software such that researchers or engineers can observe the phenomenon without anything actually happening in the real world. Although simulations are but mathematical models and could theoretically be calculated by hand, usually they are not. It takes too long. However, some famous simulations have indeed been run without computers, e. g. Schelling's (1971) simulation of social segregation. The mathematical models that make up the simulations are abstract theoretical descriptions of the *referent system*, that is, the object(s) of investigation. Calculating the results of those models is what it means to run a simulated experiment.

Agent-based modeling (ABM) is a particular simulation technique where the researcher's theoretical description of the referent system relies on describing the behavior of individual independent agents and then describing the agents' interactions, instead of describing the entire referent system as one unit. ABM is particularly well-suited to studying social phenomena where many participants interact with each other, all of whom have their own goals and strategies, and where the processes involved cannot be controlled by a single entity. In such phenomena, the participants together create a highly dynamic macrostructure (e. g. the vowel system of a language or market prices). ABM provides a way of formalizing and simulating this and of investigating the relationship between individual behavior and macrostructure. Examples of phenomena that have been studied with ABM include economic market decisions (e. g. Raberto et al., 2001), social segregation[4] (e. g. Schelling, 1971) and, crucially, also language and language change (e. g. Harrington & Schiel, 2017).

In sound change research, agent-based simulation models are a computer simulation in which real speaker-listeners are represented by virtual agents (i. e. computer programs) who interact according to certain rule sets. The agents interact with each other by exchanging tokens of de Saussurean *parole*, that is, language in use, in the form of acoustic speech signals or feature vectors describing properties of such signals. The speech signals can either be taken from laboratory phonology corpora (e. g. in Harrington & Schiel, 2017) or be synthesized (e. g. in de Boer, 2000). In the broader application of language change (rather than sound change), tokens could also be textual representations of *parole*. Upon receiving a token, *something* changes in the receiving agent's linguistic system, which we conceptualize in this thesis as an "exemplar-based phonological

[4]Technically, Schelling (1971) used a cellular automaton (CA), which is often described as a type of simulation similar to ABM. However, it can also be considered a special case of ABM, one where the interactions between agents are more constrained, because "cells" (the CA equivalent of agents) can only interact with their direct neighbors and the cells' location in the network structure cannot change.

mind," following exemplar models of speech perception (Johnson, 1997; Pierrehumbert, 2003; although ABM as a method is not at all tied to exemplar theory). What exactly changes, and which agents get to exchange how many tokens with which of their fellow agents, is subject to modeling decisions (which must in turn be based on theoretical considerations).

1.4.2 Key terminology: Choosing a microstructure and generating a macrostructure

This section introduces and defines simulation-specific technical terms used in this thesis. They are printed in boldface.

Researchers have to formalize and implement their theory of how the **referent system** works. Cioffi-Revilla (2017a) defines a referent system as a "real-world system or process that is an object of investigation (*explanandum*)", and not open-ended but rather "*defined* or *specified* by the specific *research questions* being investigated" (p. 379, emphasis in the original).

In the formalization and implementation, researchers obviously *choose* the microstructures found in the agents, e. g. details of the speech perception apparatus and interaction rules/regularities, and they choose them based on their theory. It is one property of simulations that researchers have to fully specify their theory, that is, they cannot leave out any details (Manzo, 2014). This is the simulation's **theoretical underpinnings**. They are usually coupled with a variety of **parameters** that can be adjusted. For example, if the theory dictates that some event in the referent system happen with a certain probability but does not state a value for this probability, it can be adjusted between 0 and 1 (or a narrower range if the theory dictates that).

The **input to be tested** is the material fed into the model: a macrostructure that the system departs from, e. g. the vowel space of the involved populations of agents, or the distribution of a certain phonetic cue. In our case, this macrostructure is determined by the acoustic speech signals or feature vectors that form the agents' starting memory. Technically, these are properties of individual agents and therefore microstructure, not macrostructure. However, at the same time, the entirety of the individual agents' speech tokens taken together makes up a macrostructural phonological system (e. g. a community's vowel space). So whether we consider the microstructure (speech tokens regarded as individual agents' properties) or the macrostructure (the aggregate properties of those tokens) as our *input to be tested* is really a matter of our research question.

What the simulation now does is apply the chosen microstructure, that is, run the interactions between agents using the input to be tested, and observe the resulting macrostructure. The resulting macrostructure (vowel space, cue distributions, etc.) is the **outcome of interest**. Appropriate hypotheses would be that we expect certain output macrostructures to result from a given input macrostructure. Such expectations should be informed by real data. Agent-based models "provide computational demon-

strations that a given microspecification is in fact sufficient to generate a macrostructure of interest" (Epstein, 2006, p. 8).

1.4.3 Non-linguistic examples of simulations

Now that we have described a framework of what simulations are, let us consider two examples of simulations. The examples were chosen so as to illustrate, without much theoretical detour, (a) the terminology introduced in Section 1.4.2, and (b) two referent systems with relatively little and relatively high complexity, respectively. To illustrate these two points, it does not matter whether the examples are linguistic or from any other discipline.

Looking at a physical phenomenon, air flow, one goal of simulations in engineering is to find out the aerodynamic properties of structures such as vehicles, bridges or urban buildings without (or before) physically building them. This offers a *referent system* with structural parts in a certain shape, made of certain materials, moving at a certain speed through air (or air moving around the parts at a certain speed). It is then of interest how exactly the flow of air is changed by the presence of the structure. This would be possible if the mechanisms underlying both laminar and turbulent air flow were thoroughly understood. Since this is not yet the case (Eames & Flor, 2011; Worth & Nickels, 2011), prototypes of constructions are often built and put into a wind tunnel, such that the air flow and the effect of the structure's presence on air flow can be *measured* instead of *predicted* (e. g. Allegrini et al., 2013). The *input to be tested* in this example is the newly-conceived structure, the *outcome of interest* is the change of air flow and the *theoretical underpinnings* are constituted by the given formalization of aerodynamic laws.

Another goal of simulations is the prediction of traffic flow on highways (Treiber & Kesting, 2013). The referent system here consists of the highway itself and the driver-vehicle-units. The highway can be conceptualized as a two-dimensional space, with one dimension, the lanes, being discrete and small (usually 2 or 3 lanes, sometimes 4 or more) and the other dimension being continuous and large (hundreds of kilometers), with several exits and certain legal speed limits. The driver-vehicle-units are entities that continuously move along the large dimension of the highway (in one direction only), sometimes move along its small dimension (lane-switching), with certain technical speed limits and a more or less internally consistent lane-switching, tailgating and de-/acceleration strategy. The units' strategies are constrained by the fact that they can never pass a neighboring unit without switching lanes. One question of interest here is, what kinds of strategies lead to a slowing down of the overall traffic flow, or in extreme cases to traffic jams. The *input to be tested* in this example is the strategies and distribution of strategies in the population of driver-vehicle-units, the *outcome of interest* is the change to traffic flow. The *theoretical underpinning* is much simpler than in most models – as can be seen by the fact that one possible version of it fits completely into this paragraph (note that other versions are conceivable as well, including much more

complex ones).

In the traffic flow example, the referent system is understood very well and can be represented reasonably in a simulation. However, there is still a large number of degrees of freedom in advancing a well-known single-lane simulation model (Nagel & Schreckenberg, 1992) to include two or even more lanes (Knospe et al., 2002). It is also an important challenge to model a highway not as a closed system, but rather as a part of a network of roads with cars entering and leaving the highway or even as a part of the road, public transport and pedestrian traffic network of a whole country (Raney et al., 2002). One can see that even in a well-understood referent system, it is difficult to create an accurate simulation of reality. The example where the referent system includes turbulent air flow contains an even higher number of degrees of freedom and is thus even harder to capture in a simulation. This is why much effort is undertaken to build and operate wind tunnels (and even design new types of them, as in Devenport et al. (2013)), although there is certainly a large cost-saving and time-saving incentive in replacing them with good simulations. Simulations and other methods often have to go hand in hand to solve theoretical (Eames & Flor, 2011) or practical (Howell et al., 2010) problems.

1.4.4 Possible kinds of insight that can be gained from simulation

The examples from Section 1.4.3 have illustrated that, generally speaking, the number of degrees of freedom in the referent system translates to difficulty in building and interpreting simulations. In sound change research, and generally in linguistics, a discipline at the crossroads of – importantly, but not only – social sciences, neuroscience, and humanities, we have a myriad of variables, many of which we know very little about. This brings up the question of what exactly one might conclude from running simulated experiments when the number of degrees of freedom is very high.

Produce actual predictions about the future

The most direct kind of insight one can hope to get from building a simulation model is how a conceived *input to be tested* would behave in reality: starting with a computer model of a structure, how does it change air flow? Starting with a formalization of a driving strategy on a highway, how does it impact traffic flow? Starting from a given (empirically found or completely made up) vowel space and (possibly) an extra vowel space of a contact language, how will the vowels change over the decades?

In order for a simulation to produce credible results, one of two conditions has to be met. Perhaps the theoretical underpinning is so plausible and comprehensive that there is little reason left to doubt the correctness of the simulations' outcome (of course, empirical validation is still desirable). It is generally safe to assume that any given referent system in linguistics is way too complex for this condition to be met. Alternatively, the

outcome of a particular simulation model has been shown to be within a small margin of empirical results so many times (i. e. with so many different *inputs to be tested*) that it is plausible to extrapolate and believe that the outcome will be within equally small margins of the underlying truth/referent system with new input conditions as well. This is known as the *validation* of a simulation model (cf. Cioffi-Revilla, 2017a).

A proper validation, however, would require so many laboratory studies (real time and/or apparent time) that to attain it seems completely out of reach. The situation would be easier for synchronic research questions (because in these cases, data collection is inherently easier), but even then validation will suffer from the fairly small data sets linguists deal with. Basically, this kind of insight cannot be achieved as long as the referent system includes anything as poorly understood as the human brain.[5]

Think a theory through

A variant of the first type is when you assume a given theory as correct – no matter how good its empirical support is, there may even be none at all – and want to "think through" (Gilbert & Abbott, 2005, p. 859) what the theory's consequences would be. This can be interesting for exploring one or more theories where the procedures taking place in the referent system have been formalized but the results are not readily apparent. Implementing a simulation model here can aid in attaining those results. While it is true that a computer can only produce what it has been programmed to produce, this does not mean that the results cannot be surprising. de Boer (2006) rightfully says that this is a misunderstanding, noting that a system's behavior, even when specified completely, can be "so difficult to predict that the results of simulating it are often very surprising" (p. 387). This also shows that randomness in a simulation model is not strictly necessary for the model to produce useful results; that is, results that we would not know without running the model.

For this type of insight, validation of the simulation model and underlying theory is not important, because the insight is of the type "if theory X is correct, a particular outcome Y follows".[6]

Falsify hypotheses (and compare theories)

A different type of insight a simulation can yield is to falsify hypotheses about the referent system. Once researchers have fully specified their hypothesis (in the form of a simulation model), they can test if their model is capable of generating an expected macrostructure. The expectation may come from sound changes empirically observed using methods other

[5] This is not meant to downplay the fantastic insights both biologists and linguists have attained, but rather to highlight the human brain's complexity and the mystery that it remains to this day in spite of all efforts.

[6] It is still important to make sure the model correctly implements the theory, but that is called *verification* of the simulation model, not *validation*.

than simulation (e. g. apparent time, real time, comparative method). If the model cannot produce the expected result, the hypothesis must be re-tuned – or even rejected. If it does produce the expected result, this is evidence that the model is valid; and that the underlying theory is capable of explaining the phenomenon at hand. It does not, however, inform us about whether the underlying theory is more or less valid than any other competing theory, unless the competing theory is simulated with equal rigor and success.

Simulations have a real forte here. Hypotheses about driving strategies, for example, could hardly ever be falsified in non-simulated experiments, because while researchers can observe highway traffic, it would most likely be resource-prohibitive to experimentally manipulate the strategies of a *significantly large* number of drivers.

There is one caveat, though, which is very well captured in one of Epstein's (2006) footnotes: "For expository purposes, I write as though a macrostructure is either generated or not. In practice, it will generally be a question of degree" (p. 9). Moreover, sound change researchers do not usually find themselves in the luxurious situation of being able to rely on many empirical studies about the same phenomenon. Oftentimes, there is only one data set to draw expectations from. In case of a mismatch between that data set and the simulation, it may be hard to tell which of them is closer to the truth.

Generate hypotheses

For a fourth type of insight we might get from simulation models, we need to exploit the *hypothesis-generating* capacity of *parameter sweeps*. A parameter sweep is a continuous "evaluat[ion of] the model as a single parameter changes in values while others are held constant" (Cioffi-Revilla, 2017a, p. 389), and this process can "reveal special properties within a range, such as singularities, asymptotic behaviors, oscillations, or other quantitative and qualitative patterns" (ibid.). Typically, this is used to *verify* the model, that is, to make sure that it was indeed implemented to the specification of the theory and that it is free of bugs.

However, it is also theoretically valid to use this method to sweep through the conceivable range of values for one or more parameters in a theoretically unmotivated way and see if any unpredicted but interesting changes to "quantitative and qualitative patterns" can be observed.[7] This approach allows to generate hypotheses within the given theoretical framework that can later be tested.

One must be cautious here, though. By tweaking the parameters (and possibly even procedures) of the model, researchers can – by definition – find any result they wish. This is only constrained by the available computation time, that is, how many combinations of values a researcher can try. In many cases, it will be possible to tweak the parameters until there is an outcome consistent with our theory or previous experimental findings.

[7]Note that manipulating many parameters may result in such a large number of combinations of values that it becomes impossible to find enough computation time even with very expensive machines.

This may be described as "overfitting on purpose", that is, creating a model that explains one data set very well but fails for all other data sets, which is usually a bad thing. It can, however, be theoretically valid and useful to do exactly that. It is crucial, then, that the results be regarded as *new hypotheses* that must be tested, and that the material for testing them be *independent* from material used for the parameter sweep. This requirement is a given whenever results are interpreted *after* a parameter sweep. In fact, it is always a given when interpreting a simulation's result, as long as empirical validation of said model is scarce. Note that hypotheses can be generated not only by sweeping through *parameters* in the sense of Section 1.4.2, but also by sweeping through (observed or made-up) input data or by varying the procedural details of how speech perception and production are implemented in the agents.

1.5 Wrap-up: apparent time, real time and agent-based simulation

We have seen in the previous sections that simulations can be a powerful tool in a wide range of disciplines, including phonetics. We must remember, however, that an accurate simulation of reality can only exist when the referent system is already well-understood. If it is not, simulations can still be a good tool for attaining a good understanding of the referent system; but it is then more promising to aim to generate hypotheses rather than test them. It therefore becomes necessary to rephrase the definition I gave earlier in this chapter (Section 1.4.1): simulation is a method of *working with hypotheses* when real experiments are not feasible; not necessarily and certainly not exclusively one of *testing hypotheses*.

Now, how do simulated experiments relate to "real experiments?" Let us briefly revisit the differences between the apparent time and the real time paradigm discussed in Section 1.3: In the case where data are collected specifically for the research question, the apparent time paradigm can decouple the time needed to conduct the study from the length of the period the researchers wish to observe. In the case where only unspecific real time data are available, a specific collection with the apparent time paradigm allows for experimental control of numerous variables. The downside of the apparent time paradigm is that it only measures an abstraction of the actual sound change, whereas the real time paradigm measures the change more directly.

I argue that simulation extends all these differences between real and apparent time: it has the potential of enabling even more experimental manipulation (i. e. control for a larger number of variables by running simulations with a wide range of parameters), in even less time (the existing input data can be re-used for many simulations instead of collecting additional data), but requiring even more validation (the results will generally not be credible; they will instead constitute new hypotheses). One way agent-based simulation might be used successfully in our field, much like presented in Section 1.4.4 under

the headlines "thinking a theory through" and "generating hypotheses," is to manipulate experimentally many potentially interesting variables and then, in non-simulated follow-up studies, focus on those variables that turned out to produce interesting results in the simulation. Those variables might include possible social and language-internal triggers of sound change as well as formalizations of mechanisms of change spreading through a community.

About 30 years ago, Bailey et al. (1991) stated that the apparent time paradigm had been a "basic analytical construct in quantitative sociolinguistics" (p. 241) for nearly 30 years. This has remained the same up until now and will most likely remain the same for the foreseeable future. However, simulation might potentially be used more often alongside apparent time in the years to come.

I will conclude this introduction with an outlook on the following chapters. This dissertation reports on a combination of research projects at the crossroads of theoretical, empirical and methodological work in phonetics. Chapter 2 is a methodological piece of work about software that is used to organize and deal with digital research data, the *emuDB Manager*, which is part of the *EMU Speech Database Management System*. Chapters 3 and 4 represent empirical work in the field of laboratory phonology. They share the use of the apparent time approach to analyze the stability of phonological quantity across generations. However, while Chapter 3 reports on a quantity pattern in Finnish that is found to be stable across generations, Chapter 4 reports on a quantity pattern in Central Bavarian German that is found to have changed between generations. Chapter 5, then, has both a theoretical and a methodological focus. In a study about *agent-based modeling*, I aim to test the validity of this research method (using the data from Chapter 4), while at the same time applying the method – in a manner that is already known to be valid – to inform future theoretical research on sound change. The general discussion in Chapter 6 will then wrap up the findings from the previous chapters.

2 emuDB Manager: Cloud Hosting, Team Collaboration, Automatic Revision Control

Abstract

In this paper, we introduce a new component of the EMU Speech Database Management System (Winkelmann, 2015; Winkelmann et al., 2017) to improve the team workflow of handling production data (both acoustic and physiological) in phonetics and the speech sciences. It is named *emuDB Manager*, and it facilitates the coordination of team efforts, possibly distributed over several nations, by introducing automatic revision control (based on Git), cloud hosting (in private clouds provided by the researchers themselves or a third party), by keeping track of which parts of the database have already been edited (and by whom), and by centrally collecting and making searchable the notes made during the edit process.

This chapter also appeared as:

Jochim, M. (2017). Extending the EMU Speech Database Management System: Cloud Hosting, Team Collaboration, Automatic Revision Control. *Proceedings of Interspeech 2017, Stockholm, Sweden*, 813–814. https://doi.org/10.21437/Interspeech.2017

2.1 Introduction

The software tool presented in this paper, the *emuDB Manager*, facilitates important steps in the research workflow of handling speech production data. Managing speech production data involves the collection, segmentation, annotation, and analysis of speech. These steps require many hours of manual labor and result in copious amounts of primary as well as secondary data files. To aid researchers – and especially teams of researchers – with their task, emuDB Manager allows them to systematically share their files on a server, keep track of different versions, keep track of their notes, and coordinate which parts of the data have already been edited (and by whom) and which parts remain to be edited (and by whom).

2.2 EMU Speech Database Management System

The EMU Speech Database Management System (in short EMU-SDMS) mainly comprises a web application to visualize, annotate and segment data, and two R packages to analyze data, with a sophisticated query language to allow for the analysis of database subsets based on their hierarchical and sequential annotation. The latest developments have been described in Winkelmann et al. (2017). The system's website is found at https://ips-lmu.github.io/EMU.html.

The emuDB Manager adds to this a server component for centrally storing data and assigning tasks to project members; and a web front end to control these functions and to upload/download data as necessary. The server component extends an already existing reference implementation of the EMU-webApp protocol.

Researchers wishing to host an "EMU cloud" at their own institution need to run a web server and a nodeJS server. The source code of all components is found on GitHub (https://github.com/ips-lmu).

2.3 Cloud Hosting

Cloud hosting basically means to store data (primary and secondary) on a central server, such that all team members can access and modify them (after being authenticated through their password). Cloud hosting can greatly facilitate collaboration in teams, no matter if they are located at a single university or spread over research institutions world-wide. The key advantage is that team members do not need to store portions of the data on their own computers, edit (segment and annotate) them in parallel and then put the individual portions back together. Editing is done directly on the server-stored data. This avoids much confusion about which copy is the most recent one or the "master," since the server ("the cloud") holds all modifications.

The emuDB Manager web front end is used to upload data onto the server, where all team members can see them. The EMU-webApp can then be used to visualize and to

edit the data. Once the edit process is finished or has progressed sufficiently to start analysis, the edited data can be downloaded again. The version being downloaded can also be given a label, which is especially handy for intermediary analyses (in the middle of the edit process). By referring to the labels, the team can always be sure which version of the data has been analyzed.

Data protection is a very important aspect of cloud hosting. It is in the interest of both researchers and participants that the data not become public before the researchers decide to publish them in a way that upholds the participants' anonymity and other legal rights. The emuDB Manager achieves this by requiring users to authenticate with their password before they gain access to any data. The project leader can choose accounts that are entitled to access and modify the project data. For this reason, the project leader can create new accounts or utilize ones provided by research institutions (using Shibboleth), e. g. via the CLARIN Service Provider Federation (https://www.clarin.eu/content/service-provider-federation). This is useful for teams comprising multiple universities, since each member can use the account they already possess.

Data protection also extends to the question where the data is physically stored, and that can be on a server located at an affiliated research institution (as is the case for the author's institute) or on a commercially rented server.

2.4 Automatic Revision Control

Over the course of a project, data is added and edited by several persons. The first analysis steps are often started before the edit process has been completed, and sometimes analysis reveals that more collection is due or that annotation guidelines need changing. This makes it necessary to consistently distinguish different versions of the whole data set and assign labels to these versions. The software Git provides a very stable mechanism for this. The emuDB Manager uses Git to store a snapshot of the data whenever data have been added or edited, and when work has been (re-)assigned to collaborators. This way, no changes are left undocumented. It is transparent at any time which changes have been made, when, and by whom. Any set of changes (called a *commit*) down to the resegmentation of an individual recording can be traced to its editor and time, and if need be, it can be reversed.

2.5 Team Collaboration

Coordination of these tasks becomes much more difficult as the number of collaborators and assistants grows, and also as the amount of data grows. emuDB Manager provides two tools for efficient coordination. Research teams can decide whether to grant full access to everybody including assistants, or restrict the assistants' access to those parts

specifically assigned to them.

2.5.1 Whose work is it?

The first tool is that team members can assign their colleagues (or self-assign) parts of the database for editing. This is done using *bundle lists*.[1] To create a bundle list, researchers select a part of the database using regular expressions (or the whole database) and assign them to one or more team members. When team members log in for editing, they will only see the parts currently assigned to them.

All bundle lists are gathered in a clear overview, indicating the percentage to which they have been finished. This enables users to see at a glance who is currently editing which files and how far work has progressed.

Once a bundle list has been finished, it is given an archive label. It will still be shown in the overview, but clearly marked as "finished work." The archive label can be any text freely chosen by the team members. While the software imposes no restrictions, useful choices may include the date when the bundle list was finished, or a name for the portion of work or editing stage that the bundle list represents.

2.5.2 Collecting Notes

Many times, especially while segmenting and annotating, researchers take notes about particular recordings. emuDB Manager allows researchers to store these notes centrally, and makes it possible for every team member to search these notes.

Notes can be entered in EMU-webApp, the same interface that is used for editing and visualizing data. Notes will be stored along with the bundle lists. In emuDB Manager, researchers can read and search these comments. It is also possible to restrict the view to "commented bundles only," and then use EMU-webApp for visualizing the respective bundle data (spectrogram, segmentation, formants etc.), along with the comments.

This is very helpful especially for the communication between project collaborators and student assistants, when discussing the annotation of individual recordings.

2.6 Discussion

emuDB Manager is an extension of the EMU Speech Database Management System (EMU-SDMS). EMU-SDMS "sets out to be as close to an all-in-one solution for generating, manipulating, querying, analyzing and managing speech databases as possible" (Winkelmann et al., 2017, p. 393). This new extension adds several features that are especially valuable when working in teams: It provides speech scientists with an easy-to-use interface to established (but often complicated-to-use) techniques such as

[1] In EMU-SDMS, a recording and its accompanying secondary files (e. g. annotation, derived signals) are termed a "bundle."

automatic revision control or server storage; further, it also facilitates usage of advanced features of the EMU-SDMS, such as bundle lists.

The most interesting next steps are concerned with further exploiting cloud hosting and cloud computing, as well as with inter-labeler agreement. As to inter-labeler agreement, the emuDB Manager already forms a useful tool for multiple editors to control each other's work. A useful extension to this would be to automatically evaluate inter-labeler agreement when multiple editors work on the same recordings independently of each other.

As to cloud functionality, we are considering two aspects. Since data analysis is carried out with the R package emuR, one of the two is to exploit RStudio's server version to bring the analysis step into the cloud. Currently, it is necessary to download the data from emuDB Manager in order to perform analyses, once the edit process is finished. RStudio in the browser would therefore yield two advantages: Server-side processing power could be utilized for computation-heavy statistical analyses (cloud computing); and the need for the error-prone task of repeatedly copying large data sets between machines would be further reduced.

The second aspect is who provides the cloud resources. While each research institution could use the software (which is freely available under an open source license, see https://github.com/ips-lmu) and run a cloud service for its own members, it might be more efficient if laboratories shared their resources. The CLARIN Center BAS (https://www.bas.uni-muenchen.de/Bas/), hosted by the University of Munich, already provides speech processing tools to the research community (e. g. WebMAUS for automatic segmentation) and is considering to include the emuDB Manager in its cloud services. It may also be worthwhile to integrate EMU-SDMS with infrastructures such as the Open Science Framework (https://www.osf.io/), or with commercial cloud service providers.

The emuDB Manager is already proving a valuable tool for our international team spread over three research institutions and we hope that in the future, other groups will be able to share the same advantages.

2.7 Acknowledgments

This work was supported by the DFG-DACH grant number KL 2697/1-1 "Typology of Vowel and Consonant Quantity in Southern German varieties: acoustic, perception, and articulatory analyses of adult and child speakers" awarded to F. Kleber.

3 What do Finnish and Central Bavarian have in common? Towards an acoustically based quantity typology

Abstract

The aim of this study was to investigate vowel and consonant quantity in Finnish, a typical quantity language, and to set up a reference corpus for a large-scale project studying the diachronic development of quantity contrasts in German varieties. Although German is not considered a quantity language, both tense and lax vowels and voiced and voiceless stops are differentiated by vowel and closure duration, respectively. The role of these cues, however, has undergone different diachronic changes in various German varieties. To understand the conditions for such prosodic changes, the present study investigates the stability of quantity relations in an undisputed quantity language. To this end, recordings of words differing in vowel and stop length were obtained from seven older and six younger L1 Finnish speakers, both in a normal and a loud voice. We then measured vowel and stop duration and calculated the vowel to vowel-plus-consonant ratio (a measure known to differentiate German VC sequences) as well as the geminate-to-singleton ratio. Results show stability across age groups but variability across speech styles. Moreover, VC ratios were similar for Finnish and Bavarian German speakers. We discuss our findings against the background of a typology of vowel and consonant quantity.

This chapter also appeared as:

Jochim, M., & Kleber, F. (2017). What do Finnish and Central Bavarian have in common? Towards an acoustically based quantity typology. *Proceedings of Interspeech 2017, Stockholm, Sweden*, 3018–3022. https://doi.org/10.21437/Interspeech.2017-1285

3.1 Introduction

The aim of this study was to corroborate acoustically a typology of quantity usage in different languages as part of a large-scale project studying the diachronic development of quantity contrasts in German varieties. With German varieties using both durational and non-durational cues to mark the vowel length and the so-called voicing contrast, we chose Finnish – undisputedly a quantity language – as a solid basis for comparison. Beyond the Finnish data already available (Doty et al., 2007; Engstrand & Krull, 1994; Lehtonen, 1970; Suomi et al., 2013), we need to further establish the phonetic detail pertaining to its quantity contrasts. For our typological aim, we need to be able to compare the strength of the durational cues in German phonological systems not only to the other cues inside the same systems, but also to the strength of durational cues in quantity-heavy phonological systems like Finnish. Additionally, since the use of durational cues in German varieties has changed diachronically, both on a historic timescale (from Old High German to Modern High German) (Seiler, 2005) and within recent generations (Kleber, 2017a; Moosmüller & Brandstätter, 2014), we also investigated the stability of the relevant durational cues in Finnish across generations and under the influence of system-internal variation (here different speech styles).

Bannert (1976) has proposed a typology that classifies languages based on where they allow quantity contrasts: in vowels only (e. g. Czech); in consonants only (e. g. Italian); in both vowels and consonants, independently of each other (e. g. Finnish (Bannert, 1976; Suomi et al., 2013)); in both vowels and consonants, but inter-dependently (e. g. Central Bavarian (Bannert, 1976)); or not at all.

Many languages, including those under investigation in this paper, have two quantities: long vs. short. In Finnish, this leads to four possible types of vowel-consonant (VC) sequences: VC, V**ː**C, VC**ː**, and V**ː**C**ː** (here and hereafter **ː** indicates phonologically long vowels and consonants, respectively). Central Bavarian employs complementary length with long vowels always preceding lenis (i. e. short) consonants and short vowels only fortis (i. e. long) consonants. That is, in this variety only two types are possible: V**ː**C and VC**ː**. However, there is evidence that it has started to allow a vowel length contrast before fortis stops (Kleber, 2017a; Moosmüller & Brandstätter, 2014), presumably influenced by Standard German.

While Finnish and Central Bavarian are part of the typology proposed in Bannert (1976), Standard German is not. Standard German poses a challenge for the quantity typology, because it uses durational cues for both vowel and consonant contrasts (independently, like Finnish), but it also uses non-durational cues to support them. In particular, vowel contrasts are cued by duration and by quality (Wiese, 1996) (note though that there is one vowel pair, /a aː/ that is distinguished solely by duration). For stops, German has a two-way contrast that is variously termed fortis/lenis or voicing contrast (see Braun (1988) for a discussion). Its main cue is aspiration, but in the absence of aspiration (e. g. before nasals as in [beːt̪n̩] ‘to pray’), the most important cue becomes relative duration of the stop’s closure phase and the preceding vowel (Kohler,

1979). In the remainder of the paper we refer to this cue as proportional vowel duration (PVD). Since the described vowel and consonant contrasts are (1) to some extent quantity contrasts and (2) freely combinable, it therefore seems plausible to take the same four types of VC sequences as described above for Finnish as a basis for Standard German VC sequences.

PVD has been shown to separate V:C: and V:C sequences (e. g. [bo:tn̩] 'messengers' vs. [bo:dn̩] 'floor', Kleber, 2014; Kohler, 1977) as well as V:C: and VC: sequences (e. g. [bi:tn̩] 'to offer' vs. [bɪtn̩] 'to request'], Kleber, 2017a) in different varieties of Standard German. The present study asks whether PVD is a good measure (1) to demonstrate the phonemic four-way length contrast in Finnish VC sequences and (2) for an acoustically based quantity typology as suggested by Bannert (1976). Table 3.1 gives a first impression of how similar PVD values (recalculated from previous studies) are across languages.

Table 3.1: Proportional vowel duration (PVD) in VC sequences; n/a refers to missing data for a particular sequence in the respective analysis. Vowel duration is either given as a proportion of the vowel+closure sequence (German, Central Bavarian upper row) or of the vowel+stop sequence (Finnish, Central Bavarian lower row).

Language (Source)	**V:C**	**V:C:**	**VC**	**VC:**
German (Kohler, 1979)	0.76	0.58	n/a	n/a
German (Braunschweiler, 1997)	0.75	0.68	0.61	0.50
Finnish (Lehtonen, 1970)	0.68	0.47	0.45	0.35
Central Bavarian (Kleber, 2014, 2017a)	0.76	0.56	n/a	0.31
	0.69	0.44	n/a	0.21

Three major questions arise from table 3.1. (a) German and Finnish appear to implement the four types differently in terms of PVD. Is this due to the different usage of non-durational cues in these two languages, and should they therefore be treated differently (i. e. due to the different phonetic implementation) or the same (because of similar kinds of phonemic categories) typology-wise? (b) If Bavarian is developing a third category V:C: (which may be governed by dialect leveling with Standard German but is certainly not governed by an assimilation to Finnish), will it adopt Standard German's phonetic implementation (as the values in the upper row suggest), or will it be more similar to Finnish (as the values in the lower row give reason to expect, and perhaps for the same typological reasons that set Finnish and German apart)? There are two differently calculated sets of PVD values for Bavarian: One includes aspiration in the calculation and the other does not. The difference between the two sets clearly demonstrates that an acoustically based quantity typology needs to consider both lower-level units such as phones (i. e. the closure phase) and higher-level units such as phonemes (i. e. the entire stop). And (c), since VC and V:C: are very close in Finnish but further

apart in German, can this measure be used to separate all four categories in a language like Finnish? Such a separation depends largely on the dispersion of a given data set, but from Lehtonen (1970) we only know the mean. As a first step towards an acoustically based quantity typology, we will therefore focus on question (c) in the present study.

Thus, our first research question is whether the four Finnish quantity categories can be separated by means of PVD and how this measure performs in relation to absolute duration and geminate to singleton ratio that have been investigated in previous studies (Doty et al., 2007; Engstrand & Krull, 1994; Lehtonen, 1970; Suomi et al., 2013). In order to compare (in future studies) the outcome of the present study to ongoing diachronic developments in Germanic languages, our second research question is how stable the observed patterns remain under the influence of system-internal variation. We chose to test the difference between younger and older speakers. We do not, however, expect any substantial age differences in Finnish, since we are not aware of any instability reports regarding the language. Moreover, as a within-speaker type of variation, we chose to test the difference between normal and loud speech. Differences in loudness are known to correlate with speech rate (with louder speech being slower than normal speech (Dromey & Ramig, 1998); we did not vary rate directly to allow future comparison with children's data (Kleber, 2017b)).

3.2 Method

3.2.1 Material

We analyzed 13 words (table 3.2) of a 45-word corpus. All but two of the corpus words, and all of the analyzed words, were structured $C_1V_1C_2V_2$. Within the 13 target words, V_1, C_2, and V_2 were either short or long, and both C_1 and C_2 were stops.

Special emphasis has been put on the words *taka* and *taakka*, which form the only minimal pair in our corpus that contrasts VC and V**ː**C**ː**; and on the words *kota*/*koota*/*tutti*/*tuutti*, which contrast all four types of VC while preserving the identity of C_2 and the quantity of V_2. Moreover, all V_1 in these four words are high/mid-high back vowels and show strong overlap in their formant frequencies F1 and F2.

3.2.2 Participants and Recording Procedure

13 native speakers of Finnish took part in the experiment (9 female, 4 male). They were assigned to one of two age groups: younger (born 1995-1997) and older (six born 1950-1962, one born 1971). The recordings were made in 2016. The apparent-time design was used to test the (in-)stability of the cues involved. Moreover, the young group allows for a real-time comparison with data of then-young speakers described in Lehtonen (1970). Ten participants lived in the region of Uusimaa (located in South Finland and including Helsinki) at the time of recording. The other three had also lived there, but had moved

Table 3.2: Target words analyzed in the current study.

Finnish	**English**
kota	'capsule'
koota	'put together, collect'
tutti	'pacifier'
tuutti	'cone'
taka	'back, rear or hind (prefix)'
taakka	'burden'
takka	'chimney'
kaato	'bull's eye'
katto	'roof'
kiitää	'to race'
kiittää	'to thank'
tapaa	'to meet sb.'
tappaa	'to kill'

to Munich, Germany, within two years before the recordings were made. Participants were paid.

The speakers were recorded at their own homes, using a laptop computer, mobile recording equipment (BeyerDynamic headset microphone, M-Audio audio interface) and SpeechRecorder (Draxler & Jänsch, 2004) (version 3.4.2). The digital audio signals were sampled at 44.1 kHz, with a 16-bit resolution.

Each of the 45 target words was embedded into the carrier sentence *Sano X yhden kerran* 'say X once'. Six repetitions of each sentence were presented one at a time and in randomized order on the laptop screen. The recording sessions were divided into six blocks, each consisting of all 45 words. The participants were asked to read the sentences in a normal voice in blocks 1, 3, and 5, and in a loud voice in blocks 2, 4, and 6.

3.2.3 Analysis

The recordings were automatically segmented using WebMAUS (Kisler et al., 2016). Segment boundaries were then corrected manually where necessary. Because WebMAUS has not yet incorporated Finnish training data, we used its language-agnostic mode.

All manual corrections and the analysis were conducted using the EMU Speech Database Management System (Winkelmann et al., 2017) (version 0.2.1) and R (R Core Team, 2016) (version 3.3.2).

The dependent variables we investigated were the absolute duration of V_1 and C_2, the respective proportional vowel duration (PVD), defined as $\frac{V}{V+C}$ (like Lehtonen (1970) we included aspiration in the consonant duration to allow for direct comparison of all

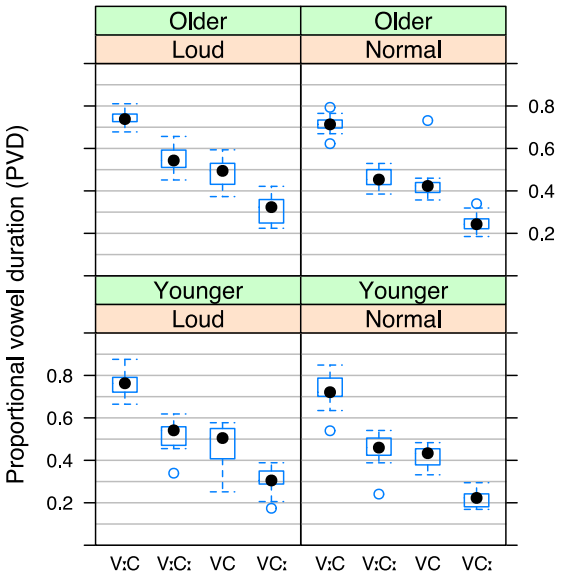

Figure 3.1: PVD for kota/koota/tutti/tuutti tokens. N per boxplot is 18 (younger), 21 (older); overall N = 312.

Finnish data available and based on the assumption that aspiration only plays a marginal role in Finnish), and the ratio of long vs. short segments (V_1 ratio: $\frac{V:}{V}$, C_2 ratio: $\frac{C:}{C}$).

Our independent variables were age (younger/older), speech style (normal/loud), and category (V:C/V:C:/VC/VC:). For some tests, category was reduced to two factors V_1 and C_2 quantity (long/short). All factors except age were varied within-subjects.

3.3 Results

3.3.1 Proportional Vowel Duration (PVD)

Commensurate with fig. 3.1, a repeated measures ANOVA with PVD as the dependent variable revealed significant main effects for category ($F[3, 33] = 435.5, p < 0.001$) and speech style ($F[1, 11] = 30.6, p < 0.001$), as well as a significant interaction effect for category × speech style ($F[3, 33] = 5.2, p < 0.01$). To prevent any potential effects of vowel height or position on PVD, the analysis was run on o/u-word tokens only.

In order to specifically test VC against V:C:, and again to ensure best comparability, we ran another analysis on the tokens of *taka* and *taakka*. Commensurate with fig. 3.2, the ANOVA revealed main effects for category ($F[1, 11] = 25.9, p < 0.001$) and speech style ($F[1, 11] = 42.9, p < 0.001$), but no statistically significant interactions.

These findings suggest a difference between the categories VC and V:C: that is subtle,

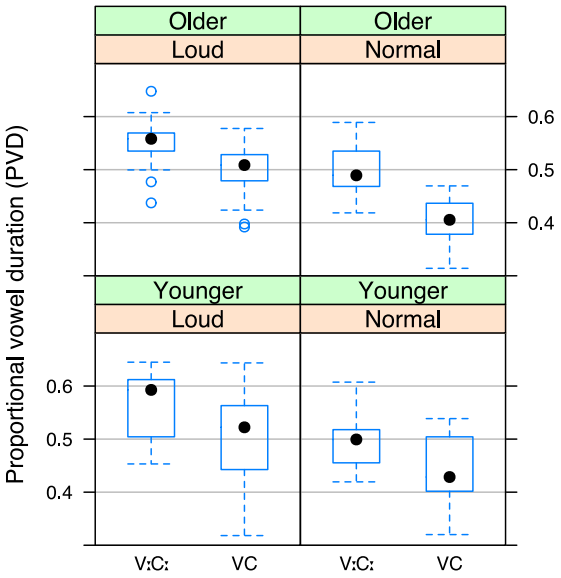

Figure 3.2: PVD for taka/taakka tokens. N per boxplot is 18 (younger), 21 (older); overall N = 156.

yet robust and statistically significant. The difference in mean (fig. 3.1 and 3.2 show the median) between *taka* and *taakka* is between 5 and 6 % for loud speech (younger and older) and for the younger speakers' normal speech, and about 9 % for the older speakers' normal speech.[1] While the younger speakers show substantial overlap between the two categories, the older speakers show very little.

In general, PVD appears to increase in loud speech in all four types of VC sequences, but it does so to the same extent in all four categories.

The two endpoint categories in normal speech show PVD means of 22 % (younger, VC:), 25 % (older, VC:), 73 % (younger, V:C), and 71 % (older, V:C), respectively. For V:C, this is similar to Lehtonen's data, but for VC:, it differs substantially (see table 3.1).

3.3.2 Absolute V_1 and C_2 duration

The absolute durations of V_1 and C_2 are shown in a scatter plot in fig. 3.3. We observe four clearly separated clusters, one for each type of VC sequence. The overlap between them is remarkably small. However, it increases in loud speech. If the durations were

[1] We also calculated the PVD measure with two types of logarithmic transformations, defining it as $\frac{ln(V)}{ln(V)+ln(C)}$ or $\frac{ln(V)}{ln(V+C)}$, respectively. With neither of them did the degree of separation between the two categories diminish.

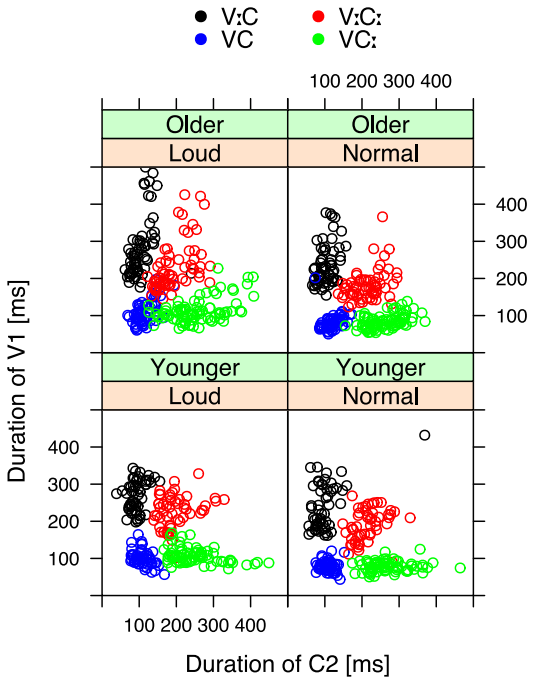

Figure 3.3: Absolute duration of V_1 and C_2. Included are tokens of all 13 target words (table 3.2), separated by speech style and age group. The colors encode the kind of VC sequence the respective token appears in. Overall N = 1,010.

completely independent of each other, we would expect the four clusters to form a rectangle along the two dimensions. This, however, is not the case. While V:C and VC: sequences show greater dispersion along the dimensions of vowel and consonant duration, respectively, V:C: and VC sequences vary along both dimensions although they differ greatly in the degree of dispersion. These category-dependent distributions suggest that (1) variation is greater in long than in short phonemes (see Mooshammer and Geng (2008) for similar results in German), (2) the two vowel categories overlap to a greater extent when preceding long as opposed to short consonants, and (3) the overlap between the two consonant categories is not affected by V_1 quantity. This observation is in line with previous accounts of a language-independent tendency of vowels being influenced by adjacent consonants, but not the other way round (Braunschweiler, 1997; Chen, 1970). In section 3.3.3 we will evaluate this observation numerically.

3.3.3 Geminate to Singleton ratio (GSR)

We calculated the geminate to singleton ratio for both V_1 and C_2, as a function of C_2 or V_1 quantity, respectively. The GSR in V_1 is higher before short C_2 than before long C_2 (mean: 2.87 before short, 2.22 before long C_2); this effect turned out to be statistically

significant ($F[1, 11] = 18.8, p < 0.01$). Not only the mean values, but also the dispersion of values in fig. 3.3 point in this direction. On the other hand, and again commensurate with fig. 3.3, GSR in C_2 did not differ significantly between short V_1 and long V_1 tokens (2.33 after short, 2.26 after long V_1).

3.4 Discussion

This study comprised two main aims: Firstly, to test whether PVD is a useful acoustic measure for an acoustically based quantity typology, and secondly, to establish the stability of duration cues used for signaling phonemic vowel and consonant quantity in Finnish. The three main findings were as follows:

(1) PVD is able to separate all four types of VC sequences in Finnish. This suggests that it may be a useful acoustic correlate for the typology. In comparison with absolute duration, the main advantage of PVD is that it constitutes a uni-dimensional measure for all four types. Absolute durations, while providing perhaps a better separation of the four categories (cf. fig. 3.3 vs. fig. 3.1 and 3.2), need two dimensions to achieve the same.

(2) The category separation provided by absolute durations is slightly reduced in loud speech. This appears not to be the case for PVD. This suggests that PVD is slightly more robust as a cue in terms of normalization across speech styles/rates (see Pickett et al. (1999) for similar results in Italian), which would seem plausible because the measure itself may integrate normalization for rate. It would be very interesting to specifically test the perceptual relevance of PVD in Finnish, especially in light of Kohler's (1979) finding that PVD is a strong perceptual cue in German.

(3) As expected, we did not find any substantial differences between the two age groups. This suggests that the acoustic basis of Finnish quantity contrasts, namely duration, has not changed within recent generations.

One problem regarding PVD, however, remains: Why would the vowel proportion in VC be smaller than in V:C:? This appears to be so in both German and Finnish and it suggests that the difference between long and short is stronger for vowels than for consonants. This might be explained in terms of the non-durational cues employed in the respective contrasts. Finnish is often regarded not to use cues such as aspiration or vowel quality to distinguish its quantity contrasts – neither in vowels nor in consonants, which would make it likely for vowel and consonant lengthening to be the same. Doty et al. (2007), however, did investigate and find some additional cues for the stop length contrast. This could explain the bias towards more vowel lengthening and thus a higher vowel proportion in V:C:. In German, non-durational cues are known to play an important role in both vowels and consonants. However, Braunschweiler (1997) only investigated the /a a:/ contrast, where vowel quality plays a minor role, making duration especially important for vowels. This could explain why in those data, PVD is particularly high for V:C:.

Finally, how does Central Bavarian – the German variety that motivated the current study – fit in the pattern? Depending on the exact definition of PVD, the Bavarian PVD values in table 3.1 are either closer to the Finnish or the German PVD values: When PVD marks the vocalic proportion of a vowel+closure sequence, the Bavarian temporal patterns of the three VC categories resemble more closely those of Standard German but such a measure leaves aside an important part of the stop (namely the aspiration phase) that may very well be a relevant factor in the auditory processing of the vowel and consonant length contrast (note that Kleber (2014, 2017a) did, unlike Kohler (1979), include words with oral releases). In fact, when PVD marks the vocalic proportion of a vowel+stop sequence then the temporal patterns of the three categories measured for Bavarian are closer to those found for Finnish. In particular, the VC**:** category – where Central Bavarian and Finnish according to Lehtonen (1970) diverge the most – in our Finnish data was much closer to the Bavarian values (our data yielded a mean of 22–25 % for Finnish VC**:**).

We are currently conducting further analyses of durational and non-durational cues in Central Bavarian and other German varieties to better understand the timing relations in VC sequences and their typological characteristics. Considering the entire stop in the PVD value might be the more appropriate measure for an acoustically based typology because it appears to better allow for generalization – both within (e. g. when comparing orally vs. nasally released stops) and across languages (e. g. when comparing languages that use aspiration with those that do not). After all, Standard German temporal patterns may also have something in common with Finnish temporal pattern when accounting for the entire stop in the PVD measure.

3.5 Acknowledgments and Data

This work was supported by the DFG-DACH grant number KL 2697/1-1 “Typology of Vowel and Consonant Quantity in Southern German varieties” awarded to F. Kleber.

An R script and a data frame containing the 1,010 observations this report is based on is permanently available as Jochim and Kleber (2017a).

4 Fast-speech-induced hypoarticulation does not affect the diachronic reversal of complementary length in Central Bavarian

A variant of this chapter has been submitted for publication as:

Jochim, M. & Kleber, F. Fast-speech-induced hypoarticulation does not considerably affect the diachronic reversal of complementary length in Central Bavarian.

4.1 Introduction

The focus of this paper is on the trigger and the implementation of a prosodic sound change currently in progress in the German dialect Western Central Bavarian (WCB). Recent work has shown that some dialect features are in the process of being dropped in favor of features that resemble Standard German more closely. This includes the phonetic implementation of stops and the phonotactics of word-medial vowel-plus-stop sequences. One aim is to investigate the potential role of dialect-internal factors that may also be involved in this externally fostered change. Another aim is to work out the specifics of how the change is progressing in the dialect, that is, to what extent the different cues are affected.

At the center of the sound change that we are interested in is the implementation of word-medial post-vocalic stops. Standard German has a phonemic length contrast in vowels (/miːtə/ 'rent' vs. /mɪtə/ 'center') and a phonemic fortis-lenis contrast in stops (/miːtən/ 'to rent' vs. /miːdən/ '(they) avoided') (Wiese, 1996). Fortis stops have longer closure phases (in word-medial position) and a higher voice onset time (VOT) than lenis stops (Jessen, 1998). Phonotactically, Standard German does not restrict the combination of vowels and stops: Both fortis and lenis stops can follow after either phonemically long or short vowels. That is, in addition to the examples above representing combinations of short vowel plus fortis stop, long vowel plus fortis stop, and long vowel plus lenis stop, respectively, Standard German also allows the combination of short vowel plus lenis stop as in /vɪdɐ/ ('ram') (adding up to a total of four possible combinations). On the other hand, Central Bavarian (CB), spoken in the south east of Germany (Western CB, hereafter WCB) and most parts of Austria (Eastern CB, hereafter ECB), puts a clear restriction on the combination of sounds: long vowels only occur before lenis stops and short vowels only before fortis stops. Most accounts claim for CB varieties that it is the stop contrast that is phonemic while vowel length is regarded allophonic, that is, predictable by the nature of the post-vocalic stop. As opposed to Standard German, the CB fortis–lenis contrast is a true length contrast (with fortis meaning long and lenis meaning short), which is why the phonotactic rule described above is often called one of complementary length. VOT plays much less of a role in the dialect (Bannert, 1976; Seiler, 2005; Wiesinger, 1990).

The rule of complementary length has been described in the literature for a long time (Bannert, 1976; Hinderling, 1980; Pfalz, 1913; Seiler, 2005; Wiesinger, 1990) (and has variously gone by the names of "Pfalz's law," "complementary length," or "(Central) Bavarian quantity relations"). Hinderling (1980) leaves no doubt that the rule is adhered to even when borrowing words from Standard German into the dialect. According to him, borrowers avoid illegal combinations by adjusting either the vowel or the consonant. This is done in free variation. That is, a word such as *Pudding* 'pudding' – a loan from Standard German,[1] where it is produced as /pʊdɪŋ/ – becomes either /puːdɪŋ/ or /pʊtɪŋ/

[1]Hinderling (1980) notes that *Pudding* is a loan from Standard German.

in Central Bavarian. But the same can also be observed in words native to the dialect lexicon, such as *Vater* 'father', /faːter/ in Standard German, which is produced in the dialect either as /fater/ (i. e. with a short vowel plus fortis stop) or as /fɔːder/ (i. e. with a long vowel plus lenis stop).

While the literature on CB was unanimous about the above-mentioned phonotactic restriction for a long time, recent work suggests that the co-dependency of vowel and following stop is likely on the retreat, giving way to two independent phonemic contrasts between vowel length and consonant strength (i. e. fortis vs. lenis), respectively. Based on auditory analyses in the framework of traditional linguistic fieldwork, Schikowski (2009, p. 44f.) notes that the co-dependency can be weakened in younger speakers, although adherence to the rule is still absolutely dominant. Moosmüller and Brandstätter (2014) presented acoustic evidence that a combination of long vowel plus fortis consonant does form part of ECB[2], in spite of traditional accounts. Two further studies presented experimental evidence for more pronounced dialectal traces in older compared to younger WCB speakers' production and perception of the Standard German vowel length (Kleber, 2017a) and fortis–lenis contrast (Kleber, 2018), respectively. More precisely, older but not younger WCB speakers use consonant duration in both modalities to cue the Standard German vowel length contrast before fortis stops in this regionally-accented standard register. Moreover, only older WCB speakers adopted this strategy to differentiate vowel length in the contexts of post-vocalic sonorants. Regarding the fortis–lenis contrast, in comparison with the older cohort of the same speech community, younger WCB speakers were less affected by the prevalent dialectal complementary length feature again when asked to operate in a standard register and in particular in perception. At the same time, these younger speakers relied to a greater extent on VOT. While all of these observations stem from sources that used very different methodologies and materials, they lend strong support for an acoustic apparent-time investigation of how WCB speakers from two different generations nowadays implement the postvocalic fortis–lenis contrast when operating in the dialect. The present study also fills the gap regarding the status of short vowel plus lenis stop combination (which to our knowledge has not been investigated before) by including this sequence in the acoustic analyses.

In investigating such vowel-plus-stop sequences, several phonetic studies (including those in the previous paragraph) have used combined measures like the vowel-to-closure duration ratio to describe the entire sequence. This has been claimed to be perceptually relevant (Kohler, 1979), more stable across speech styles (Pickett et al., 1999), or advantageous for typological reasons (Jochim & Kleber, 2017b). The present study, however, is largely built on the separate measure of closure duration in order to be able to track the change within the two speech sounds in greater detail.

Kleber (2017a) argued that dialect levelling, which is defined as a diachronic process during which regional varieties become more similar to the standard language (Kerswill, 2003; Trudgill, 1986) or a close dialect (Hinskens, 1998) as a result of language-external

[2]More precisely, the Viennese dialect.

factors such as changing community network structures and speaker mobility (Britain, 2010), accounts best for this observed apparent-time change. This process has also been related to a number of other sound changes currently in progress in Germany (cf. Harrington et al., 2012 and Bukmaier et al., 2014) and other European countries (see e. g. Kerswill, 2002 on British English). For example, the voicing effect by which vowels are shorter before fortis than before lenis stops and which is characteristic of Standard Anglo-English has been claimed to spread via dialect contact to Scottish-English regions where the so-called Scottish Vowel Length Rule (SVLR) had been operating before (Hewlett et al., 1999; Scobbie, 2005). For two reasons, this example is of particular interest in the context of the present study: first, the SVLR has a similar domain as the phonotactic restriction of Central Bavarian discussed above, in that it restricts long and short vowels, respectively, to certain phonological contexts. Second, Rathcke and Stuart-Smith (2016) presented data suggesting that in Glaswegian English the diachronic weakening of the SVLR was more likely to be related to the language-internal factor prosodic deaccentuation than to language contact.

The three main aims of the present paper are therefore to test (1) whether long vowel plus fortis stop sequences emerge in the WCB dialect (as shown for ECB in Moosmüller and Brandstätter (2014) and suggested by WCB speakers' usage of the regionally-accented standard register (Kleber, 2017a, 2018)); (2) whether this change extends to short vowel plus lenis stop sequences; and (3) whether such a change can also be related to language-internal factors (as was the case for Glaswegian English).

The language-internal factor chosen for the present investigation is that of speech rate-induced hypoarticulation which appears especially relevant to the collapse of duration-based contrasts and which appears in everyday speech. More precisely, hypoarticulation arising from fast speech constitutes a phonetic bias (Garrett & Johnson, 2013) able to trigger a change towards short vowels and lenis consonants. This builds on ideas from Kohler (1984) and Ohala's (1993a) model of listener errors leading to sound change. The model differentiates how listeners usually handle phonetic variance from an unusual, erroneous way that can give rise to sound change but is only observed infrequently. One source of phonetic variance is, for example, coarticulation, where a phonological property of one speech sound is physically present in another speech sound (Farnetani & Recasens, 2010). Usually, in Ohala's model, listeners will compensate for this displacement and thus be able to attribute the property to the speech sound it originated from, rather than the speech sound it physically appeared in. In infrequent situations, however, listeners fail to achieve this compensation: They wrongly attribute the property to the speech sound it physically appears in and may eventually adjust their mental representation of the respective speech sounds (as happens, for instance, in tonogenesis, see Kingston, 2011).

Transferring this model to duration-based contrasts, we aim to test whether fast-speech-induced hypoarticulation – or to be more exact: failing compensation for it – can be regarded a trigger mechanism of the sound change in question. Fast-speech-induced hypoarticulation is ubiquitous in everyday speech (Lindblom, 1990) and listeners are usu-

ally very well able to compensate for fast speech (Reinisch, 2016). However, fast speech is a particular peril to phonologically long sounds that contrast with a short counterpart. Mitterer (2018) showed that in Maltese, "the singleton-geminate distinction is endangered by speech-rate variation" (p. 1). However, he also highlights cross-linguistic differences in the variation between singleton and geminate durations. This suggests that in some languages, speech rate variation may not be large enough to endanger the contrast and these languages might thus have no need to employ compensation for speech rate. This, in turn, would mean that only languages where speech rate variation is large enough in the first place are susceptible to the failure-of-compensation-based account of shortening outlined above, which prompts us to explore such effects of "endangerment" in the present study. Bukmaier and Harrington (2016) tested whether fast-speech-induced hypoarticulation could be considered a potential trigger mechanism for the diachronic neutralisation of /s̻/ within the Standard Polish three-way contrast /s, s̻, ɕ/ which has been observed in a number of Polish varieties, but found no support for this hypothesis. However, they tested the hypothesis with speakers of a variety that exhibit the contrast. In the present study we therefore will test again the effects of fast speech induced hypoarticulation on the phonetic implementation of phonemic contrasts, but with speakers of both stable (here the non-changing German standard variety) and unstable varieties (here WCB) and with a duration-based (rather than spectrum-based) contrast. The prediction is to find greater effects of speech rate in the form of greater within category variability and between category overlap in the speakers of the unstable variety than in speakers of the stable variety. No a priori predictions are made with respect to potential age group differences within the unstable variety; in this regard, the study is exploratory.

In a controlled speech production experiment, we collected original recordings from both dialect speakers and standard speakers. One variable we controlled for was speech rate, eliciting the speakers' usual tempo as well as the highest tempo they would comfortably employ. The words elicited were the same for both varieties. This allowed us a direct comparison of dialect realizations with the regional standard variety as a control group.

The results complement and extend previous findings on the dialects and regional standards spoken in the Central Bavarian dialect area (Bannert, 1976; Hinderling, 1980; Kisler & Kleber, 2019; Kleber, 2017a; Moosmüller & Brandstätter, 2014; Pfalz, 1913; Schikowski, 2009; Seiler, 2005; Wiesinger, 1990). To our knowledge, this is the first study to present experimental findings on VC words (i. e. short vowel plus lenis consonant) in these varieties; and the first controlled experiment after Bannert that deals with dialect rather than regional standard data from the Western Central Bavarian area. This study adds new and important findings about the sound change in progress that was already suggested by Kleber (2017a, 2018), Moosmüller and Brandstätter (2014) and to some extent also Schikowski (2009).

4.2 Method

4.2.1 Participants

This analysis includes data from 30 speakers in three groups: ten younger dialect speakers (aged 20-29, mean 25.3, standard deviation 2.91; 6 female, 4 male), ten older dialect speakers (aged 49 and above, mean 60.5, SD 8.15; 5f, 5m), ten younger standard speakers (aged 19-30, mean 24.1, SD 3.51; 4f, 6m). The standard speakers were all from Munich and served as a control group. For the standard group, we only selected speakers who did not speak dialect according to their own assessment (they had varying degrees of passive knowledge of the dialect). Their variety of Standard German can be described as Southern Standard German.[3] For the dialect group, we selected speakers from the Western Central Bavarian dialect region, mostly from the district of Upper Bavaria. Their dialect competence was assessed by the first author, a native speaker of WCB. Many of them were also fluent in a standard register.[4]

4.2.2 Materials

V:C	V:C:	VC	VC:
wieder	Bieter	Widder	bitter
Puder	Pute	Pudding	Butter
Tube	Lupe		Suppe
Hagen	Haken		hacken
Kader	Kater		Cutter
Rabe		Rabbi	Rappe
Tiger		Tigger	Ticker
	bieten		bitten

Table 4.1: Words used in the present study, grouped by their phoneme types in Standard German. V denotes a short vowel, V: a long vowel, C a lenis consonant and C: a fortis consonant. Every row is one minimal set or near minimal set.

[3] Standard German can be broadly divided into Northern and Southern Standard German. All speakers, except perhaps trained speakers, can be assigned to one of the two groups. Speakers of Southern Standard German exhibit, for example, [s] for word-initial /z/ phonemes but apart from such minor deviations from the dictionary pronunciation, they cannot be identified perceptually as speakers of a regional accent by phonetically naive people (i. e. they cannot be classified more precisely than northern vs. southern). More importantly, there are no noticeable differences between the northern and the southern standard variety regarding the vowel length or the fortis-lenis contrasts.

[4] Many WCB dialect speakers are able to use a standard register. When they do, however, they can be divided into speakers of a regionally-accented standard (i. e. they can be classified more precisely than northern vs. southern) and speakers of Southern Standard German.

The analysis included the 25 trochaic two-syllable words listed in Table 4.1 (a larger corpus was recorded), which are part of both the Standard German and the Central Bavarian lexicon. All words had the structure C_1VC_2X, with C_1 being any consonant, C_2 being a stop, and V being a vowel. X was either a vowel, the sequence /ən/ (where speakers sometimes elided the Schwa), or the sequence /ɪŋ/ (the latter only in the word *Pudding*). The target sounds were V and C_2. The word list comprises 8 minimal or near minimal sets. The target stops include all three places of articulation (labial, alveolar, velar) where Standard German and Central Bavarian have stops; the target vowels include the long and short variants of an i-like, u-like, and a-like vowel.

The words were embedded in carrier sentences in a way that made it likely for them to carry the sentence accent. The sentences varied per word, such that the sentence plus target word combination was meaningful. All ten repetitions (five per condition, see Section 4.2.3) of a word were embedded into the same carrier sentence.

While the regional standard speakers were presented with prompt texts in standard orthography, the dialect group were presented with carrier sentences written in a way that non-linguists are likely to use when writing dialectal text messages, using the letters of the standard language's alphabet (e. g. *Sie woit an Pudding kocha.* 'She wanted to cook pudding.'). However, to elicit natural productions of the target words, we had to minimize the effect of the exact way words were spelled. We therefore used standard orthography for the target words, but not the carrier sentences. Moreover, for both dialect and standard, we made the prompt texts disappear before participants started reading, such that they did not see the written words while talking. Dialect speakers but not standard speakers were given the full list of prompt sentences immediately before the experiment to familiarize themselves with the orthography.

4.2.3 Procedure

Speakers produced the sentences in alternating speech rate blocks. To determine speaker-specific sentence durations, we asked participants, in a test phase prior to the experiment, to read out six different sentences, three *at their usual speaking rate*, and three *as fast as they could while still feeling comfortable*[5]. We then measured the length of these utterances and averaged the length per condition, rounding to a multiple of 100 ms. 400 ms were added to allow the participants some time for preparing to speak. This typically resulted in values between 1,000 and 2,000 ms for both conditions. The difference between the two conditions was 200 or 300 ms for 28 participants and 400 ms for the remaining two participants.

The main phase began with a normal speech rate block. Participants were given 1.5 seconds to read the sentences silently from a screen. The text was then replaced by a progress bar that visualized the predetermined speaker-specific sentence duration at a

[5]With some participants, we repeated the fast condition, because they failed to accelerate their speech measurably in the first try.

normal speech rate. They now had to reproduce the sentence aloud from their memory. The task was to utter the sentence during the time the progress bar completed. After all sentences were spoken that way, the block was over. In the next block, participants were presented with the same prompts again, but while the time to silently read the sentence remained the same as in the previous block, participants were given less time to produce the sentence. The time given in this fast speech block was again indicated by the progress bar – now set to the predetermined speaker-specific fast speech rate. This procedure of alternating speech rate blocks was repeated 10 times, resulting in 5 repetitions of each token at a normal speech rate and 5 repetitions at a fast speech rate. Speakers were informed about the targeted speech rate prior to each block. Prompts were presented in randomized order, with each block containing a different order of prompts. The randomization was the same for all participants.

The SpeechRecorder software (Draxler & Jänsch, 2004) was used to present prompts and make recordings. The acoustic recordings were conducted either in a sound-attenuated recording booth at the Institute of Phonetics and Speech Processing in Munich, or with mobile equipment in quiet environments in participants' homes. In either case, a head-mounted Beyerdynamic Opus 54 condenser microphone was used. The audio signal was digitized at a sampling rate of 44.1 kHz and a resolution of 16 bit, using a PreSonus audio interface in the studio and an M-Audio device in the mobile setting.

4.2.4 Segmentation and measurements

The complete utterances were automatically segmented with MAUS (Kisler et al., 2016). Based on the automatic results, the relevant segment boundaries (start and beginning of the utterance, start and beginning of each phoneme in the target word, burst in the target word stops) were checked and adjusted manually with the EMU Speech Database Management System (Jochim, 2017; Winkelmann et al., 2017). We considered the end of each segment to be the start of the following segment. The bursts were identified using the intensity spike in the wave form. The beginning and end of the target vowels were set to the center of the first and last visible vertical bar, respectively, in the spectrogram. This criterion was chosen to allow the best possible consistency across different labelers at the expense of possibly (but systematically) underestimating vowel durations and overestimating the durations of vowel-adjacent segments.[6] These vowel boundaries, then, also determined the boundaries of the adjacent segments, notably the start of the closure in post-vocalic stops and the end of aspiration in pre-vocalic stops. In cases where the vertical bar coincided with the burst of the preceding stop, aspiration was set to 0. Negative VOT never occurred in our data. Some lenis stops in the dialect speakers appeared as approximants, which is a typical surface form of hypoarticulated lenis stops in Central Bavarian. These were included in our analyses as stops with 0 VOT.

[6]Note that the data discussed here are a subset of a larger corpus collected and annotated by three research institutions with several labelers involved.

4.2.5 Data analysis

In this section, we will describe the variables derived from the measurements described above, proceeding from simpler to more complex variables.

$closure_{norm}$ and VOT_{norm}

The two simplest variables were defined as follows. $closure_{norm}$ (word-normalized closure duration) was defined as the duration of the stop closure (i. e. from the offset of the preceding vowel to the burst) divided by target word duration. VOT_{norm} (word-normalized voice onset time) was defined as the duration of stop aspiration (i. e. from the burst to the onset of the following vowel) divided by target word duration. The divisions were done to normalize for the token-specific general speech rate.

Optimal category boundary

In order to investigate whether some effects do or do not put the fortis–lenis distinction at risk, we calculated the optimal category boundary, following the procedure described by Miller et al. (1986) and also employed by Mitterer (2018). This method searches a threshold for a given unidimensional acoustic feature (in our case stop closure duration) that divides a set of tokens in two categories (in our case, fortis and lenis). To this end, the researchers choose a range of values (in our case 0 to 280 milliseconds) and calculate the classification accuracy for each of these values. The value with the highest accuracy is then considered the optimal category boundary.

Category expansion

In order to investigate the (in-)stability of a category, we calculated a measure we call *category expansion*. It is closely linked to the dispersion of the category's acoustic parameters. We consider a category that extends over a wide range of values in an acoustic space to be *large*, and a category that is limited to a small range of values to be *compact*. The difference in dispersion between the fast and the normal-paced speaking condition, then, accounts for category expansion. Categories that are larger in fast speech have positive expansion values, whereas categories that are more compact in fast speech have negative expansion values. We theorize high values to be indicative of instability in the respective phonological category, because we think that fast speech rate puts a pressure on the categories that a stable category should easily absorb, while an unstable category should be negatively affected.

The expansion value is defined as the difference between the Coefficient of Variation (CoV, $\frac{StandardDeviation}{Mean}$) in fast speech and the CoV in normal-paced speech. In the present study, we only report on category expansion based on the means and standard deviations of $closure_{norm}$, although other acoustic measures are conceivable as a basis as well.

Fortis–lenis overlap (FLO)

For our investigations of word-specific effects, we use a measure we call fortis–lenis overlap (FLO) to quantify the overlap between (near) minimal pair words that have a fortis–lenis contrast in their stop. It defines the acoustic region typical of lenis stops (the lenis region) to include any closure duration below the third quartile of lenis closure durations (in a given group of tokens). The measure, then, represents, among a given group of fortis tokens, the share of tokens that fall in the lenis region. It can take values between 0 and 1, meaning no fortis tokens or all fortis tokens, respectively, were produced in a lenis-like fashion.

4.2.6 Statistics

All statistical analyses were conducted with the statistical software R (version 3.4.4, R Core Team, 2018) and the R packages lmerTest (version 3.0-1, Kuznetsova et al., 2018), lme4 (version 1.1-17, Bates et al., 2018), and emmeans (version 1.3.1, Lenth et al., 2018).

For the analyses in Sections 4.3.1, 4.3.2, and 4.3.3, we fitted three linear mixed-effects models on our data. The main purpose of this was to be able to estimate marginal means for the various factors. This was not used for the word-by-word analysis in Section 4.3.4, because the measure used as a dependent variable there (fortis–lenis overlap, FLO, see Section 4.2.5) strongly reduces the raw data; this enables us to consider the entirety of FLO data points in an analysis using visualization and basic descriptive tools (thus eliminating the need for estimating marginal means). A second reason for fitting linear mixed-effects models was to test the statistical significance of factors. In the framework of null hypothesis significance testing (NHST), we compare the p values of fitted models to predefined alpha values (5%, 1%, and 0.1%) in order to reduce the risk of reporting false positive effects. One major problem of this procedure (comparing the p value) is that it operates on the assumption that the tested factor does not have an effect in the population underlying our sample.[7] Conversely, it yields no numerical result about the case where the tested factor does have an effect in the population, which limits the procedure's ability to decide whether the tested factor does or does not have an effect in the population. However, the procedure indeed has the advantage of being a standard procedure in the field of phonetics (and beyond). Pragmatically speaking, it also helps in spotting biased interpretations of raw data visualizations.

Two linear mixed-effects models included the dependent variables word-normalized closure duration ($closure_{norm}$) and word-normalized voice onset time (VOT_{norm}), respectively. Both of them included the fixed factors *speaker group* (younger standard speakers, younger dialect speakers, older dialect speakers), *quantity category* (V**ː**C, V**ː**C**ː**, VC, VC**ː**, with V denoting a vowel, C a consonant, and **ː** being the length marker), and

[7] The p value is defined as the conditional probability of observing either the observed data or data that deviate more strongly from the null hypothesis, *given that the null hypothesis is true* (Fahrmeir et al., 2016, p. 388).

speech rate (fast, normal)[8]; and the random factors *speaker* and *target word*. The third model included the dependent variable *category expansion* (based on $closure_{norm}$); it included the same set of fixed and random factors as the other two models, with the exception that *speech rate* was not included. The models were specified as shown in equations 4.1, 4.2, and 4.3, respectively. Based on these models, we carried out pairwise comparisons using Tukey's method for correcting family-wise errors.

$$\begin{aligned} closure\_norm \tilde{} \; speaker\_group * category * rate \\ + (category + rate|speaker) + (speaker\_group + rate|target\_word) \end{aligned} \tag{4.1}$$

$$\begin{aligned} vot\_norm \tilde{} \; speaker\_group * category * rate \\ + (category + rate|speaker) + (speaker\_group + rate|target\_word) \end{aligned} \tag{4.2}$$

$$\begin{aligned} category\_expansion \tilde{} \; speaker\_group * category \\ + (category|speaker) + (speaker\_group|target\_word) \end{aligned} \tag{4.3}$$

[8]Normalization was applied to $closure_{norm}$ and VOT_{norm} precisely to remove the effect of token-specific speech rate (see Section 4.2.5). It may seem counter-intuitive, then, to test for the effect of *speech rate* on these normalized measures. See Section 4.3.3 for an explanation.

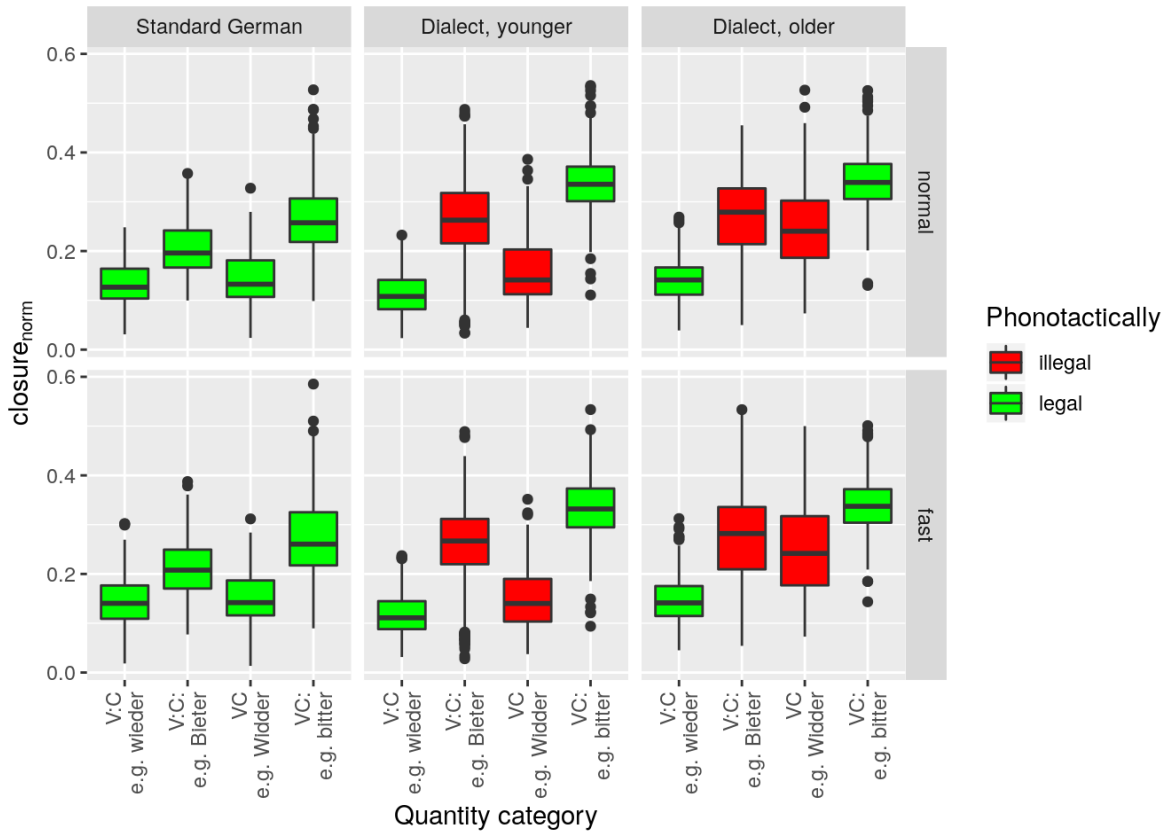

Figure 4.1: Word-normalized stop closure duration in the four *quantity categories* (V**ː**C, V**ː**C**ː**, VC, VC**ː**) separately for the three *speaker groups* (columns) and *speech rates* (rows). Bavarian phonotactically illegal clusters are highlighted in red.

4.3 Results

4.3.1 Closure duration

The *closure*$_{norm}$ data in Fig. 4.1 support the assumption that our control group, the standard speakers, would produce higher closure durations in fortis (C**ː**) than in lenis (C) stops; and higher closure durations after short vowel (V) than after long vowel (V**ː**). The younger but not the older dialect speakers exhibit the same pattern as the control group. The key difference is that the older speakers produce short vowel + lenis (VC) words with almost fortis-like closure durations. Fig. 4.1 further suggests that the closure duration in fortis stops is higher for dialect speakers (both younger and older) than for standard speakers.

The linear mixed-effects model revealed statistically highly significant main effects for *speaker group* ($F[2,42.1] = 12.2$, $p<0.001$) and *quantity category* ($F[3,22.9] = 21.9$, $p<0.001$), and a significant main effect for *speech rate* ($F[1,36.8] = 4.6$, $p<0.05$]). Sta-

tistically significant interactions were revealed between the factors *speaker group* and *quantity category* ($F[6,32.3] = 4.5$, $p<0.01$), and between *speaker group* and *speech rate* ($F[2,110.5] = 6.6$, $p<0.01$).

Pairwise comparisons in the model corroborate the observation that the VC category is similar in the control group and the younger dialect speakers, but different in the older dialect speakers (cf. Table 4.2). As for the fortis categories (VC: and V:C:), the model revealed that $closure_{norm}$ is higher in dialect than in standard, and slightly higher in older dialect than in younger dialect speakers (cf. Table 4.3; this is also reflected in the statistical significances of pairwise comparisons, with the exception that, for V:C:, the difference between control group and younger dialect speakers does not reach statistical significance). For the V:C category, the difference between the two dialect groups is statistically significant but not the difference between control group and either of the dialect groups.

The speech rate effect found in the data is very subtle, with a fast–normal difference of no more than 0.8 percentage points in any of the three speaker groups. Pairwise comparisons revealed that the estimated difference is 21.4 vs. 21.6 % for younger and 24.9 vs. 25.1 % in older dialect speakers (neither of them statistically significant). The estimated difference for the control group is 18.8 vs. 19.6 % and turned out to be statistically significant ($p<0.001$).

These findings indicate that the VC category, which has been described as phonotactically illegal in Central Bavarian, does indeed not occur in older dialect speakers, but very much so in younger dialect speakers. This can be interpreted as a sound change in progress, by which long stops (C:) are shortened after short vowels (V), lifting the phonotactical restriction of *no lenis after short vowel*. The findings further suggest that V:C:, contrary to predictions, already exists in the phonological system of all dialect speakers. In line with the literature, the findings indicate that the dialect features higher closure duration in fortis stops than the regional standard.

contrast	estimate	SE	df	t.ratio	p.value
Y_SD - Y_WB	-0.008	0.023	36.22	-0.352	0.9342
Y_SD - O_WB	-0.100	0.025	33.48	-3.932	0.0011
Y_WB - O_WB	-0.092	0.017	45.11	-5.345	<.0001

Table 4.2: Pairwise comparisons of estimated marginal means for $closure_{norm}$ of VC tokens (short vowel plus lenis), averaged across the levels of *speech rate*.

4.3.2 Voice onset time

The model with VOT_{norm} as the dependent variable revealed a significant main effect for *quantity category* ($F[3,26.4] = 10.7$, $p<0.001$), but neither for *speech rate* nor *speaker group*. It also revealed one significant interaction and that is between *quantity category*

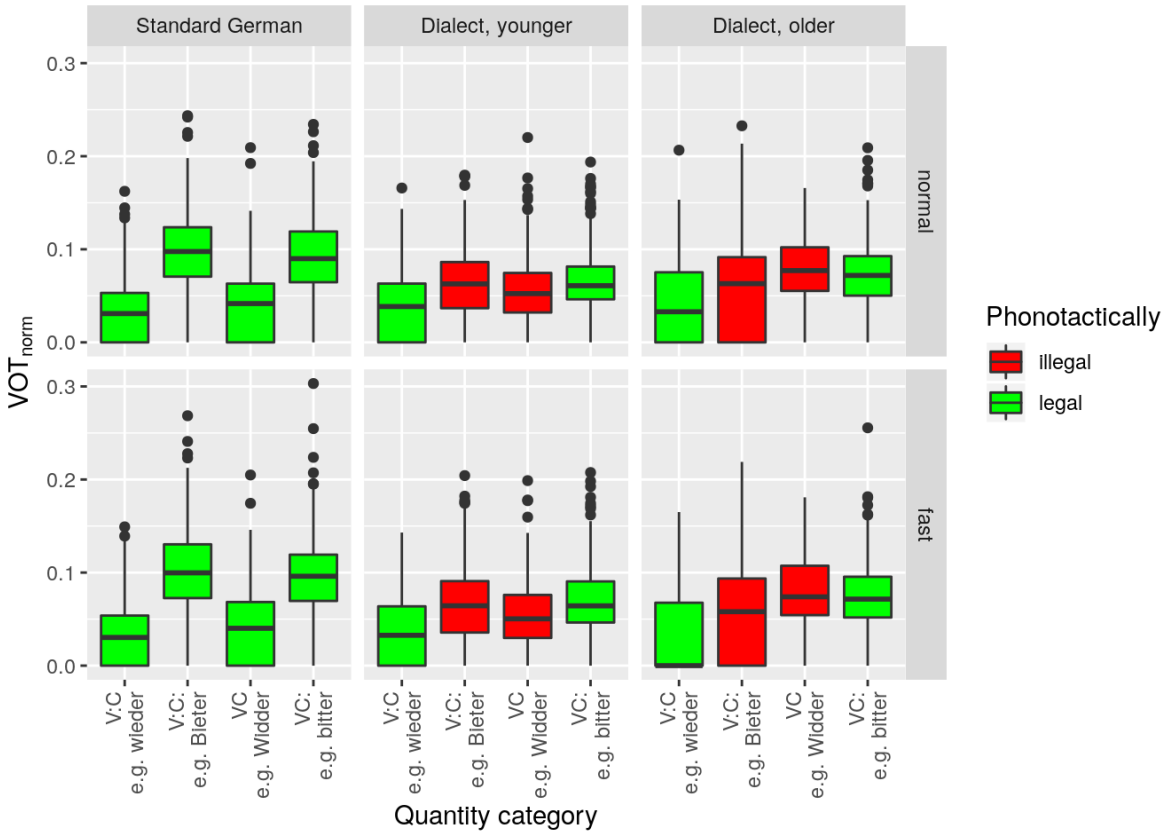

Figure 4.2: Word-normalized voice onset time by *speaker group*, *speech rate*, and *quantity category*.

speaker_group	emmean	SE	df	lower.CL	upper.CL
V:C:					
Y_SD	20.97 %	0.021	30.97	0.1671	0.2525
Y_WB	25.54 %	0.022	30.22	0.2112	0.2997
O_WB	26.48 %	0.022	29.29	0.2185	0.3112
VC:					
Y_SD	27.10 %	0.019	35.10	0.2325	0.3094
Y_WB	33.49 %	0.020	34.12	0.2951	0.3746
O_WB	34.19 %	0.020	32.89	0.3004	0.3834

Table 4.3: Estimated marginal means for $closure_{norm}$ in fortis consonants (after long vowel (V:C:) and after short vowel (VC:)), averaged across the levels of *speech rate*. Dialect speakers have higher closure durations than regional standard speakers.

contrast	estimate	SE	df	t.ratio	p.value
Y_SD:					
V:C - V:C:	-0.0647	0.010	39.95	-6.553	<.0001
VC - VC:	-0.0501	0.010	34.37	-4.972	0.0001
Y_WB:					
V:C - V:C:	-0.0279	0.009	41.58	-2.959	0.0252
VC - VC:	-0.0124	0.010	35.97	-1.302	0.5675
O_WB:					
V:C - V:C:	-0.0233	0.013	32.17	-1.818	0.2838
VC - VC:	0.0058	0.013	28.12	0.434	0.9721

Table 4.4: Pairwise comparisons of estimated marginal means for VOT_{norm}, averaged across the levels of *speech rate*.

and *speaker group* ($F[6,34.6] = 7.0$, $p<0.001$). This suggests that the different speaker groups employ VOT as a cue for phonological quantity in different manners.

Commensurate with Fig. 4.2 and Table 4.4, pairwise comparisons show a clear and unsurprising pattern in the standard control group: (1) Fortis consonants have a higher VOT than lenis consonants. (2) The VOT difference turned out significant both after long and short vowels. (3) VOT did not change as a function of speech rate ((3) is also true of the two dialect groups). This indicates that VOT is a very stable cue for the fortis–lenis contrast in the regional standard. The younger dialect group exhibits the same general pattern as the standard group, with fortis VOT above lenis VOT in all contexts. However, the fortis–lenis difference is extremely small and only reaches statistical significance after long vowels. This finding suggests that VOT may not yet be used as a robust cue in the production of the phonological fortis–lenis contrast by younger dialect speakers. This interpretation is also supported by the trend that younger dialect

speakers produce lower VOT for fortis stops than standard speakers.[9] Older dialect speakers did not produce a statistically significant VOT difference between fortis and lenis stops, neither after long nor after short vowels, supporting previous accounts by which VOT is said to play no role in Bavarian. Fig. 4.2, however, also shows a greater tendency towards longer VOT after short vowels. This observation reflects the dialectal pattern of stop fortition after short vowels.

Taken together, the findings suggest that VOT was not used as a cue in the older state of the WCB dialect as represented here by the older generation (which is in line with previous accounts, cf. Bannert, 1976; Seiler, 2005; Wiesinger, 1990), but that younger dialect speakers are starting to adopt the cue (which is in line with Kleber's (2018) observations on these speakers' regionally-accented standard register).

4.3.3 Fast-speech-induced hypoarticulation and variation in closure duration

After having established in section 4.3.1 that the VC category is becoming legal in younger dialect speakers we will now consider in more detail whether this sound change in progress is at least to some extent system-internally driven. Given that words such as *Widder* were previously produced with a fortis stop and now exhibit a lenis stop, this is a case of lenition and therefore a particularly relevant candidate for a sound change triggered by fast-speech-induced hypoarticulation. This section will shed light on the effect of speech rate on the fortis–lenis contrast from three different angles: First, we will explore contrast endangerment to show that speech rate variation is indeed an important factor in maintaining the fortis–lenis contrast; second, we will test whether the speech rate effect is disproportionately strong in the category that has changed between generations (VC); and third, we will test whether dispersion is affected disproportionately strongly.

Contrast endangerment

Figure 4.3 shows each speaker's optimal category boundary between the closure duration of lenis and fortis stops and separately for the two speech rate conditions and speaker groups. It demonstrates a large range of between-speaker variability, especially in the older dialect group, where speakers range from 70 to 127 milliseconds (ms). The younger dialect group ranged from 52 to 86 ms, the control group from 39 to 62 ms. Also commensurate with Figure 4.3, this between-speaker effect is much larger in our data than the within-speaker effect explored in detail in the previous sections in the form of the fixed effect *speech rate*.

The minor acceleration effect within speakers thus may suggest at first glance that the fortis–lenis distinction is not at risk. However, when taking into account the range

[9] This trend appears very clear in Fig. 4.2, but fails to reach statistical significance.

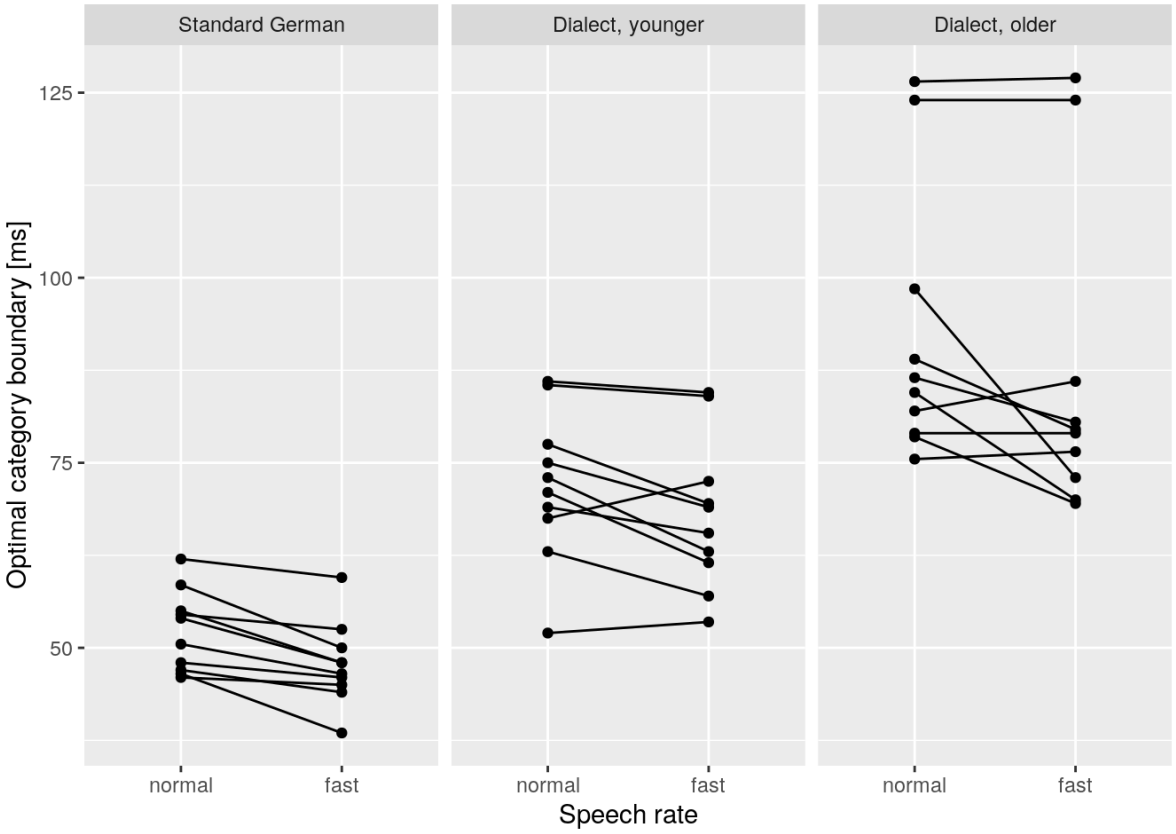

Figure 4.3: Optimal category boundary between fortis and lenis stop closures per speaker and speech rate. The lines connect the pairs of data points that belong to the same speaker.

of idiosyncratic speech rates observed across speakers, the fortis–lenis contrast perhaps may become endangered after all at the group level, namely then when speakers-turned-listeners do not compensate for speech-rate-induced variation in closure durations unknown to them; that is, values outside their own scope of rate-induced variation.

Lenition in fast speech

The older dialect speakers have a long consonant in the VC, VC**ː**, and V**ː**C**ː** words. The younger speakers have retained the long consonants in VC**ː** and V**ː**C**ː**, but have shortened those in VC. If this consonant shortening had been caused by failing compensation for speech rate, we would expect the long consonant in the older speakers' VC to be less stable across speech rates (and therefore, we suppose, harder to normalize) than in the other two categories. This should surface in a stronger effect of *speech rate* on (the already word-normalized measure) $closure_{norm}$ in VC, compared to the other two categories. We hypothesize:

In the fast speech condition, older dialect speakers reduce $closure_{norm}$ *in VC but not in V**ː**C**ː** and not in VC**ː** – possibly to the extent that VC is merged with V**ː**C (with V**ː**C being the only category where older speakers have a short consonant to begin with).*

However, pairwise comparisons in our model from section 4.3.1 revealed a statistically significant speech rate effect only for 3 out of 12 pairs (3 speaker groups * 4 quantity categories), and even these effects are tiny: V**ː**C**ː** in older dialect speakers (26.1 % in normal vs. 26.9 % in fast, $p<0.05$), V**ː**C in younger standard speakers (13.5 % vs. 14.4 %, $p<0.01$) and VC**ː** in younger standard speakers (26.6 % vs. 27.6 %, $p<0.01$).

These findings suggest that faster speech rate only shortens stop closures in a manner exactly proportional to word shortening. Contrary to our prediction, older dialect speakers did not shorten closure phases (i. e. lenite) in the VC category disproportionately strongly.

Category expansion

To extend the test whether older dialect speakers show signs of instability in fast speech, we measured the *category expansion* (cf. section 4.2.5) based on $closure_{norm}$ in the data. We expected unstable phonological conditions to be associated with expanding categories (i. e. categories that are more dispersed in fast speech than in in normal speech).

However, commensurate with Fig. 4.4, no group differences were found in how speech rate affects the dispersion of the phonological categories. The corresponding statistical model (with *speaker group* and *phonological category* as fixed factors) did not reveal a significant main effect or interaction. This finding corroborates the result from the previous section (on lenition in fast speech) that there is no evidence of instability in the dialect speakers' fast speech.

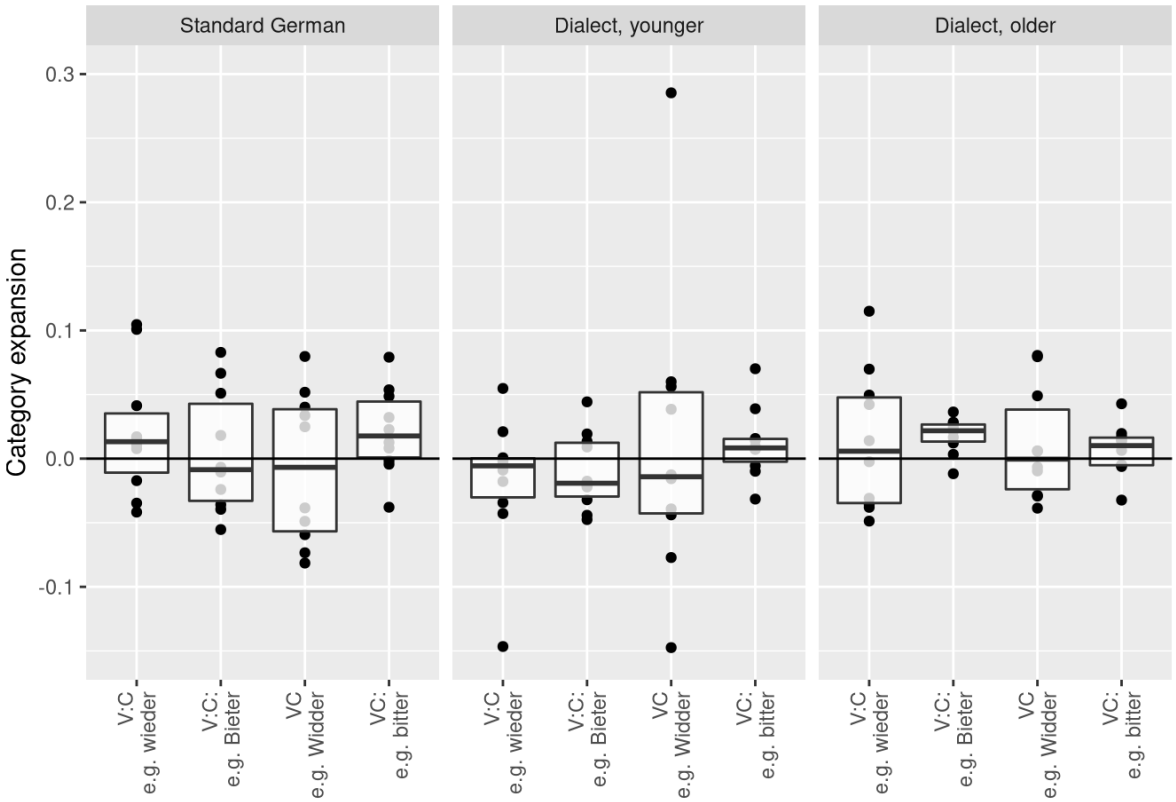

Figure 4.4: *Category expansion* based on $closure_{norm}$. Positive values indicate more dispersion of the acoustic parameter $closure_{norm}$ in the fast condition as opposed to the normal-paced condition. *Note that the dots are not outliers as would be typical for box-and-whiskers plots. The dots represent raw data points (one per speaker). The whiskers have been omitted for clarity.*

4.3.4 Lexical diffusion

In this final analysis, we will explore our data on a word-by-word basis. In Figures 4.5 and 4.6, we use the *fortis–lenis overlap* (FLO, see Section 4.2.5) to show how often "fortis words" (i. e. words that contain fortis stop phonemes in Standard German) were realized with a lenis-like closure duration. The FLO can take values between 0 and 1. We consider a FLO between 0 and 1 to be a sign of unstable categories (0 means that no fortis tokens were produced lenis-like, while 1 means that all tokens were produced lenis-like).

Commensurate with Fig. 4.5, there is a group effect for alveolar stops. For alveolar fortis stops, the 10 control group speakers have a low FLO (mean 0.04, standard deviation 0.04), while the 10 younger dialect speakers have a high FLO (mean 0.13, SD 0.06) and the 10 older dialect speakers have an even higher FLO (mean 0.23, SD 0.10). For bilabial and velar stops, all speakers have a relatively low FLO (highest among them are the control group's labial stops at a mean of 0.05 (SD 0.08)).

Figure 4.6 therefore focuses on the alveolar stop, depicting the FLO for each individual word with an alveolar fortis stop. We can see that the word *Kater* 'tomcat' is almost exclusively produced as lenis by the older dialect speakers (FLO being 1 for 9 out of 10 speakers), but has a slight tendency towards fortis in some younger dialect speakers (with one speaker exhibiting a FLO of 0.6, four speakers a FLO of 0.8 and five speakers a FLO of 1; that is, 3 of 5, 4 of 5, and 5 of 5, respectively, of the speaker's repetitions exhibiting acoustic values typical of lenis stops). For the words *bieten, Bieter, Pute* ('to bid', 'bidder', 'turkey hen'), the older dialect group is very heterogeneous in whether they have lenis or fortis stops. The younger dialect speakers, however, have pretty much settled on fortis.

These findings suggest that the change is governed by lexical diffusion[10], that is any given phoneme can reflect a conservative state in some words but an innovative state in other words, and this can even vary within speakers (as shown by FLO values far away from both 0 and 1 for certain words).

[10]We follow Crystal's 2008 (p. 145) definition of lexical diffusion, treating it as a surface phenomenon (namely that different lexical items can be affected to different degrees) and associating no assumption about the underlying mechanism with the term.

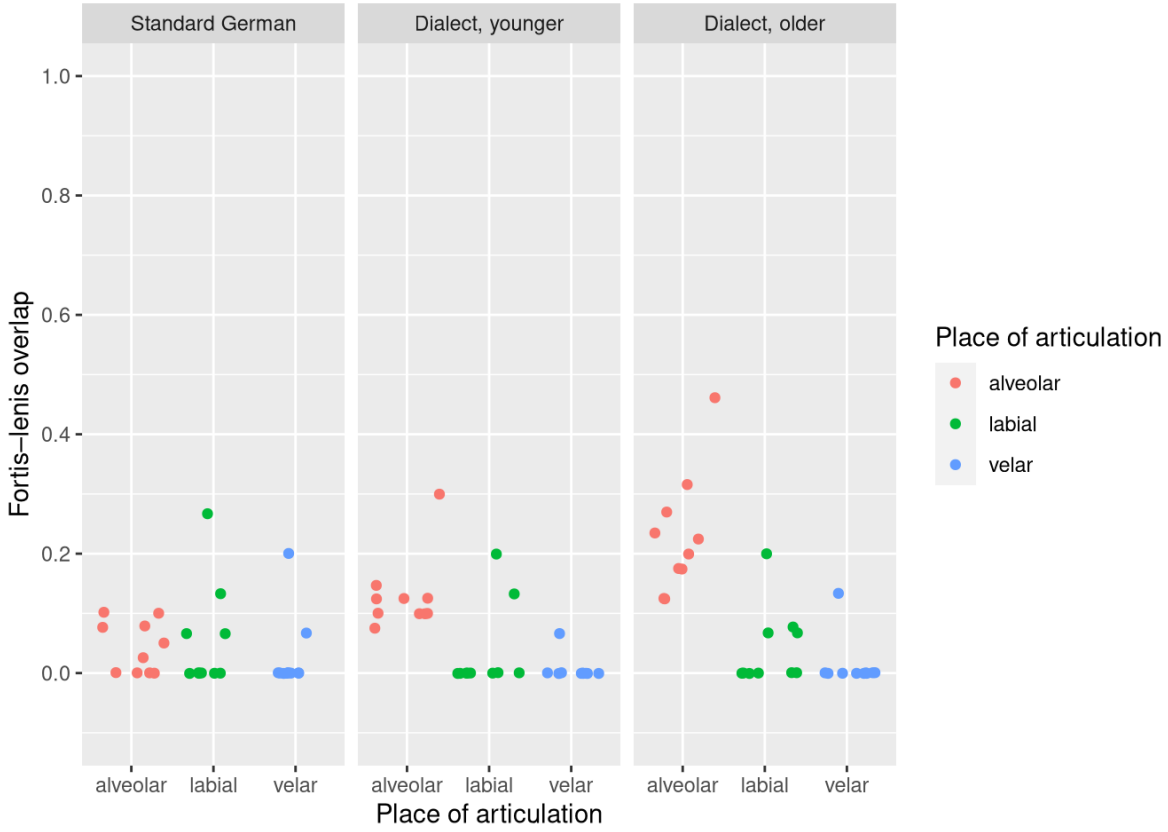

Figure 4.5: Endangerment of category boundary (by place of articulation). Each data point represents the combination of one speaker and one word group. The word groups comprise words containing a fortis stop. The y axis represents the share of fortis stops (among all fortis tokens in the respective word group by the respective speaker group) whose closure durations are "typical of lenis stops," that is, below the third quartile of closure durations of the corresponding lenis stops.

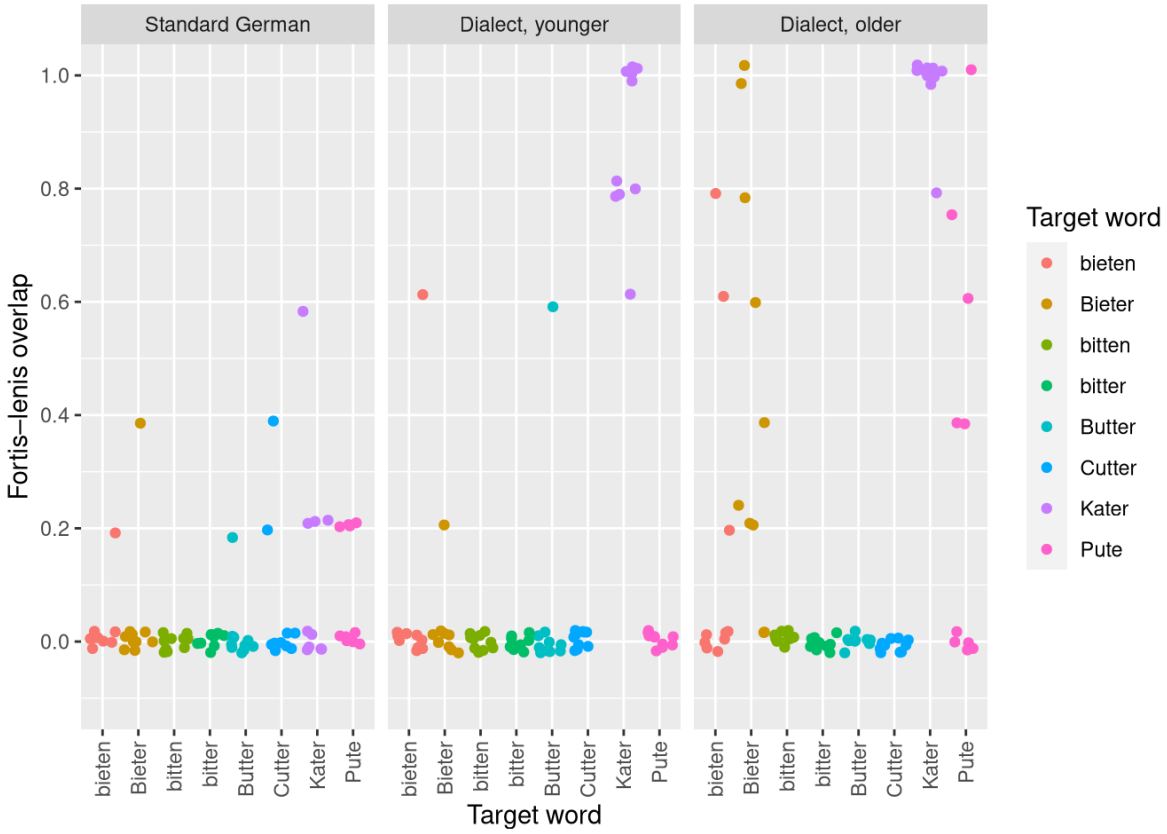

Figure 4.6: Endangerment of category boundary (by word – alveolar only). Same as Figure 4.5, but separated by word instead of word group. The labial and the velar group are not included here.

4.4 Discussion

Our three main aims in this paper revolved around an ongoing change in a German dialect. We aimed (1) to test whether long vowel plus fortis stop sequences emerge in the Western Central Bavarian (WCB) dialect (as shown for Eastern CB (ECB) in Moosmüller and Brandstätter (2014) and suggested by WCB speakers' usage of the regionally-accented standard register (Kleber, 2017a, 2018)); (2) to test whether this change extends to short vowel plus lenis stop sequences. Finally, we aimed (3) to model naturally-occurring fast speech to test whether such a change can also be related to language-internal factors (in this case, fast-speech-induced hypoarticulation). To achieve these aims, we used a newly collected corpus of dialect recordings that included all four phonotactic categories involved in the sound change (including the previously understudied short vowel plus lenis stop (VC) combination). In the corpus, we further required speakers to doubtlessly be operating in a dialect register in spite of the laboratory setting. We achieved this by writing down the prompts in non-orthographic forms and having speakers recite those from memory shortly after seeing them in writing.

The five main findings are as follows: stop closure durations indicate that (a) the combination of short vowel plus lenis consonant has become legal in younger dialect speakers, and (b) the combination of long vowel plus fortis is legal even in older dialect speakers. As to voice onset time (VOT), we found that (c) younger dialect speakers' usage of this acoustic parameter in speech production is between that of older dialect speakers and younger standard speakers. (d) An analysis of two acoustic parameters and further derived measures did not yield a plausible reason why fast-speech-induced hypoarticulation can be considered a trigger of the observed sound change. Furthermore, (e) analyses of such derived measures, specifically fortis–lenis overlap (FLO), showed that some words exhibit markedly more between-speaker and within-speaker variation than others.

Our results confirm that the observations made for the Viennese dialect (ECB, see Moosmüller and Brandstätter, 2014) and the WCB regional accent (Kleber, 2017a, 2018) (i. e. Standard German, but noticeably produced by WCB dialect speakers), also hold true for the WCB dialect: The combination long vowel plus fortis consonant (V:C:) already forms part of the older dialect speakers' phonological system (participants of the present study born 1944–1968). Our results further confirm that the other supposedly illegal combination, short vowel plus lenis consonant (VC), does indeed not occur in the older speakers, but it does occur in the younger speakers (participants of the present study born 1987–1997). On the whole, this is in line with the results reported in recent years (Kleber, 2017a, 2018; Moosmüller & Brandstätter, 2014; Schikowski, 2009) suggesting that the complementary length feature of Central Bavarian dialects is levelling out: Both lenis and fortis consonants have existed before, but the combinations of short vowel plus lenis consonant (VC) and long vowel plus fortis consonant (V:C:) are new options in the dialect's phonology.

For the VC combination, our results provide empirical evidence both that the com-

bination does exist now, and that it did not exist in older generations. For the V:C: combination, our findings indicate that it does exist now, but they also suggest that it existed in earlier generations, too. The latter point contradicts Kleber's (2017a, 2018) reports, where an age cohort of dialect speakers very similar to ours was found not to employ V:C: combinations. However, the participants she tested were dialect speakers operating in a (regionally-accented) standard register. It is counter-intuitive to some extent that older speakers (who mainly speak dialect) would borrow a grammatical form from the standard language into their dialect, but then not employ said form when switching to their standard register. It does not seem wholly implausible, however, especially when they have a lot of standard language input (e. g. from mass media) to shape their everyday language (which is the dialect), but only rarely actively use the standard language (i. e. speak/produce in this variety). It seems quite plausible that an individual's rarely-used register remains unaffected by change processes for longer than their most-often-used register. It must be noted, however, that bilinguals have been reported to change the fine phonetic detail of one language during times when they predominantly use the other language (Sancier & Fowler, 1997). Yet it is unclear whether this can be transferred to a nuanced contact situation like the one investigated here, with a standard language that people often hear, a dialect that people often speak, and a regionally-accented standard that is somewhere in between in terms of usage frequency. Since this is quite speculative, another explanation for the differences might be more robust: Kleber (2017a, 2018) tested a different set of words and the mechanism cannot necessarily be generalized over both sets of words (see below for a discussion of lexical diffusion).

In addition to changing phonotactics, our results also indicate that usage of voice onset time (VOT) as a cue to the lenis–fortis contrast in stops is becoming stronger in Western Central Bavarian. While the younger dialect speakers still produce shorter VOT in fortis stops and also a smaller degree of separation between lenis and fortis (in terms of VOT) than the standard speakers, they also deviate clearly from the older dialect speakers, who produce yet shorter VOT and almost no separation between lenis and fortis (again, in terms of VOT). This is in line with Kleber (2018), who found that younger speakers of the WCB regional accent (rather than dialect) employ VOT in their production more strongly than older speakers. She also found that in perception, the relative importance of closure duration[11] and VOT has shifted towards VOT in the younger listeners. We interpret these apparent-time observations as a hint of the emergence of a new acoustic cue (VOT) borrowed from the regional standard – in addition to the lifted phonotactic restriction.

Our research was focused on the phoneme level, grouping all words from the corpus into the four categories V:C, V:C:, VC, and VC:. A close look inside these groups reveals that, on the whole, they are pronounced homogeneously within each speaker group. However, one word in particular – *Kater* 'tomcat' – contradicts this generalization. Like, for

[11]Note that she tested closure duration as a part of the vowel-to-closure duration ratio.

example, *Haken* 'hook' and *Lupe* 'magnifying glass', *Kater* was in the V:C: group due to its canonical pronunciation in Standard German. However, in both the older and the younger dialect speakers, the closure durations in this word severely stand out from the other words in the V:C: group. The observed lenis-like realizations (of the alveolar stop in *Kater*) are in line with the grammar of Western Central Bavarian and exactly what many dialect researchers would expect. However, the other words in the group did not match this expectation, often clearly exhibiting both a phonetically long vowel and long stop. There were other deviations from homogeneity as well. The words *bieten* 'to bid', *Bieter* 'bidder', and *Pute* 'turkey hen' exhibited considerable within-speaker and between-speaker variation in the older dialect group, with some speakers leaning towards a fortis stop, some towards a lenis stop, and some without a clear preference for either. This observation suggests that in future studies, lexical diffusion must be considered as a mechanism for the observed sound change. With the present study having established the relationships of the four categories within and between speaker groups, a follow-up design does not depend on minimal pairs contrasting the categories. The fact that only few such minimal pairs exist limited the amount of words we could test in the present study. A follow-up study now could test a large set of words, apt for investigating the spread of the change through the lexicon.

Kleber (2017a) discussed in more detail the plausibility of this (and other changes for that matter, e. g. Bukmaier et al. (2014), Harrington et al. (2012)) to be driven by external factors (dialect levelling) rather than internal factors. With the present paradigm of modelling naturally-occurring fast speech, we put to the test an alternative explanation based on Kohler (1984) and Ohala (1993a): fast-speech-induced hypoarticulation providing a phonetic bias to diachronically lenite fortis phonemes. One of our main findings is that the four words that make up the VC group (e. g. *Pudding*) have a word-medial fortis stop in the older dialect speakers but that stop is lenis in the younger dialect speakers. If this shortening had been triggered by the above-mentioned phonetic bias, we would have expected to find synchronic lenition in the older group's fast VC words (more than in the other word groups, where no diachronic lenition has been observed in the past decades), particularly in the consonants. However, we did not find this kind of lenition. We therefore extended our analyses to see if any instabilities could be found in the words affected by the change, either in the older dialect speakers (to suggest a phonetic bias leading to the change) or in the younger speakers (as a sign that the new phonology has not yet stabilized). No such instabilities were found, either – in none of the investigated cues closure duration, VOT, burst intensity, vowel duration, and vowel formants.[12] There are several problems that might have concealed the expected effect. (1) Generally, consonants have been found to be affected by speech rate increases much less than vowels (Gay, 1981), and (2) short phonemes (i. e., the vowels in the first syllable of the VC words) have been found to be affected less than long phonemes (because

[12]Due to spatial limitations, not all of the analyses we conducted are presented in detail in the results section. All data and code (in the form of an *R Notebook*) to reproduce them are found in Section 4.5.

they are short already; Hoole and Mooshammer, 2002). However, these studies, just like ours, *did* find shortening effects on consonants and on lax vowels. The effects are simply small in size. (3) Bukmaier and Harrington (2016) used a similar paradigm to test fast speech rate as a trigger mechanism of a fricative-merger in some Polish dialects. They failed to find such an effect, although dialect levelling as an alternative explanation could be ruled out in their case. None of these limitations explain why the expected effect was not found at all, in spite of looking at the data from many phonetic angles. This leads us to the conclusion that this particular instance of a sound change in progress is directly triggered by language contact (dialect levelling), as hypothesized in Kleber (2017a), with no need for a specific phonetic bias in the older state of the language to foster the change.

To sum up, we found new phonotactics as well as one new cue in the phonology of the German dialect Western Central Bavarian. It is very likely that these innovations were brought about by dialect contact alone and not fostered by phonetic bias. In future studies, it will be more important than before to account for lexical diffusion in the mechanism of this type of sound change.

4.5 Acknowledgments and Data

The data and scripts to reproduce the analyses of the present paper have been published as Jochim and Kleber (2022a).

5 Agent-based modeling

A variant of this chapter has been submitted for publication as:

Jochim, M. & Kleber, F. Reconstructing the timeline of a prosodic change in a German dialect: Evidence from agent-based modeling.

5.1 Introduction

This chapter has a twofold aim that revolves around a particular agent-based simulation model of sound change. We will call this model the ABM_{IPS}. One aim, methodological in nature, is to evaluate the generative sufficiency of the ABM_{IPS}; that is, how well the assumptions that make up the model are able to predict data observed in the laboratory. In the simulation literature, this evaluation process is called the validation of a model (Cioffi-Revilla, 2017a). The other aim is to use the very same model in combination with data from Chapter 4 to contribute to our knowledge of sound change theory and the particular sound change in progress observed in Bavarian (Kleber, 2017a; Moosmüller and Brandstätter, 2014; Chapter 4 of this dissertation). The ABM_{IPS} has been developed and has received initial validation tests in Harrington and Schiel (2017), Harrington et al. (2018), Stevens et al. (2019), and Harrington, Gubian, et al. (2019).[1] With a simulation model at this early stage of development, we are interested in contributing to the model's development, and at the same time we want to carefully explore how results of the model can already help us to generalize from the "model world" to the "real world" (i. e. phonetic theory). This exploration will be strongly guided by the epistemological overview given in Chapter 1 (especially Section 1.4.4) of this dissertation. Given these aims, the present study rests at the interface of methodological and theoretical work in the field of phonetics.

In the following paragraphs, therefore, we will introduce some general background on methods (validation and generative sufficiency) and theory (sound change; Bavarian phonotactics). We will then introduce some aspects more specific to the present study (agent populations; validity criteria), describe the ABM_{IPS} (the simulation model under investigation; intake strategies) and finally describe what we are testing in the present study (dispersion of duration distributions; manipulated parameters and scenarios; hypotheses).

Validation and generative sufficiency A valid agent-based simulation model of sound change demonstrates computationally that a given theory and the *microstructures* it describes – the details of how humans perceive and produce speech and how speakers both within and between different varieties interact with each other – in fact lead to a certain *macrostructure*. The macrostructure, in our case, is the phonological system that we observe in our groups of participants (or, more precisely, the acoustic parameters we measured in the participants). In Chapter 4, we have observed this phonological system to have changed between two groups of participants that represent different generations of dialect speakers. We will test whether the model, in a variety of configurations, predicts this change by simulating interactions between the older group of dialect speakers and a group of Standard German speakers. The Standard German group served as a control group in Chapter 4, but in the present chapter, this role changes. The role of

[1] With ongoing work that is not yet published.

control group is now assumed by the younger dialect speakers, whose acoustic features represent the "ground truth" that the older dialect group is expected to develop towards. The new control group does not take part in the simulated interactions.

The property of a model that correctly predicts the described between-group change is called generative sufficiency (Epstein, 2006). While generative sufficiency does not compare the underlying theory of the model to alternative explanations (like many experimental designs do), it does show in a very compelling way that the underlying theory is strong enough to explain the observations made about its referent system (the sound change in progress). Further below, we will develop criteria to decide whether or not a model can be considered as predicting the change.

Sound change The present study is concerned with sound change. The big questions in this field of research include why sometimes sounds change while more often they do not; why certain sounds change in some languages but remain the same in other languages; what new sounds they change into; and what must happen for a change to spread through the language community (and be adopted by many speaker-listeners after it initially occurred in only a few or even a single one of them) (cf. Garrett & Johnson, 2013; Harrington, Kleber, Reubold, Schiel, et al., 2019; Labov, 1994). Sound change, and in fact language in general, has been termed "a phenomenon of the third kind" (Keller, 1990): something that is the result of human action, although it was not planned by any individual. It is therefore neither purely natural nor purely artificial. It is rather a consequence of collective action. To simulate this kind of phenomenon, agent-based simulations are particularly apt, and more so than other types of simulations – because actions at the level of an individual are explicitly modeled as agents, while an agent population can represent a macrostructure independent of the individual level, a "phenomenon of the third kind," that just emerges.

This chapter is focused on a particular sound change in a dialect of German, namely Western Central Bavarian (henceforth, simplified, *Bavarian*), which has been observed to be currently in progress (Kleber, 2017a; Moosmüller and Brandstätter, 2014; Chapter 4 of this dissertation). This sound change affects the phonotactics and the status of phonemic vowel length in said dialect. The change is an instance of dialect levelling, that is, a dialect losing some of the features that distinguish it from the surrounding standard variety and instead adopting that standard variety's features (cf. Kerswill, 2003; Trudgill, 1986). The available laboratory-phonological evidence was collected using an apparent-time approach (Bailey et al., 1991; Labov, 1963). It suggests that some features (see below for details) in the Bavarian phonological system have changed and now resemble the phonological system of Standard German as a consequence of interactions between the two language communities over the decades. The evidence further suggests that the major driving force that triggered the change was borrowing from Standard German into Bavarian, rather than phonetic bias in the Bavarian linguistic system itself (see Chapter 4 of this dissertation).

Bavarian phonotactics The affected dialect feature is the distribution of vowel length and consonant strength within one word. Many phonological accounts of Bavarian have described that long vowels only occur before lenis consonants and short vowels only before fortis consonants (Hinderling, 1980; Pfalz, 1911; Seiler, 2005). This has become known as "Pfalz's law." The laboratory data available for this study covers three groups of speakers:

A) Dialect speakers aged 50 and above, who implement Pfalz's law in their speech production.

B) Dialect speakers aged 20 to 30, who do not implement Pfalz's Law and in some respects (e. g. stop closure duration, voice onset time) fall measurably between group A and group C.

C) Speakers of (regional) Standard German, aged 20 to 30 and not subject to Pfalz's law in the first place.

Note that this characterization of groups A and B with regard to Pfalz's law only holds true for short vowels: they occur before any kind of consonants in group B, but only before fortis consonants for group A. Long vowels, however, occur before any kind of consonant for both groups. This indicates that even the older generation is probably too young to be completely unaffected by the sound change in progress.

Agent populations In our validation of ABM_{IPS}, we employ two populations of agents. One is initialized with the data from the older dialect speakers (group A); the other with the data from the younger standard speakers (group C). Having these agent populations interact with each other, then, should model the interaction between the standard and the dialect community seen in the past decades. One of the most important independent variables is a measure of how much the populations interact with each other and whether the amount of input is symmetrical or asymmetrical. The dependent variables of interest should demonstrate how the phonological systems of the agent populations change during the course of thousands of interactions.

Validity criteria We will consider a simulation model quantitatively valid if the phonological system of the dialect agent population resembles that of the younger dialect group (group B of real speaker-listeners) after the simulated interactions. We also consider it quantitatively valid if this resemblance is reached after a certain number of interactions and then the phonological system of the dialect agent population changes further. We will consider a simulation model quantitatively invalid if the state of group B is never reached.

Since a model's validity cannot necessarily be determined in a binary decision (Epstein, 2006), we will consider a simulation qualitatively valid if the dialect population

changes towards the standard population while the standard population remains stable. A less strict variant of this validity criterion requires that the model exhibit the general tendency that the populations change in an asymmetrical manner, with the dialect population changing more towards the standard population than vice versa.

The simulation model under investigation [2] To simulate sound change, the ABM_{IPS} implements virtual speaker-listener agents whose systems for speech perception, phonology and speech production are modeled around exemplar theory (e. g. Pierrehumbert, 2003) and the interactive-phonetic sound change model (Harrington et al., 2018). The literature on self-organizing systems in phonetics dates back at least as far as Liljencrants and Lindblom (1972). Over the past twenty years, an increasing number of phoneticians have published variants of agent-based models to investigate sound change. Some authors, such as de Boer (2000), have taken the approach to synthesize speech as input for the agents. In the ABM_{IPS}, the speech input is taken from laboratory phonology corpora. The models also differ, for example, in whether groups of agents talk to each other or an individual agent talks to itself; or in whether and how tokens are removed from the agents' memory. A very good, extensive overview of agent-based modelling in sound change research is presented in Harrington, Kleber, Reubold, Schiel, et al. (2019).

Simulation models must be fully specified, and the only complete description of their specification is, by definition, the code itself. This is because the code may have side effects not intended or even known by its authors and users. The following is a high-level description of what the ABM_{IPS} is intended to do.

Each agent has an *exemplar cloud*, which is a memory of *phoneme exemplars*, each of which has three properties: (1) a phoneme label (e. g. /d/), (2) a label of the word this token of the phoneme appeared in (e. g. *Pudding*), and (3) a low-dimensional (sometimes unidimensional) feature vector describing acoustic features of the phoneme (e. g. closure duration in milliseconds (ms), voice onset time in ms, duration of the preceding vowel in ms). Before starting the simulation, each agent's memory is initialized and filled with a collection of tokens extracted from an individual, real speaker. Then, optionally, the memory is enlarged via a process called memory resampling (see Section 5.2.3).

An interaction is such that one token of a word is exchanged between an agent-speaker and an agent-listener. The two agents for an interaction are chosen at random, but adhere to a configurable general principle. The agents can be configured to only interact within their own population, or only with the other population, or some in-between strategy (see *manipulated parameters and scenarios* below).

The exchange of a word token consists of a modeled speech production process on the part of the agent-speaker and a modeled speech perception process on the part of the agent-listener. For the speech production process of a given word, the agent-speaker

[2] The code of the ABM_{IPS} is subject to constant work and improvement. The description of the version used in this chapter is based on personal communication with Johanna Cronenberg, Michele Gubian, and Jonathan Harrington.

considers all tokens of the respective word in its own memory. It estimates a Gaussian distribution of the acoustic features of all tokens and then creates a new sample based on this Gaussian. This sample will be the *new token*. This means that the produced token will not be one that existed before. The new token is then passed to the agent-listener, including the phoneme label, the word label and the feature vector. Passing the word and phoneme label means that the agent-listener always knows what category the agent-speaker intended to produce. Misperceptions are not a part of this model. The speech perception process is modeled somewhat differently, because it only considers the phoneme label but ignores the word label. Since misperceptions are not modeled, the relevant outcome of the perception process is whether or not the perceived token will be included in the agent-listener's memory or exemplar cloud. The ABM_{IPS} implements a number of different ways to decide this and calls them *intake strategies*.

Intake strategies, asymmetry in phonological categories, and directionality in sound change Since the chosen intake strategy so directly affects the formation of the agents' exemplar clouds, it is a very important variable in modeling a particular theory of sound change. The simplest intake strategy is called *acceptAll* in the ABM_{IPS}. Here, agent-listeners update their exemplar cloud with every token of a phoneme they perceive.

The important intake strategy in this paper, however, is called *mahalanobisDistance*. It reflects a central idea that has driven the development of the interactive-phonetic model of sound change and with it the ABM_{IPS} (Harrington et al., 2018; Harrington & Schiel, 2017). The idea emphasizes that phonological categories have a certain orientation, both with respect to other phonological categories within the same speaker's inventory, and with respect to the same phonological category in another speaker. This orientation can be captured in the Mahalanobis distance and is often asymmetric as illustrated in Figure 5.1. The figure shows two phonological categories, A and B (exemplified twice: in a unidimensional acoustic space, and in a two-dimensional acoustic space). The Mahalanobis distance is different when measuring from A to B or from B to A, because it depends on the dispersion; in the two-dimensional case, on the dispersion in the direction of the dashed line. B's Mahalanobis distance from A is inversely correlated with A's (small) dispersion and is therefore larger than A's Mahalanobis distance from B, which is inversely correlated with B's (large) dispersion.

This is why with the *mahalanobisDistance* intake strategy, a token that lies exactly in the middle between two categories may be accepted into the exemplar cloud of one category but not the other. And this, in turn, may give rise to directionality in sound change: over the course of time, category B accepts more and more tokens from category A, thereby moving the whole cloud towards A's cloud, while category A never accepts any token from B, thereby keeping its cloud in place. Note that the two clouds may be within the same speaker (i. e. one phoneme changing towards another) or in different speakers (i. e. one speaker adopting the acoustics of another speaker).

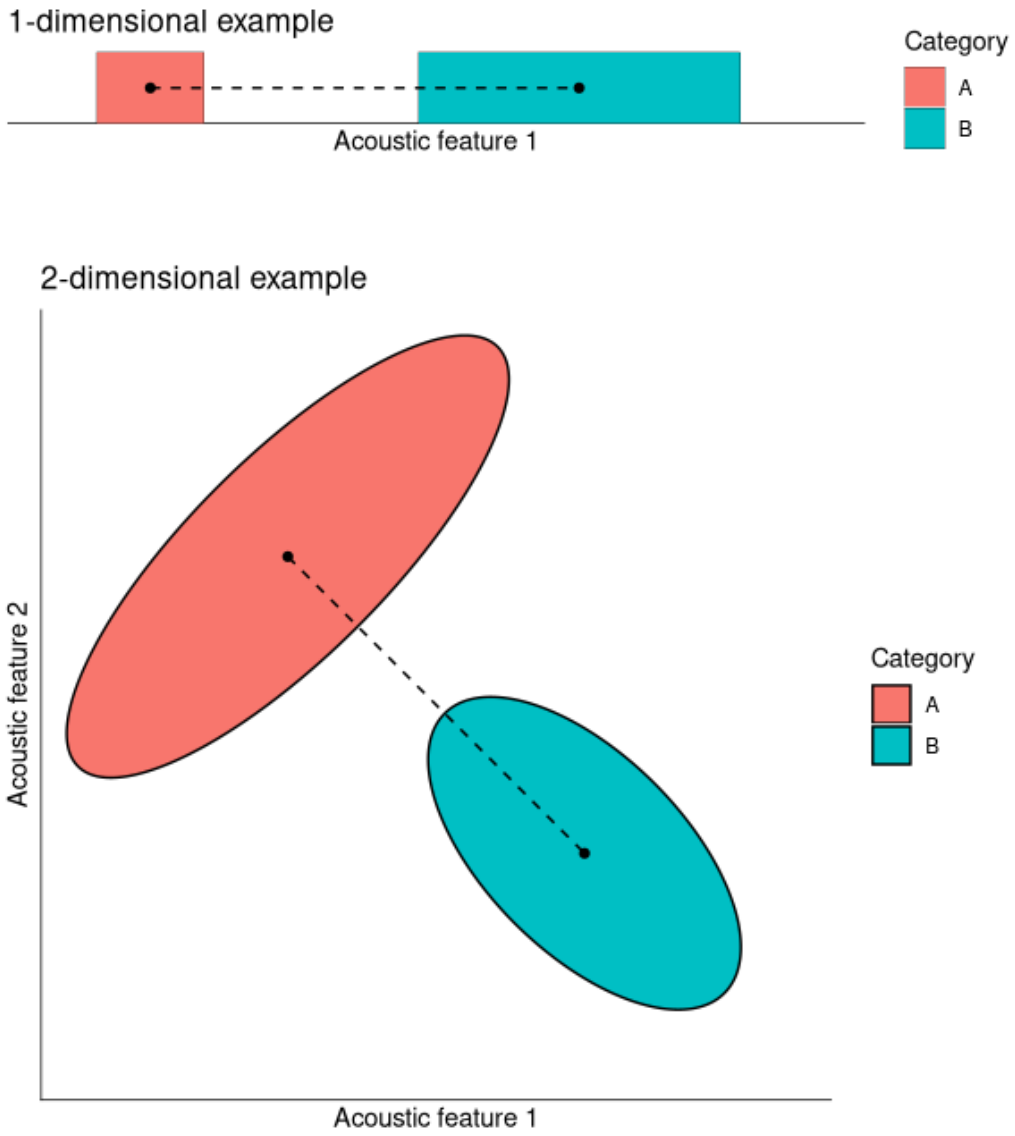

Figure 5.1: Two examples (one unidimensional, the other two-dimensional) of phonological categories A and B with a certain orientation with respect to each other. The dots represent the categories' centroids, the dashed lines represent the distance between those centroids. In the unidimensional example, A has a lower dispersion than B. In the two-dimensional example, A and B have similar dispersions, but A has a lower dispersion *in the direction of B* than B in the direction of A.

Dispersion of duration distributions As described above, the interactive-phonetic model (IP) relies on one distribution having a higher dispersion than another. In a unidimensional acoustic space with two categories (which is what we will model, see below for details), it predicts that the more highly dispersed category will change towards the less highly dispersed one. For durations of sounds, many studies have found that longer durations are associated with higher dispersion than shorter durations (see Rosen, 2005 for a discussion). This would appear to be a natural tendency: in a phoneme that is 200 milliseconds (ms) long on average, speakers can most likely produce a deviation of, say, 10 ms without the risk of the sound becoming unnatural or even crossing the boundary of another category. For a 50 ms sound, a 10 ms deviation is much more likely to produce such an effect. This means that when the ABM_{IPS} (using the *mahalanobisDistance* intake strategy) is used with a durational contrast, it will mathematically always favor the short phoneme (because the short phoneme is associated with a smaller dispersion) and lead to the long phoneme becoming shorter over time. This may, in fact, be a mechanism explaining sound changes where long phonemes are shortened (or "fortis" phonemes lenited, when the fortis–lenis distinction is based on durational cues). However, it may also be no more than a mathematical artifact. In fact, human perception has often been reported to respond to logarithmic rather than linear changes in stimuli (Rosen, 2005; Varshney & Sun, 2013). We are inclined to believe that human perception incorporates some kind of normalization for these mathematical differences in dispersion and that this normalization has to do with logarithmic transformation. This will affect our choice of manipulated parameters.

Manipulated parameters and scenarios In order to validate a simulation model, we need to compare its outcome to a set of *real*, i. e. non-simulated, data. This is only possible with real data that allow clear expectations about what the outcome of the simulation should be. If the interpretation of the real data is already ambiguous, the interpretation of the simulated data (with regard to the simulation's validity criterion) becomes nearly impossible, because there are too many degrees of freedom. We therefore chose to simulate only *stop closure duration*, the one parameter from Chapter 4 of this dissertation that allows the most unambiguous interpretation: stop closures in the word type under investigation have been shortened diachronically. We also considered other parameters investigated in Chapter 4 (e. g. voice onset time) and especially what would seem the most realistic choice: a combination of those cues (cf. Beddor, 2006). While the ABM_{IPS} technically allows cue combinations to be simulated, we decided against it for better interpretability.

Based on stop closure duration, we simulated a number of different scenarios. Three scenarios are concerned with the distribution of contact and input. The one we call *maximum contact scenario* is the simplest of them but at the same time the most unrealistic. At maximum contact, agents only interact with the opposite population and never with agents from their own population. The other two aim to be more realistic in terms of

input distribution. However, their exact construction must be considered exploratory for the lack of realistic numbers. In the *symmetric contact scenario*, all agents receive an equal amount of input from both populations. In the *asymmetric contact scenario*, the dialect population receives more standard input (40%) than the standard population receives dialect input (10%). The last simulated scenario can be called a *null contact scenario* because the agents only interact with other agents from their own population. We will base this simulation on the *fast speech condition* data described in Chapter 4, because the theoretical aim here is to see not the effect of language contact but rather the effect of speech rate, or more precisely, of fast-speech-induced hypoarticulation. All other scenarios will be based on the normal-paced condition.

Aside from varying the type of language contact, we also decided to run the simulations both on raw (i. e. linear-scaled) stop closure durations and on their natural logarithm, $ln(duration)$. Since we consider the logarithmic duration to be closer to a realistic perception apparatus, we expect logarithmic-scaled models to perform better in terms of the validity criteria than their linear-scaled counterparts.

Figures 5.2 and 5.3 show the distribution of closure duration in our dialect and standard speakers, who, as per our design, represent the input for our agent populations. They clearly show a strong between-group asymmetry in the dispersion of linear-scaled stop closure durations. This type of asymmetry is the key mechanism by which the interactive-phonetic model of sound change and the ABM_{IPS} predict change. In the log-scaled variant, however, this asymmetry is very weak. This may result in a bad performance of the log-scaled models, despite our theoretical expectation that the log-scaled models perform better.

Hypotheses Our main hypothesis is that the ABM_{IPS} will meet the above-mentioned *validity criteria* by reproducing a pattern derived from observations in Chapter 4: that after a large number of interactions between two agent populations (representing standard speaker-listeners and older dialect speaker-listeners, respectively) the dialect agent population resembles the acoustics found in the reference group of younger dialect speaker-listeners (i. e. the dialect agents will have much shorter stop closures than before). This younger group is not part of the simulation and thus has no corresponding agent population. The standard agent population is not expected to change. In the speech rate (or *null contact*) simulation, applying the a priori assumption that underlies the experimental design from Chapter 4 would predict that the dialect agent population will reach the validity criterion even by interacting only with themselves. However, since the laboratory findings from Chapter 4 did not support this, our hypothesis here is that the simulation will not bring about any change in stop closure duration when agents only interact with their own population.

Our second hypothesis is that the more contact we model, the faster this validity criterion will be reached; that is, fewer interactions should be required in the *maximum contact scenario* than in the *symmetric* and *asymmetric contact scenarios*.

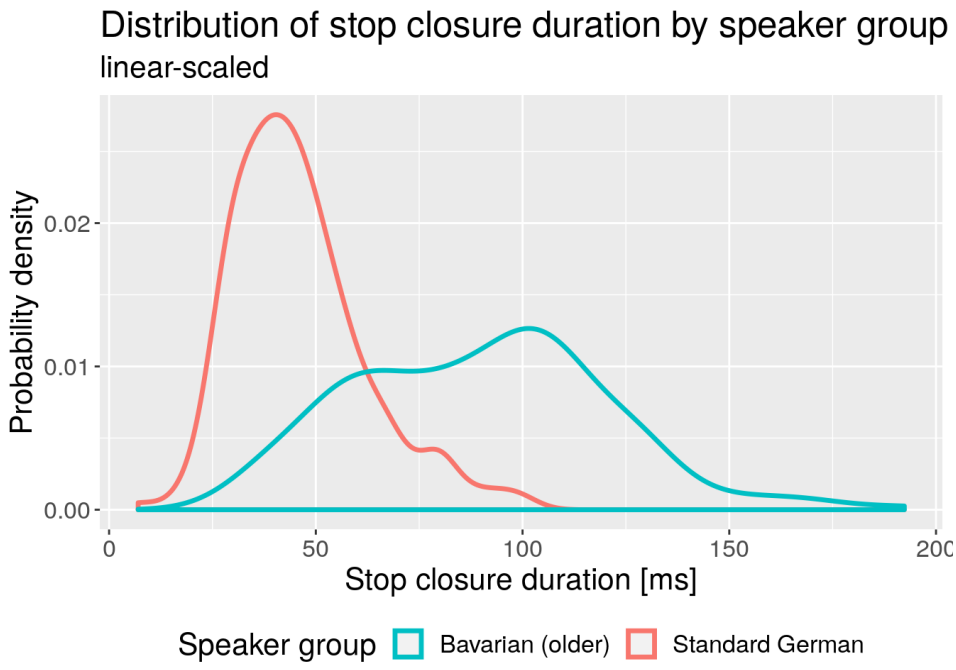

Figure 5.2: Probability density function of the stop closure durations found in the real speakers that formed the input to our dialect agents. Linear-scaled.

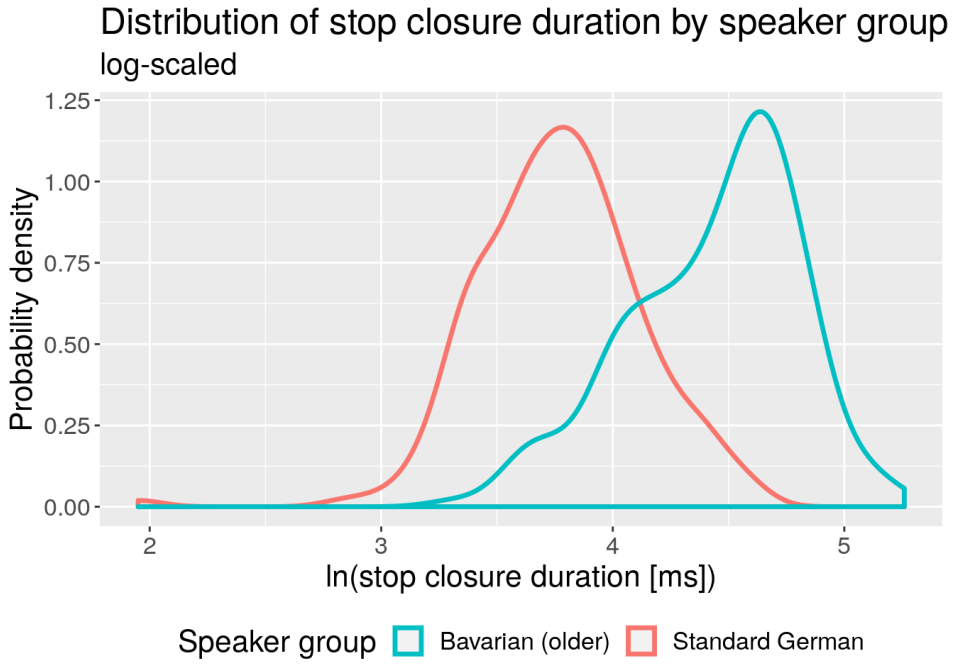

Figure 5.3: Probability density function of the stop closure durations found in the real speakers that formed the input to our dialect agents. Log-scaled.

Our last expectation is that models with logarithmic duration values will perform better than models with linear duration values, assuming that the logarithmically transformed values are perceptually more relevant.

5.2 Method

5.2.1 Participants and agents

The models were run with 20 virtual agents distributed over two populations: dialect and standard. The agents were initialized with the same speech recordings described in Chapter 4 of this dissertation. Each agent was initialized with the speech production data from one real speaker. The ten real speakers represented by the dialect agent population were aged 49 and above (mean 60.5, standard deviation 8.15). The ten real speakers represented by the standard agent population were aged 19-30 (mean 24.1, SD 3.51). The simulations' outcomes were compared against a group of ten dialect speakers aged 20-29 (mean 25.3, SD 2.91). The dialect speakers were selected from the Western Central Bavarian dialect region, particularly from the regions (*Landkreis*) of Altötting, Mühldorf and Miesbach. For the regional standard group, we selected speakers who did not speak dialect according to their own assessment.

5.2.2 Materials

We used recordings of the words *Pudding* 'pudding', *Widder* 'ram', *Rabbi* 'rabbi', and *Tigger* 'proper name of a cartoon character'. The recordings were automatically segmented using MAUS (Kisler et al., 2016) and then the segment boundaries were corrected by hand. Compare the methods section in Chapter 4 for details on how the words were elicited from the speakers and how the manual corrections were carried out.

5.2.3 Procedure and choice of parameters

All the models were run with the ABM_{IPS}'s source code publicly available from GitHub[3]. The version used was git revision 3bc0dfb. The ABM_{IPS} is programmed in R. Section 5.5 contains the auxiliary scripts used to run the simulations.

Since the model contains many stochastic elements, we ran it between 7 and 10 times for every configuration we tested. Some configurations required more computation time and those were only repeated 7 times.

The *mahalanobisThreshold* parameter defines the largest Mahalanobis distance between a token and the agent-listener's distribution of the respective phoneme class that still leads to the agent accepting that token. In the present study, we chose 2.07, which

[3]https://github.com/ips-lmu/ABM

equals 85% probability mass in a Chi-Square distribution with one degree of freedom (commensurate with the fact that we tested a one-dimensional acoustic feature).

The initialization of agents is coupled with a process called *memory resampling* in the ABM_{IPS}. After each agent is fed the tokens from one real speaker-listener, its memory is enlarged by estimating a Gaussian distribution over each word class and then picking, at random, new samples from this distribution. These are then added to the agent's starting memory, such that the starting memory is formed by a combination of tokens that actually occurred in a real speaker-listener and tokens that were artificially sampled from the estimated Gaussian. In the present study, a resampling factor of 10 was chosen, which means that the starting memory has ten times the number of available real tokens. The memory resampling procedure is used in order to reduce the likelihood of fluctuations in the agents' development over time that can be caused by classes with few members.

The *forgettingRate* parameter defines how often tokens are deleted from the agents' memories. In the present study, we kept that at 0 (i. e. agents never forget a token), because we were not interested in comparing different models of how a real speaker-listener's memory loses information over time.

5.2.4 Interaction plots

The most important results of the simulations are presented in interaction plots. These plots capture how the groups' stop closure durations change over the course of the interaction. They can be thought of as box plots (without whiskers) as a function of time: the three blue lines represent the dialect agents' first quartile, median and third quartile of stop closure durations at each point in time. The three red lines represent the same for standard agents. The three black lines represent the values of the reference group (which are not affected by the simulation and thus remain constant). The dots represent sample values, the lines are interpolations. The unit for durations is milliseconds (ms) for the linear-scaled models and $ln(ms)$ for the logarithmic-scaled models.

5.3 Results

The maximum, asymmetric, and symmetric contact scenarios (Sections 5.3.1, 5.3.2, and 5.3.3) are initialized with the same data and therefore share the same starting values. Due to the initial *memory resampling*, however, the values plotted at interaction 0 in Figures 5.4–5.18 vary slightly between the simulations.

In the linear-scaled models, the median starting values before memory resampling (i. e. the stop closure durations measured in the real speaker-listeners) are 90 milliseconds (ms) for the dialect agent population and 43 ms for the standard agent population. The figures reflect how the values of these two populations change over the course of the interactions. The younger dialect speaker-listeners, represented in the plots by straight yellow lines, have a median of 48 ms. Their values are constant, because they serve as the reference group and do not form part of the simulation.

In the logarithmic-scaled models, the medians before memory resampling are 4.50 for the dialect agents, 3.76 for the standard agents, and 3.88 for the younger dialect speaker-listeners[4]. Note that the numerical difference on this scale is very small, even when we know that the values 3.76 and 4.5 represent substantial phonetic differences in stop closure duration (43 vs. 90 ms).

5.3.1 Maximum contact scenario

Linear-scaled models For the linear-scaled models, we have 10 repetitions, running over 100,000 interactions each. Figures 5.4 and 5.5 represent repetition 5 and repetition 10, respectively. They were chosen to illustrate the fairly small amount of variation within the ten runs. See Section 5.5 for a complete listing.

Commensurate with Figures 5.4 and 5.5, the two agent populations in the linear model grow closer to each other, with the dialect population changing much more than the standard population. At around 50,000 to 70,000 interactions, the two populations' medians become virtually indistinguishable. In Figure 5.5, their medians are then also indistinguishable from the younger dialect group. In Figure 5.4, the younger dialect group maintains a difference of about two milliseconds, which is of a negligible magnitude. It must be considered coincidence to find a match as good as the one in Figure 5.5.

All ten runs can be interpreted as reaching the validity criterion, since the dialect population ends up almost exactly at the value of the reference group. This is in line with our hypothesis. The standard agents exhibited a slight change, although no change was expected.

Logarithmic-scaled models For the logarithmic-scaled models, we also have 10 repetitions, running over 100,000 interactions each. Figure 5.6 represents repetition 8. The

[4]We report the logarithmic values as unitless; strictly speaking, their unit is $ln(ms)$.

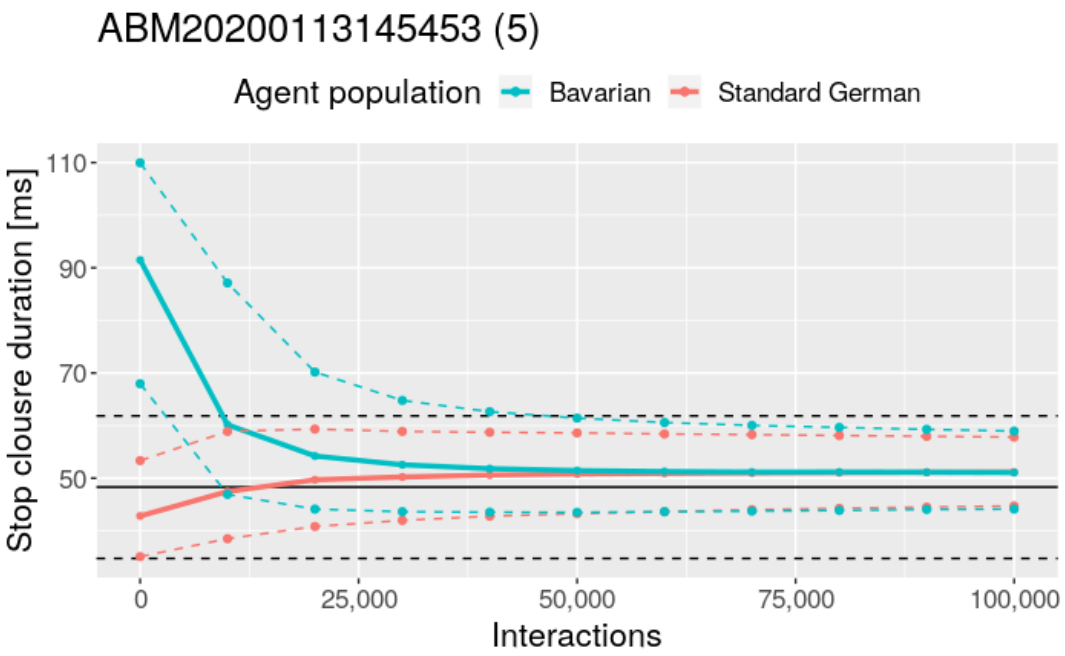

Figure 5.4: Interaction plot: maximum contact scenario, linear scale, repetition 5. Interaction plots are described in the methods in Section 5.2.4.

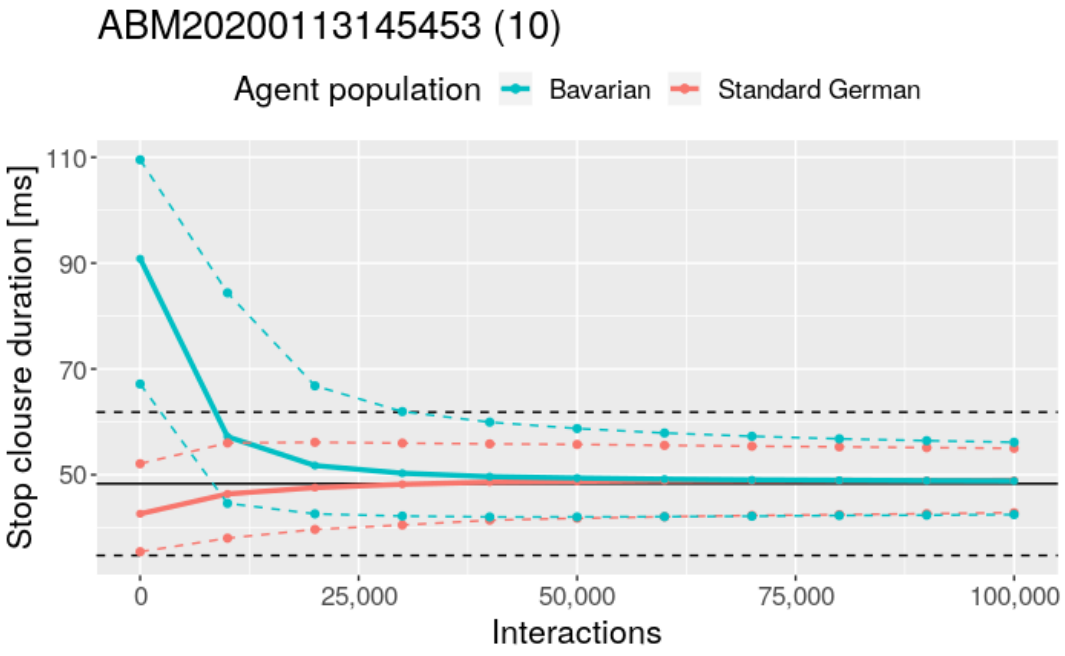

Figure 5.5: Interaction plot: maximum contact scenario, linear scale, repetition 10. Interaction plots are described in the methods in Section 5.2.4.

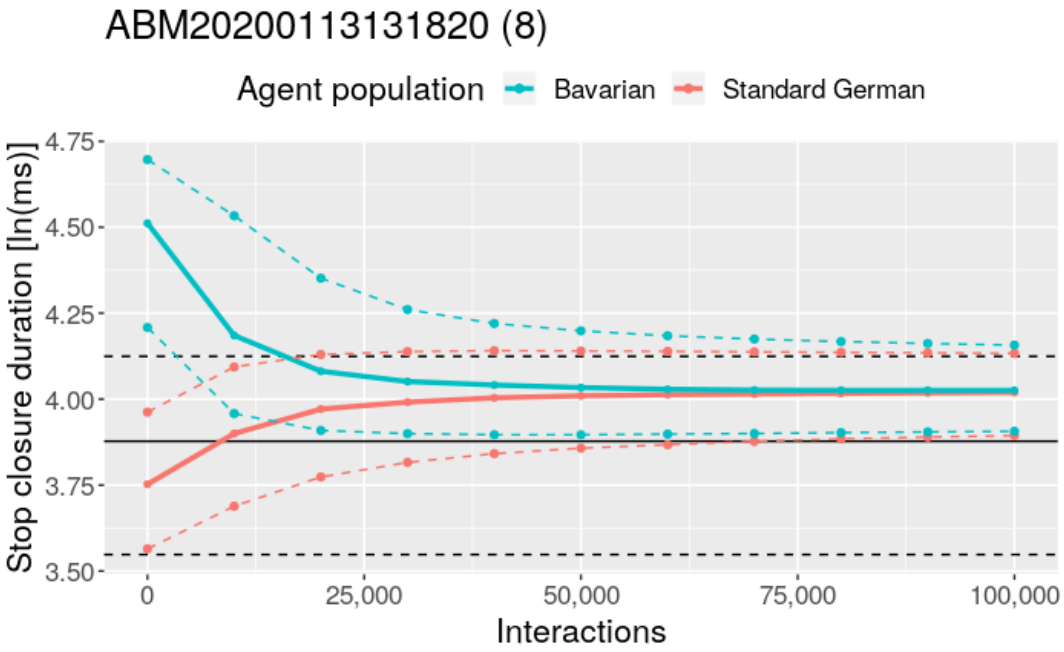

Figure 5.6: Interaction plot: maximum contact scenario, logarithmic scale, repetition 8. Interaction plots are described in the methods in Section 5.2.4.

repetitions exhibit a fairly small amount of variation. See Section 5.5 for a complete listing.

Figure 5.6 shows a logarithmic-scaled simulation run. The main similarities to the linear-scaled simulation runs depicted in Figures 5.4 and 5.5 are: (a) the two agent populations move toward each other, and (b) their medians become indistinguishable after around 50,000 to 70,000 interactions. The first main difference is that the changes in the two populations are now more symmetrical, with the standard population changing almost as much as the dialect population. However, they still meet just above 4.0, when a completely symmetrical course would have them meet around 4.125. This is consistent across all ten simulation runs with this set of parameters: in some runs, the meeting point is higher than in Figure 5.6, but it is never as high as 4.125. The other main difference is that the two populations' meeting point (4.0) is far away from the starting point of the reference group (3.875).

Do these models meet our validity criterion? The quantitative difference between the meeting point (just above 4.0) and the reference point (3.875) opposes such an interpretation. The qualitative pattern, however, does reflect our expectation of asymmetric change (i. e. the dialect population changes more toward the standard agent population than vice versa), although the effect is not far off from being symmetrical. It also does not meet the strict variant of our expectation, which would have meant no change in the standard agent population.

Figure 5.7: Amount of dialect and standard input, respectively, received by each agent in the asymmetric contact scenario (linear scale, repetition 1).

5.3.2 Asymmetric contact scenario

For illustration purposes, Figure 5.7 represents the amount of standard and dialect input, respectively, that each agent received in one run of the asymmetric contact scenario. The pattern confirms that our configuration of the model in this respect is consistent with our intentions. It is also consistent across all runs in this scenario.

Linear-scaled models For the linear-scaled models, we have 7 repetitions, each running over 3 million interactions. Their outcomes are not entirely uniform. Five of the seven yielded the qualitative pattern depicted in Figures 5.8 and 5.9, which we will call the "converging pattern," while two yielded the pattern depicted in Figures 5.10 and 5.11, which we will call the "non-converging pattern." See Section 5.5 for a complete listing.

In the converging pattern, the standard population barely changes at all, with the exception of a slight decrease of the median stop closure duration and a narrowing of dispersion at the beginning of the interactions. The dialect population, on the other hand, changes towards the standard population very steeply during the first approx. 300,000 interactions and then at a much slower pace after that. In one of the five runs exhibiting the converging pattern, depicted in Figure 5.8, the dialect population has almost reached the standard population after 3 million interactions. The other four runs exhibiting the converging pattern are qualitatively very similar to each other, with one of them being depicted in Figure 5.9: after 3 million interactions, the dialect population has reached

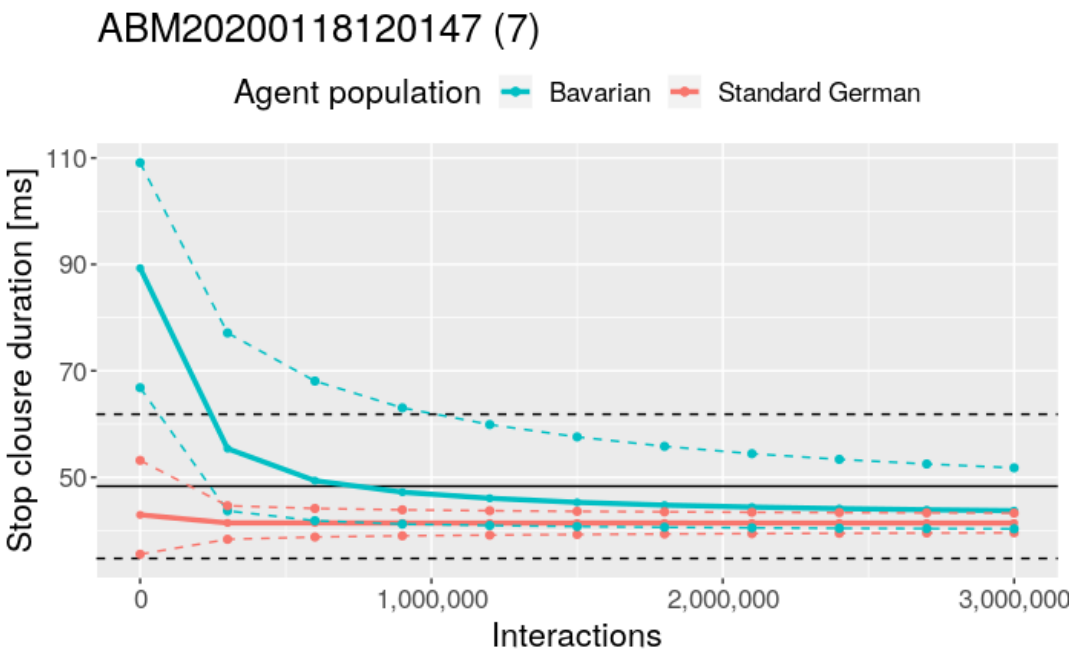

Figure 5.8: Interaction plot: asymmetric contact scenario, linear scale, repetition 7. Interaction plots are described in the methods in Section 5.2.4.

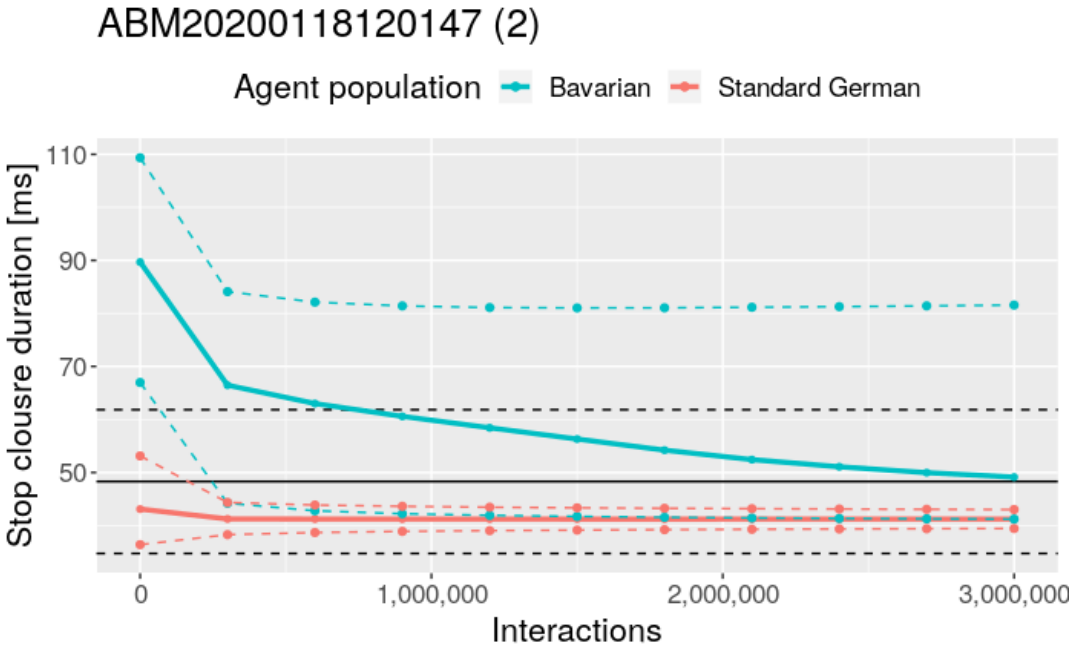

Figure 5.9: Interaction plot: asymmetric contact scenario, linear scale, repetition 2. Interaction plots are described in the methods in Section 5.2.4.

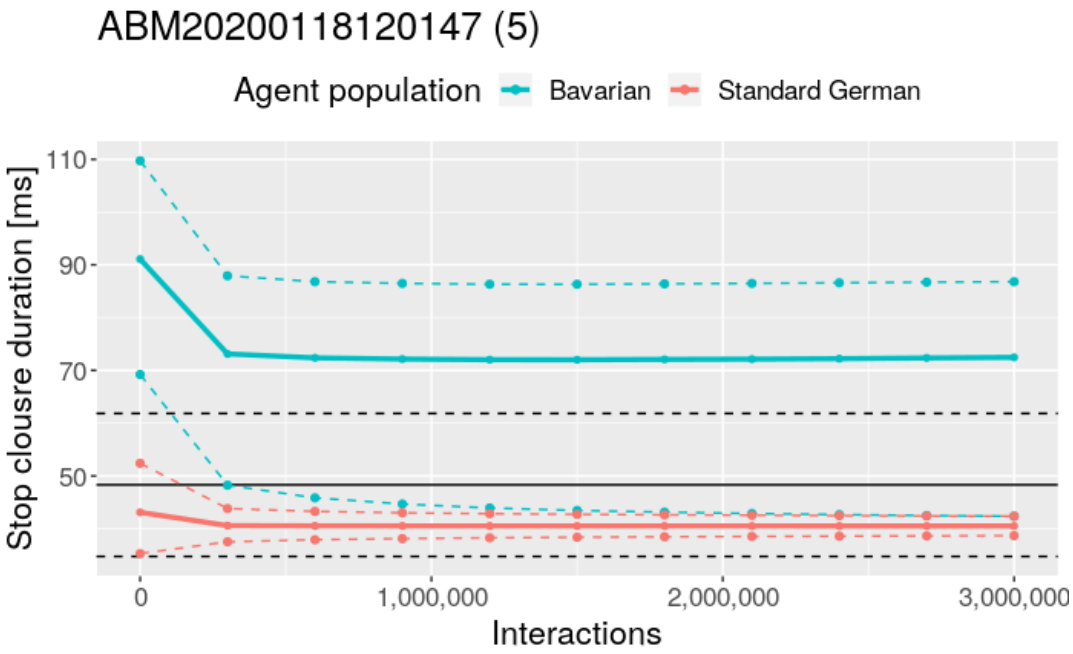

Figure 5.10: Interaction plot: asymmetric contact scenario, linear scale, repetition 5. Interaction plots are described in the methods in Section 5.2.4.

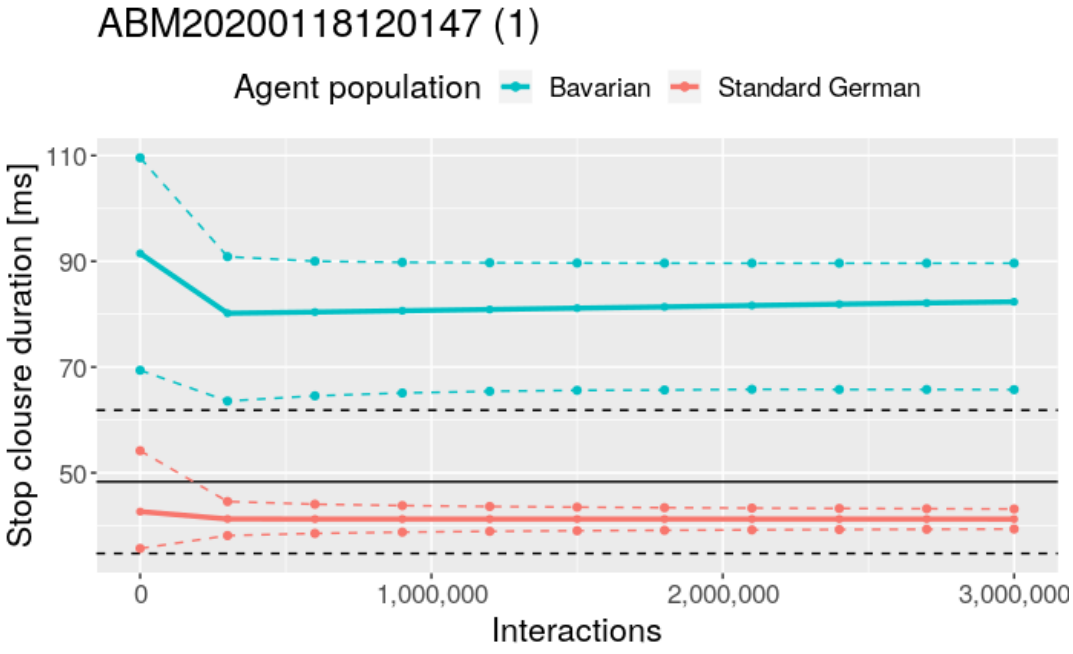

Figure 5.11: Interaction plot: asymmetric contact scenario, linear scale, repetition 1. Interaction plots are described in the methods in Section 5.2.4.

around 50 milliseconds closure duration and still has a clear downward slope. Had these simulations been run for more than 3 million interactions, the dialect population would almost certainly have exhibited the same pattern as in Figure 5.8 (and reached the standard agent population's median of about 41 ms). However, we chose not to do this because it would have cost an enormous amount of computational resources.

In the non-converging pattern, commensurate with Figures 5.10 and 5.11, the first approx. 300,000 interactions are similar to the non-converging pattern: the standard agents show a slight decrease in median stop closure duration and a narrowing in dispersion, while the dialect agents show a steep decrease in median stop closure duration. After that, however, the dialect population does not continue the decrease at a slower pace. Instead, it levels out and shows no further change for the remaining approx. 2.7 million interactions. Its median, at this point, is still far away from both the reference group and the standard population. The two repetitions exhibit one qualitative difference: while in Figure 5.10, the first quartile of the dialect population is about the same value as the standard population's third quartile, the dialect population's first quartile in Figure 5.11 is still far away from the agent population.

The converging pattern (i. e. 5 out of 7 repetitions) meets the validity criterion both quantitatively and qualitatively. The dialect population's median hits the value of the reference group. Unlike the maximum contact simulations, however, the dialect population does not stop there, but rather moves further towards the standard population. The standard population barely moves at all, which is in line with the strict variant of our asymmetry expectation. Moreover, the validity criterion is hit at a much later point than in the maximum contact scenario, which is also in line with our hypothesis.

The non-converging pattern does not meet our validity criterion to any substantial extent. In line with our expectations, the standard population does not change. The dialect population does change, and even in the correct direction. However, the dialect population's sudden stop, especially at a value far away from the reference group, was not predicted.

Logarithmic-scaled models For the logarithmic-scaled models, we have 8 repetitions running over 1 million interactions each. The repetitions exhibit a fairly small amount of variation. Figure 5.12 represents repetition 1. See Section 5.5 for a complete listing.

Commensurate with Figure 5.12, neither of the two agent populations in these models exhibits substantial change. Both show a narrowing in dispersion, and the dialect population shows a minor drop in stop closure duration during the first interactions. This drop is very small, yet very consistent across all repetitions.[5]

This pattern is very different from our expectations and very different from all the other results presented so far. It does not meet our validity criterion in any way.

[5] At the time resolution provided in Figure 5.12, the magnitude of the drop is slightly underestimated. A better time resolution shows that the curve drops down to between 4.4 and 4.8 (depending on the repetition) after around 50,000 interactions (in all repetitions) before rising again.

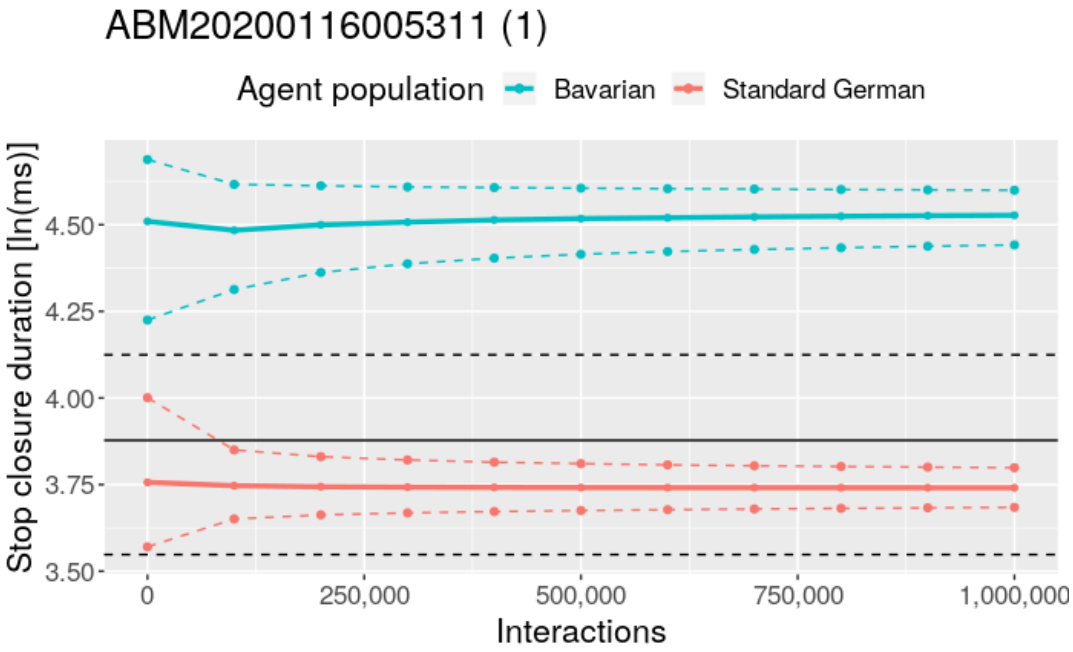

Figure 5.12: Interaction plot: asymmetric contact scenario, logarithmic scale, repetition 1. Interaction plots are described in the methods in Section 5.2.4.

5.3.3 Symmetric contact scenario

For illustration purposes, Figure 5.13 represents the amount of standard and dialect input, respectively, that each agent received in one run of the symmetric contact scenario. The pattern confirms that our configuration of the model in this respect is consistent with our intentions. It is also consistent across all runs in this scenario.

Linear-scaled models For the linear-scaled models, we have 7 repetitions, running over 3 million interactions each. The repetitions exhibit a fairly small amount of variation. Figure 5.14 represents repetition 1. See Section 5.5 for a complete listing.

In this scenario, commensurate with Figure 5.14, the dialect agent population changes towards the standard agent population. It reaches the reference point after about 500,000 interactions and then goes on to move further towards the standard agent population.The standard agent population, on the other hand, only has a small increase in stop closure duration at the beginning of the interactions but then does not move much further. The dialect population does not change all the way towards the standard agents by the end of the 3 million interactions, but gets extremely close after around 1 million to 1.5 million interactions. The two populations would probably meet at some point beyond 3 million interactions.

These models meet our validity criterion and the strict version of our asymmetry expectation, with the standard agent population hardly changing at all. They reach the

Figure 5.13: Amount of dialect and standard input, respectively, received by each agent in the symmetric contact scenario (linear scale, repetition 1).

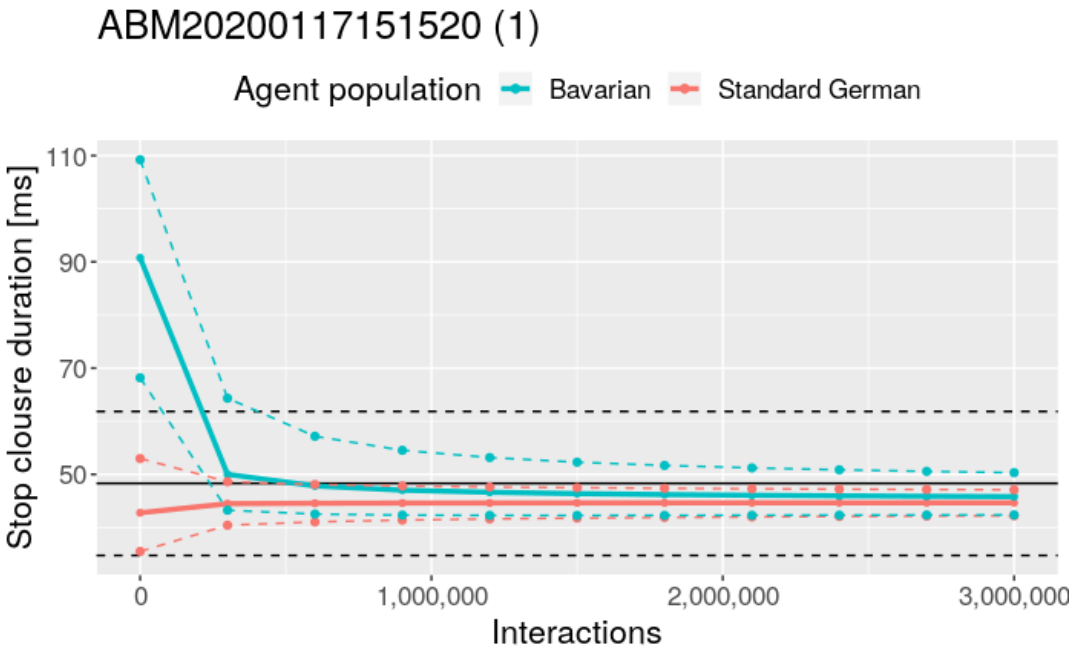

Figure 5.14: Interaction plot: symmetric contact scenario, linear scale, repetition 1. Interaction plots are described in the methods in Section 5.2.4.

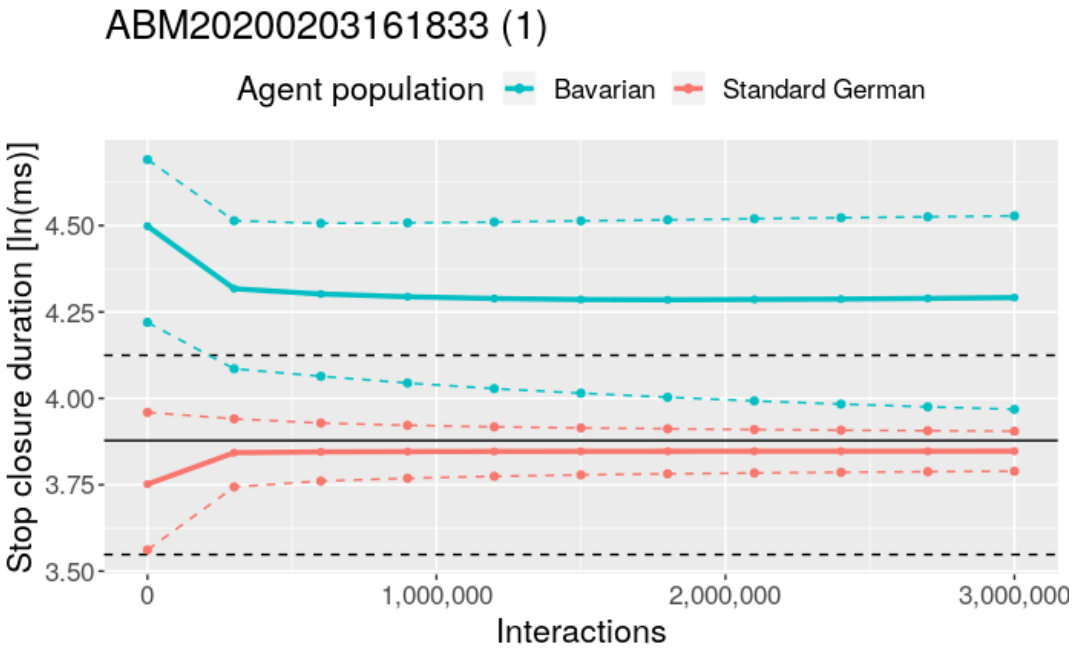

Figure 5.15: Interaction plot: symmetric contact scenario, logarithmic scale, repetition 1. Interaction plots are described in the methods in Section 5.2.4.

validity criterion much later than in the maximum contact scenario, which is in line with our expectations.

Logarithmic-scaled models For the logarithmic-scaled models, we also have 7 repetitions, running over 3 million interactions each. The repetitions are not entirely uniform. Moreover, they are problematic to evaluate, because they do not stabilize within 3 million interactions. Running them for more interactions was beyond the scope of this study.[6] See Section 5.5 for a complete listing of all 7 repetitions.

Four of the seven repetitions exhibit the qualitative pattern depicted in Figure 5.15: the dialect agents started out with a strong downward slope in the first approx. 300,000 interactions, but after that, the change almost leveled off, with the median still far away from that of the reference group. The first quartile of the agent data, however, continues to decline slowly, even after 3 million interaction. The standard agents start out with an upward slope in the first approx. 300,000 interactions and level off after that.

The other three repetitions exhibit the qualitative pattern depicted in Figure 5.16. Here, the dialect agents also start out with a strong downward slope for the first ap-

[6]The simulation as presented here took about 53 hours to complete. Runtime increases linearly with number of interactions and is mostly tied to the available CPU power. The CPU used was an Intel i7 (2018 model), which is currently among the most powerful available in terms of single–thread performance.

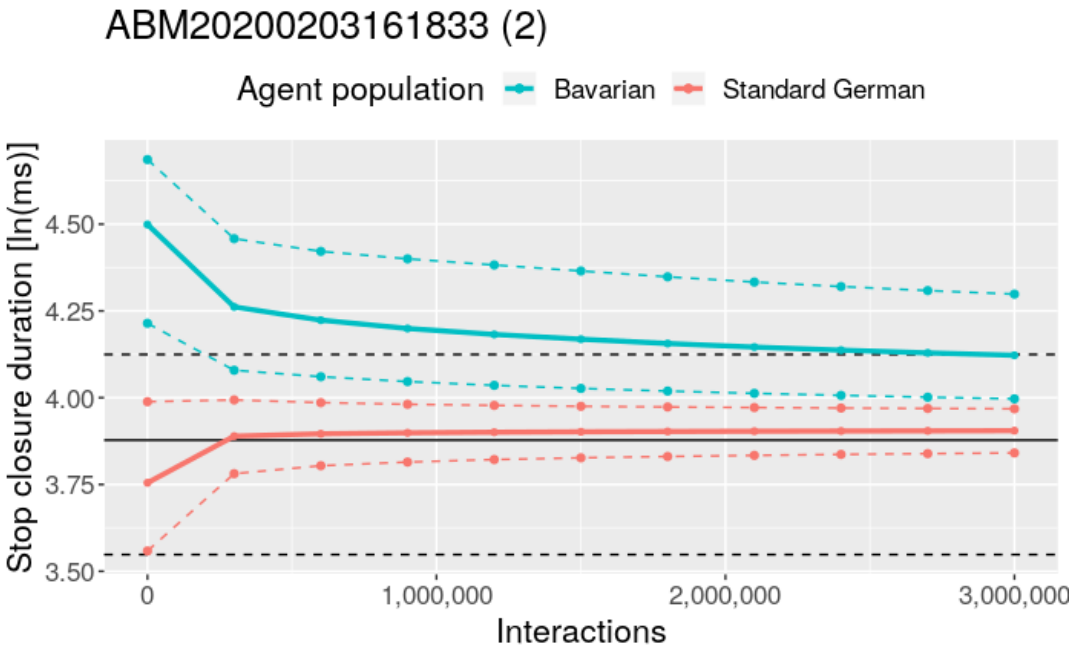

Figure 5.16: Interaction plot: symmetric contact scenario, logarithmic scale, repetition 2. Interaction plots are described in the methods in Section 5.2.4.

prox. 300,000 interactions. After that, the slope becomes less steep, but the downward trend continues even after 3 million interactions. The standard agents exhibit the same behavior as in the other four repetitions.

Within the 3 million interactions, none of the 7 repetitions hit quantitative validity. It remains unclear whether any of them would do so after more interactions. It appears likely that the 3 repetitions similar to Figure 5.16 would, while the four repetitions similar to Figure 5.15 would not. The set of three would probably also meet qualitative validity (not the strict form, though). No decision can be made about qualitative validity in the set of four.

5.3.4 Speech rate (null contact scenario)

Unlike in the other three scenarios, the agent populations in the null contact scenario were initialized with data from the "fast speech" experimental condition. Thus, the median starting points before memory resampling are 85 ms for the dialect agents, 40 ms for the standard agents, and 42 ms for the reference group. In the logarithmic-scaled models, this is 4.44 for the dialect agents, 3.68 for the standard agents, and 3.74 for the reference group.

For both the linear-scaled and the logarithmic-scaled models, we have 10 repetitions, running over 100,000 interactions each. Both sets exhibit a fairly small amount of

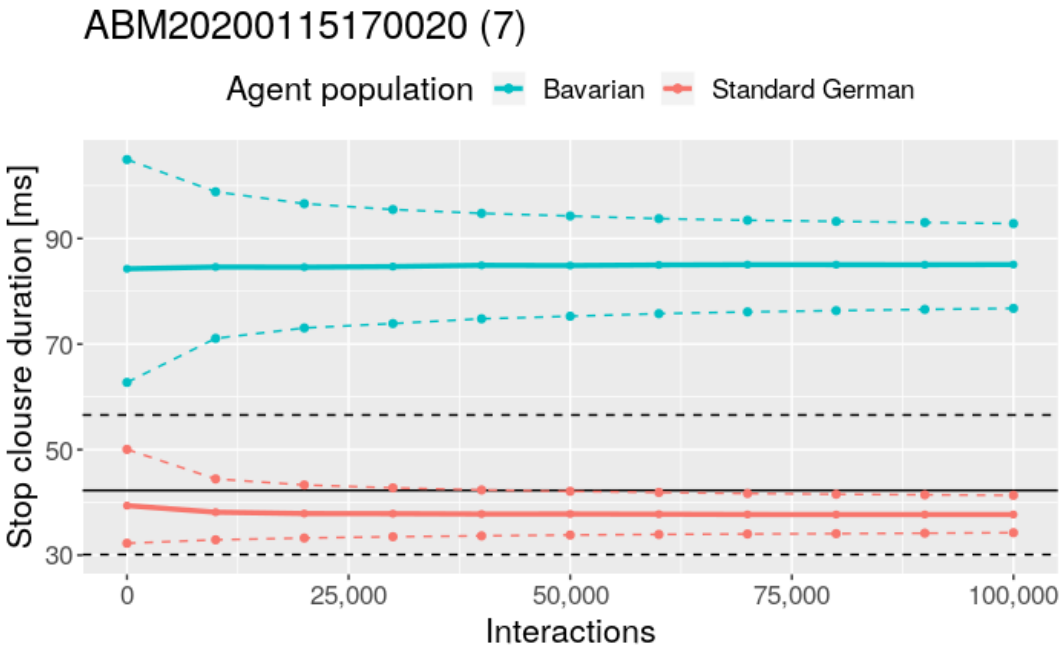

Figure 5.17: Interaction plot: null contact scenario, linear scale, repetition 7. Interaction plots are described in the methods in Section 5.2.4.

variation across the repetitions. Figure 5.17 shows the 7th repetition of the linear-scaled model, while Figure 5.18 shows the 7th repetition of the logarithmic-scaled model. See Section 5.5 for a complete listing.

In this scenario, agents only interacted with other agents of their own group and never with agents of the other group. Commensurate with Figures 5.17 and 5.18, this does not lead to any substantial change in the populations' stop closure durations. There is only a narrowing in dispersion (in both agent populations in both the linear and the logarithmic model).

These data do not meet our validity criterion in any way, which is in line with our expectations.

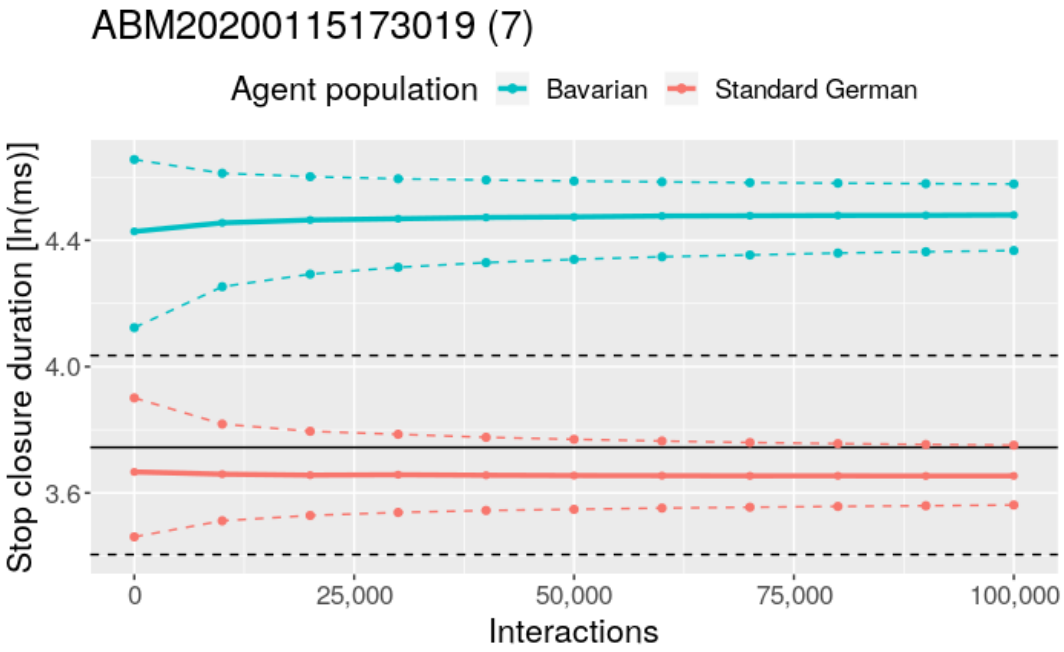

Figure 5.18: Interaction plot: null contact scenario, logarithmic scale, repetition 7. Interaction plots are described in the methods in Section 5.2.4.

5.4 Discussion

The first of two main aims of this study was to test how well the ABM_{IPS} would predict the Bavarian sound change described in Chapter 4 (i. e. the model's generative sufficiency) and whether some types of configuration would yield a stronger prediction than others. The second main aim was to explore what kinds of findings we could extrapolate from a relatively novel simulation model (i. e. one that has only seen a limited number of validation tests) to phonetic theory, as well as what hypotheses for future research we could generate using the model. We compared four different language contact scenarios (maximum, asymmetric, symmetric, and null contact), each tested with linear-scaled and logarithmic-scaled stop closure durations, resulting in a total of eight configurations of the ABM_{IPS}. Each configuration included one population comprising 10 Standard German agents and one population comprising 10 dialect agents; and each configuration was run for a total of 7 to 10 repetitions.

The first hypothesis was that the Bavarian sound change would be predicted in the maximum, asymmetric, and symmetric contact scenarios, but not in the null contact scenario. The second hypothesis was that in the maximum contact scenario, the sound change would be established after fewer interactions than in the asymmetric and symmetric contact scenario. The third hypothesis was that the logarithmic-scaled models would perform better than the linear-scaled models. The key results of all eight models are summarized in Table 5.1. The main findings are as follows:

A) The ABM_{IPS} predicted the sound change in the linear-scaled versions of the maximum, asymmetric, and symmetric contact scenario, but not the null contact scenario, all of which is in line with our expectations.

B) In line with our expectations, the models in the maximum contact scenario were by far the quickest to reach validity. They predicted the sound change in less than 70,000 interactions, while the asymmetric and symmetric models required 300,000 interactions at the very least, but more typically ranging in the millions.

C) Contrary to our expectations, the logarithmic-scaled models in the maximum, asymmetric and symmetric contact scenario performed worse than their linear-scaled counterparts. In the maximum and symmetric contact scenarios, the logarithmic-scaled versions predicted the general quality that the dialect changes more towards the standard than vice versa, but the effect was not far off from symmetry.[7] In the asymmetric contact scenario, the logarithmic-scaled model did not predict the sound change at all.

[7] NB: Symmetry in how much the two populations change. This is not to be confused with what symmetry means in the term "symmetric contact scenario," where it refers to symmetry in how much input the populations receive from the other population.

Contact	**Maximum**		**Asymmetric**		**Symmetric**		**Null**	
Scale	Linear	Log	Linear	Log	Linear	Log	Linear	Log
Quantitative validity Dialect agents' median reaches reference group's	✓	✗	CP: ✓ NP: ✗	✗	✓	?	✗	✗
Qualitative validity Populations are asymmetric, i. e. dialect agents change more than standard agents	✓	slightly	CP: ✓ NP: ✓	✗	✓	?	✗	✗
Strict qualitative validity Standard agents remain constant	✗	✗	CP: ✓ NP: ✓	n.a.	✓	?	n.a.	n.a.
Number of interactions to reach validity	$<$ 70,000	$<$ 70,000	CP: 600,000–3,000,000 NP: 300,000	n.a.	500,000	$>$3,000,000	n.a.	n.a.

Table 5.1: Summary of key results of the eight models presented in section 5.3. ✓ denotes "yes," ✗ denotes "no." Numbers are approximate. CP and NP denote "converging pattern" and "non-converging pattern," respectively (cf. Section 5.3.2). "n.a." denotes "not applicable" and is used (a) for strict qualitative validity for those models that do not reach (simple) qualitative validity and (b) in place of a number for those models that meet no validity criterion. "?" is used for models that did not stabilize even after 3,000,000 interactions.

D) In the linear-scaled models of the asymmetric contact scenario, the ten repetitions resulted in two different outcome patterns. The same is true of the logarithmic-scaled models of the symmetric contact scenario, but those additionally did not stabilize during the 3 million interactions they have been run for. For all other variants, the results were stable across repetitions.

We will now discuss these findings against the background of what each of them means for the ABM_{IPS} on the one hand and for phonetic theory on the other hand.

5.4.1 Logarithmic-scaled vs. linear-scaled models

All simulations in the present study are based on the assumptions that form the interactive-phonetic (IP) model of sound change[8] (Harrington et al., 2018). They all add the assumption that closure duration is a defining acoustic property of the stops in the languages in question. Half of the simulations, then, assume that the natural logarithm of closure duration instead of closure duration proper is that acoustic property. Our expectation was that the logarithmic-scaled models would be a better approximation of reality and would therefore better explain the observed data. However, the opposite was the case. The logarithmic-scaled models performed much worse than the linear-scaled ones, failing almost en bloc to predict any aspect of the Bavarian sound change (i. e. they met neither the quantitative nor the qualitative validity criterion). Figures 5.2 and 5.3 already showed that the method of log-transforming all data before the models are run leads to a strong reduction of between-group asymmetry in the dispersion. The ABM_{IPS}, however, relies on between-group asymmetry to predict a change; the results suggest that the remaining asymmetry in the log-scaled data is too weak. At this point, we will carefully try to distinguish what conclusions this allows us to draw about reality and about the model, respectively. This requires considering a variety of aspects.

Development of the ABM_{IPS}

At first, the discussion will remain within the constraints of the model before moving on to phonetic theory later. Under the assumption that the theory underlying the ABM_{IPS} is correct, the above result could be interpreted in two ways. According to variant (a), it is possible that the normalization for dispersion difference, which we assume is incorporated in human perception, does not actually exist[9]; this would suggest just using linear duration values instead of log-transforms in future applications of the model. Variant (b), however, suggests that the result is an indication that log-scaling the raw duration values before the models are run is not a good way of modeling the normalization procedure. It was, in fact, a very simple model to begin with, because we

[8] Especially because we only used the *mahalanobisDistance* intake strategy.

[9] Or does generally exist, but not or to a lesser degree in populations who are partaking in a sound change in progress. See Section 5.4.2 for further discussion of modeling failing normalization.

did not include the log-transform in the simulation procedure. We only log-transformed all values before the experiment and initialized our agents' memories with the results. As this squashed the asymmetry in input data dispersion even bevore the models were run, this may be the wrong point to apply log transformation. A more complete model might mean to make the agent-speakers produce linear durations and incorporate the log-transformation in the processing procedures of the agent-listeners. A new degree of freedom would then be whether the logarithmic or the linear variant gets saved to memory (i. e. the agent's exemplar cloud).

As per Occam's razor, we should go with the simpler of two alternatives and prefer variant (a). If we can explain the sound change data without any kind of transformation, then why bother implementing one? The fact that the linear-scaled models predict the real observations give reason to believe that they are good models already. However, this evidence has two problems, as outlined below.

Firstly, our data about the Bavarian sound change describe a lenition (i. e. shortening) of fortis stops. The simpler set of our models, that is, those that use closure duration directly instead of applying an additional logarithmic transform, explain the data well. This might suggest the conclusion that these simpler models, in fact, explain the Bavarian sound change well. The problem, however, is that we also know of a purely mathematical explanation about why the ABM_{IPS} might have predicted this specific type of change: the dispersion difference in long vs. short durations introduces a bias towards lenition. What we do not know, however, is whether this mathematical explanation reflects anything in the human brain. Did the ABM_{IPS} predict the data well because it is a good model of the sound change process underlying the data or because it has a mathematical inclination to predict lenition? Or does the sound change process underlying the data also favor lenition and this mathematical inclination is the reason why?

Secondly, with many parameters and procedures to tweak in the model, we run the constant risk of overfitting it to one data set (i. e. making it a perfect explanation for that data set but a poor explanation for others). The amount of existing data sets that have been used for validating the ABM_{IPS} currently ranges in the single digits (Harrington, Gubian, et al., 2019; Harrington et al., 2018; Harrington & Schiel, 2017; Stevens et al., 2019), which is extremely little when compared to an area like meteorology, where data about the weather have been collected world-wide for more than a century and at a resolution of not only daily but far better. Perhaps the only way a linguistic model could get anywhere near the richness and size of these data sets is if the discipline found ethical and efficient ways of using the vast amount of speech data collected by modern speech technology (e. g. personal assistants sold by Google, Apple and Amazon). So as long as there is no further validation that the variant with linear-scaled durations also explains other data sets, we cannot dismiss variant (b), even though variant (a) is the simpler one.

Phonetic theory

The question of what we can conclude for phonetic theory will have to be tackled a little differently. In suggesting further developments of the ABM_{IPS}, we only needed to make sure that these would work and make sense within the framework of the model. For phonetic theory, we must be stricter in the sense that we can only give interpretations of the data that potentially hold true beyond this specific software implementation of the IP model.

Let us consider how the generalizability of a simulation study's results is different from that of a perception experiment's results. It would seem that, in the present study, we have run an experiment that assumed the IP model to be correct (which is not a bad thing, since all experiments, in one way or another, assume that their underlying theory is correct), and this experiment compared a model where humans process logarithmic durations to a model where they process linear durations. The model with linear durations performed better on our test data, so this might suggest that linear durations reflect better what humans process than logarithmic-scaled durations. But we must not jump to this conclusion. Our simulation study, unlike a perception experiment, cannot be interpreted this way. Had we run a traditional perception experiment based on the same theoretical assumptions, there would be one key difference affecting generalizability. The perception experiment would contain the real human perception apparatus in action, and we could be certain that no crucial details of said apparatus had been left out. On the one hand, simplification is the very aim of building a model, but on the other hand, we cannot be certain whether any of the left-out details were crucial ones (without non-simulated empirical evidence).

We could conclude from our study that the linear-scaled variant better reflects the reality of speech perception if and only if we were certain that the specific implementation that forms the ABM_{IPS} (including the configuration of the implementation) was a correct model of reality, which is an extremely specific assumption. This will always be a big question mark in any simulation study in any discipline; but in a discipline that wants to understand a part of something as complex as the human brain, it is a huge question mark (see Section 1.4.3 for a discussion of varying degrees of complexity in simulation). And with a simulation model at this early stage of development (that was only published three years ago in Harrington and Schiel, 2017), it is also a huge question mark. Because a traditional perception experiment does not and cannot leave out any details of human speech perception, the conclusions drawn from it are not tied to such a specific assumption. Moreover, it has a chance of revealing results incompatible with our theory, while in the simulation study, this cannot happen. We can only ever get what we implemented, and that is constrained by the theory.[10] It may make sense to

[10] What we can also get is unplanned side effects of our programming. These can theoretically turn out to be necessary side effects of our theory, in which case it would be interesting to learn about them, but more commonly they will turn out to be software bugs that lead to us modeling something other than what we planned.

be looking for this when we are either very confident in the theory or have some other reason to investigate speech perception as constrained by a certain theory (rather than speech perception as constrained by reality). We must simply remember that the results are only valid as long as the theory is a good model of reality. And not only that: it is the specific software implementation of the theory and the specific configuration of said implementation that must be a good model.

Based on these considerations, we can design future research based on the hypothesis that log-transforming durations is not important in human speech perception. Perhaps this hypothesis will be tested explicitly. Alternatively, this result might simply contribute to shaping expectations in future research about the perception of durations. Note that we should regard the present study primarily as shaping this hypothesis (by providing the direction in it) – we should not regard it as providing evidence that the hypothesis is correct. This is in line with Sections 1.4.4 and 1.5 of this dissertation, where we explained that generating hypotheses is one of the principal aims we are able to attain with simulation studies; and among the most important of such aims.

5.4.2 Language-internal factors

Neither of the two models in the null contact scenario predicted the Bavarian sound change. In this scenario, no language contact at all is assumed, that is, dialect agents only ever interact with other dialect agents, while standard agents only ever interact with other standard agents. Sound change could arise in such a scenario if there were language-internal factors strong enough to trigger a phonetic change in a language community. Already in Chapter 4, we reasoned that fast-speech-induced hypoarticulation was a candidate for such a language-internal factor. According to Ohala's (1993a) concept of hypocorrection, listeners might in some cases fail to compensate for the phonetic shortness of the phonemes they hear in everyday speech (a shortness that is caused by the fastness of everyday speech, not by phonological shortness); thus phonologizing shortness and leading to the diachronic shortening of phonologically long phonemes. This makes fast speech combined with hypocorrection a possible explanation for the sound change we are dealing with (lenition of fortis stops). However, in Chapter 4, we were not able to find empirical evidence for this explanation.

In this simulation study, we tested fast speech again by initializing the agents in the null contact scenario using only the data set of fast speech from Chapter 4. Moreover, the IP model of sound change has, as a key ingredient, another language-internal factor: the asymmetric dispersions of the phonological categories involved in the change. So, this specific configuration of the ABM_{IPS} has input data that make sure one language-internal factor (fast-speech-induced hypoarticulation) could trigger a change; it also has the mechanism to make sure another language-internal factor (asymmetric dispersions) could trigger a change; and it does not model any language contact.

The fact, then, that the model did not predict the sound change at all shows that these two particular sets of assumptions (IP model plus fast-speech-induced hypoarticulation

plus no language contact; plus either logarithmic-scaling or linear-scaling) do not explain the Bavarian sound change. The two sets of assumptions are therefore probably both wrong (using simulation to falsify hypotheses, see Section 1.4.4 of this dissertation). This is in line with Chapter 4 and Kleber (2017a), where language-internal factors without the support of language contact were dismissed as the mechanism that triggered the Bavarian sound change we are investigating. Note, however, that only the set as a whole is considered wrong here. No statement is made about the individual assumptions on their own. Also note that this should, again, be read as a very specific statement that is tied to the specific software implementation of the assumptions.

One option for further development of the ABM_{IPS} might be to explicitly model Ohala's listener errors (i. e. hypocorrection), possibly as a new intake strategy, to be able to simulate a comparison of different trigger mechanisms with as many of the other details as possible (e. g. the speech production mechanism) being kept constant. Ohala (1993a) stresses that this type of listener error "represents a very small fraction of all the interactions between speaker and listener" (p. 246). Such a simulation might reveal how many misperceived interactions, possibly as a fraction of total interactions, it takes for the Ohala model to predict the Bavarian sound change. Simulation would thus again help in shaping hypotheses by filling in some details and thereby contribute to phonetic theory.

In Section 5.4.1, we argued that the logarithmic-based normalization procedure discussed there might only be (completely or partially) absent in populations partaking in a sound change in progress, but present in populations that are stable with respect to the particular phonological property under investigation (stop closure duration). This is basically one interpretation of what it could mean when one population (a phonologically unstable one) has a greater tendency towards hypocorrection than another population (a phonologically stable one). "Modeling Ohala's listener errors," as suggested in the previous paragraph, could therefore amount to implementing a variation of some agents using logarithmic transformations (i. e. corrected perception) and some using linear durations (i. e. hypocorrected perception). Or it could amount to adding a certain (low) probability of assigning a perceived token to the wrong category (which would require a notion of different phonological categories in the simulation). Alternatively, it could amount to a combination of these.

Leaving these development suggestions aside, we appear to need language contact in the equation to successfully predict the Bavarian sound change in our simulations.[11] We will therefore, after a short aside on other language-internal factors, turn to language

[11] Of course, it would also be possible to run the null contact scenario with a mix of normal-paced and fast-paced tokens. This would appear useful if one assumed that the variance in fast tokens alone is not necessary to trigger a change – despite our demonstrations that quantity in Bavarian is an unstable system. However, the stop closure durations of the control group are far below those of the Bavarian agents, even when normal-paced control values are compared to fast-paced agent values. There is currently no reason to believe that a simulation with mixed speech rates would come anywhere near the control condition.

contact in the next section.

Other language-internal factors Fast-speech-induced hypoarticulation is only one example of a language-internal factor potentially able to trigger a change. Other conceivable options include such things as phoneme density in an acoustic space (Bradlow, 1995; Ettlinger, 2007) or microprosody (Kingston, 2011). Apparently, however, none of them was present in the data, because if any had been present (and strong enough within the speech production and perception framework defined by the ABM_{IPS}), the null contact scenarios should have produced a sound change effect. Note, though, that we explicitly initialized these models with fast speech data only, in order to emphasize the effect of fast-speech-induced hypoarticulation. This emphasis potentially has the side effect of deemphasizing or even eliminating other language-internal factors that might have been present before (because focusing something particular necessarily takes focus away from other things). We must leave it at the fact that we only tested explicitly the combination of fast-speech-induced hypoarticulation and asymmetric dispersions (which we reasoned would provide a phonetic bias for the type of sound change we are dealing with, fortis stop lenition), but not other language-internal factors.

5.4.3 Language contact scenarios and timeline

In contrast to the previously discussed null contact scenario, the three linear-scaled models that include language contact to varying degrees do predict the Bavarian sound change (also to varying degrees).

The symmetric contact scenario is the one that best fits our expectations. It predicts that the standard agents remain constant (i. e. the model meets our strict validity criterion), while the dialect agents change their phonology to meet the numerical expectation set by the reference group (i. e. the model meets our quantitative validity criterion) and then develops further, all the way toward the standard agents. In this scenario, each agent had 50% input from standard agents and 50% input from dialect agents (thus symmetric contact).

The maximum contact scenario also meets both the quantitative and the qualitative validity criteria. The main difference to the symmetric contact scenario is that the maximum contact scenario also predicts a slight change in the standard agents (the two also differ in number of interactions required for reaching the validity criteria; this will be discussed below). In this scenario, dialect agents had only standard input, while standard agents had only dialect input (thus maximum contact).

For the asymmetric contact scenario, the 7 repetitions produced different result patterns (which is not the case for the other scenarios). In this section, we will focus on what we earlier called the converging pattern, because that one predicts the Bavarian sound change best. In Section 5.4.4, we will turn to why multiple patterns have been produced. The converging pattern predicts results very similar to the symmetric contact

scenario (which performed very well), again meeting the quantitative and the strict qualitative validity criteria. There is also a difference in number of interactions required to reach validity. In this scenario, the input that standard agents received was about 90% from other standard agents and about 10% from dialect agents. The input that dialect agents received, however, was only about 60% from other dialect agents and about 40% from standard agents (thus asymmetric contact: the dialect agents have more contact, i. e. other-variety input, than the standard agents).

For the ABM_{IPS}, this means that we have identified a number of configurations that correctly predict the data. This provides evidence that these configurations can be considered valid models of the type of sound change we are looking at. The major difference between the present study and previous validation studies of the ABM_{IPS} (Harrington, Gubian, et al., 2019; Harrington et al., 2018; Harrington & Schiel, 2017; Stevens et al., 2019) is that these studies all modeled what we here termed the maximum contact scenario (i. e. they did not model within-group contact). While this scenario also worked for the current data set, explicitly modeling within-group contact led to better results, as shown above. The reason for this might be that in most dialect contact situations, the maximum contact scenario can readily be classified as no more than a first approximation of the real dialect contact,[12] while our other contact scenarios are clearly closer to reality. It is not so clear, however, how much other-variety input real speaker-listeners receive and this surely differs between different dialect contact situations. Moreover, overfitting as an alternative explanation as to why the symmetric and asymmetric contact scenarios worked better than maximum contact cannot be ruled out either.

For phonetic theory, this means that within the assumptions of the ABM_{IPS}, we need language contact to explain the Bavarian sound change. We can only drop the strict tie to the IP framework and other assumptions of the ABM_{IPS} by turning this from a piece of evidence to a hypothesis for future research: if we need language contact to explain the Bavarian sound change in the IP framework, chances are that we also need language contact in other frameworks. Not necessarily, of course, but this is a hypothesis that can be tested in future research.

But what about the amount of language contact? It was clear from the beginning that "maximum contact," where agents have no input from their own variety at all, is an exaggerated model. It has become clear that maximum contact can still generate valid results (reaching both quantitative and qualitative, but not strict qualitative validity). But it has also become clear that the other contact scenarios can generate better results (additionally reaching strict qualitative validity). As to the difference between asymmetric and symmetric contact scenarios: the asymmetric scenario was constructed because we believe that in reality, speakers of Standard German receive less dialect input than dialect speakers receive standard input. We therefore believe that it is a somewhat

[12] In the context of second language learning, a scenario with no own-variety input at all might, in fact, be appropriate.

more realistic scenario, although it is hard to estimate realistic numbers to construct it. The results show that the asymmetric contact scenario, within the framework of the ABM_{IPS}, is somewhat less stable in predicting the change than the symmetric contact scenario, because it produced multiple different patterns across the repeated runs. It is also substantially slower to reach validity (possibly by a factor of 5; but we quantified this only very roughly). It is interesting that when the dialect agents' input changes from 50% to 60%, the slowing down is this drastic. This could help in shaping our ideas about how often sound change is to be expected, because the sound change seems to be pretty sensitive to the amount of contact.

How does the timeline in the simulation model, which is made up of consecutive interactions, relate to real time? To answer this question, we need to address at least one other point: memory size coupled with amount of input. In the ABM_{IPS} (and probably in most or all models), the effect of a newly-heard token is larger when the agent-listener previously had fewer tokens. In other words, the larger an agent-listener's exemplar cloud is, the more input is needed to achieve the same effect. However, we do not know realistic numbers of how many tokens are part of a typical exemplar cloud. Possibly, but very speculatively, we might turn this reasoning upside down and speculate about exemplar cloud size in real speaker-listeners by looking at how many interactions our model needs in order to find a particular effect (and cross-check with real-time data, that is, how long the effect actually took – but this is probably not very reliable because we have nowhere near as many data points as, for instance, meteorologists).

5.4.4 Consistency across repetitions

For six (or seven) out of eight models, the results were consistent across repetitions. This was not the case for the linear-scaled asymmetric contact scenario and possibly for the logarithmic-scaled symmetric contact scenario (which did not stabilize). The results in these configurations were, however, not very noisy, either. Instead, the repetitions appeared to form a number of groups, each reflecting a particular pattern.

We do not need to limit our interpretations to the patterns observed most often. In terms of generative sufficiency, any pattern that was produced by a given configuration can be thought of as explainable by the respectively configured model. Moreover, if we wanted to dig deeper, we could empirically establish the probability of a particular configuration leading to a particular pattern: since the repetitions appeared to produce outcomes that can be assigned to either of a low number of patterns, we are able to fulfill the law of large numbers by running the simulation again, repeating it a number of times that is sufficiently high in comparison to the number of observed patterns (the repetitions are clearly statistically independent). This is not usually possible in non-simulated linguistic experiments, because it would be too expensive to repeat them, say, 100 times.

In cases where we identify a configuration that sometimes leads to a "sound change pattern" and sometimes to a "stability pattern," we might use this procedure to empiri-

cally determine the probability with which a certain model predicts change or stability, respectively. This probability, of course, would have to be taken with a grain of salt: like many of the points discussed here, it only holds true under the specific set of assumptions that make up that model.

Generally speaking, the 7 to 10 repetitions that we conducted appear to be a meaningfully large number of repetitions, since the amount of variation between them is so small. Running a higher number of repetitions, however, could potentially reveal patterns that can be explained by the model, but only with a low probability.

5.5 Acknowledgments and Data

Michele Gubian and Johanna Cronenberg, who are currently the main developers of the ABM_{IPS}, have provided help in analyzing the simulation results and in understanding the model's exact procedures. Specifically to facilitate the present study, Johanna Cronenberg has implemented the configuration options required to design the symmetric and asymmetric contact scenarios.

The raw data this report is based on as well as the code and configuration necessary to replicate the experiments are permanently available online. This includes interaction plots for the repetitions referred to but not included in Section 5.3. They have been published as Jochim and Kleber (2022b).

6 General discussion

This discussion is to wrap up the five previous chapters, which variously dealt with empirical-theoretical aspects and methodological aspects of the phonetics of sound change. One main aim of this dissertation as a whole was to contribute to an explicit discussion of the strengths, weaknesses, and explanatory power of simulation techniques in phonetics (cf. de Boer, 2006). The focus of this discussion will therefore be on comparing what we were able to learn using the apparent time method and agent-based modeling, respectively. At first, however, we will summarize the main findings from the individual chapters.

6.1 Summary

In Chapter 1 (Introduction), we systematically discussed the types of insights we can hope to find when using agent-based modeling. This led to a list including (a) producing actual predictions, (b) thinking a theory through, (c) falsifying hypotheses, and (d) generating hypotheses. Types (b) and (d) were identified as the most promising to pursue when applying agent-based modeling to phonetic research questions.

In Chapter 2, we developed the emuDB Manager, an extension to the EMU Speech Database Management System (Winkelmann et al., 2017). It has proven an invaluable time-saving tool in organizing and editing, across three research institutions in three countries, the empirical data used in the remaining chapters. The Manager has been used with equal success in many other phonetic studies as well, mainly with a focus on speech production (Franzke et al., 2019; Kleber, 2017b; Klingler & Moosmüller, 2017; Wolfswinkler & Harrington, 2020) but also on speech perception (Jochim et al., 2018) or on both (Klingler et al., 2019).

In Chapter 3, we investigated vowel and consonant quantity in Finnish using the apparent time paradigm. Finnish, along with other members of the uralic family, is often cited as an example of a language employing quantity in both vowels and consonants (Bannert, 1976; Crystal, 2008; Dasinger, 1997). We found that phonetic duration has been a stable cue to vowel and consonant quantity in Finnish across generations. Moreover, we found proportional vowel duration (PVD) to be a useful measure to compare vowel plus stop sequences both across languages and across speech styles.

In Chapter 4, we investigated the Western Central Bavarian dialect of German using the apparent time paradigm and non-simulated laboratory-phonological methods. A sound change in progress in this variety had been proposed by Kleber (2017a), and Moos-

müller and Brandstätter (2014). The data set we presented corroborates this proposition; it is also the first data set to show that this change affects vowel-plus-stop sequences where both the vowel and the stop are short (note that the type of stop we are referring to here is often called lenis instead of short, but that is mostly a terminological issue). There is also some evidence suggesting that the change is spreading on a word-by-word basis. Moreover, we investigated fast-speech-induced hypoarticulation as a trigger mechanism of the change but found no compelling evidence for it. This supports Kleber's (2017a) conclusion that the sound change is best explained as dialect convergence towards a standard language.

In Chapter 5, we investigated the same data set about Western Central Bavarian using agent-based simulation; specifically using the simulation model we termed ABM_{IPS} (which was first presented in Harrington and Schiel, 2017). We found some configurations of the ABM_{IPS} that are well able to explain the data, and we identified a number of useful ways of proceeding with agent-based modeling in general and the ABM_{IPS} in particular.

6.2 Comparison

We will now turn to comparing Chapters 3, 4 and 5. With Chapter 2 describing the development of a tool for phonetic research rather than an instance of phonetic research, it would not be useful to include this chapter in the comparison.

For the purposes of this comparison, a couple of words remain to be said about the typological work presented in Chapter 3 and two aspects of it that were not entirely adopted in Chapter 4. One aim of that study was to investigate the factor *speech rate* and its effect on vowel and consonant duration. In the experimental design, however, we chose to work with a proxy measure, *loudness*, instead of speech rate proper. The reasoning was that at some point we might have been able to compare these data to children's speech data. With child participants, it is more feasible to elicit variation in loudness than variation in speech rate, so we wanted to do the same with adult participants. In hindsight, however, it would have been more fruitful to work with speech rate directly instead of a proxy measure. This would have facilitated a cross-linguistic comparison of the effect of speech rate in adults. A second apparent incoherence is that, in Chapter 3, we used the measure *proportional vowel duration*, which combines a vowel and an adjacent consonant into one numerical measure. The reasoning here was to reduce numerical complexity for cross-linguistic comparisons of phonological categories. Chapter 4, however, was a deep dive into one dialect of German and its convergence towards the standard. A reduction of this kind was not desirable here and so we used individual measures for the vowel and the consonant in this study.

In Chapter 3, we set out to find a firm point of reference for studies of changing quantity in other languages. While this dissertation does not have a strong focus on typology, this point of reference has remained important in two respects. Firstly, it employs a very

similar apparent time method to the study in Chapter 4. The fact, then, that very similar methods showed generational change for Bavarian in Chapter 4, but generational stability for Finnish in Chapter 3, corroborates that the generational change in Bavarian is not merely an artifact of the method. Secondly, the order of magnitude of the fortis–lenis contrast in terms of phonetic duration puts our data into better perspective. We have found a roughly 2:1 duration ratio for fortis vs. lenis consonants in Finnish, and our Bavarian data get close to such a ratio and also close to the absolute closure durations of Finnish, although this does not appear in an explicit analysis in the present dissertation. Standard German remains farther away from both the ratio and absolute durations of Finnish. This means that if the Bavarian cross-generational trend of closure shortening reported on in Chapter 4 were to continue in the future, this might actually constitute a change from quantity language to non-quantity language.

We will now proceed to comparing in more detail the two chapters that dealt with the Western Central Bavarian sound change using an apparent-time approach and agent-based modeling, respectively. We will start by revisiting the speech production experiment from Chapter 4.

We have conducted an acoustic speech production experiment with 30 speakers and 25 target words. These words were split into four groups based on the phonemes they contain in Standard German. Two of these groups contained a long vowel plus fortis stop sequence (V:C:) and a short vowel plus lenis stop sequence (VC), respectively; such words, according to the rule of complementary length (Bannert, 1976; Hinderling, 1980; Pfalz, 1913; Seiler, 2005), should not exist in the dialect; we called these two groups phonotactically illegal in the dialect. We called the other two groups (V:C and VC:) phonotactically legal in the dialect. We then went on to construct a list of words that predictably constitute minimal pairs in Standard German and are known to and used by dialect speakers as well. The words in the illegal group are probably best described as native words in Standard German and as loan words or borrowings in the dialect; however, there is no authoritative source on what constitutes a loan word in the dialect[1]. Moreover, according to Hinderling (1980), all words native to the dialect and even all borrowings from the standard should be pronounced in either of the two phonotactically legal ways. Some of the words used may also be lexically more frequent in Standard German than in the dialect; but again, there is no authoritative source on lexical frequency in the dialect.

Our 30 speakers were evenly divided into the three groups Southern Standard German speakers, younger dialect speakers, and older dialect speakers. Via the use of clearly dialectal carrier sentences for the two respective groups, we made sure that they were, in fact, speaking their dialect during the experiment (and neither Southern Standard German nor regionally-accented Standard German).

One finding was that the older dialect speakers did indeed pronounce the words in

[1]For Standard German, there exist not only dictionaries, but even specific etymological dictionaries (e. g. Pfeifer, 2005); for Bavarian, only a number of general-purpose dictionaries do exist.

the VC group with short vowel, but with fortis-like stop, as predicted by the dialect's reported grammar.[2] The younger dialect speakers, however, pronounced these words with short vowels and lenis-like stops. Considering that the speaker groups were only ten in size, but the groups were pretty consistent internally (especially the younger group, see the fortis–lenis overlap in Section 4.3.4), we regard this as pretty strong evidence that the rule of complementary length used to be but is no longer as strong as described by Hinderling (1980).

As for the V:C: word group, we found that both younger and older dialect speakers pronounced them with long vowels and fortis-like stops; that is, neither of the dialect groups acted as predicted by the rule of complementary length. We consider the most likely explanation to be that even our older generation is too young to be completely unaffected by the change. While we cannot completely rule out the possibility that V:C: (unlike VC) was never illegal in the first place (cf. also Seidelmann, 2013), we deem this explanation highly unlikely; firstly due to the overwhelming body of literature describing the "illegalness" and secondly due to very frequent words like *Vater* 'father' that have V:C: in Standard German, but either V:C or VC: in the dialect.

We also tested, but did not find support for, fast speech being a trigger of this change. We therefore believe that dialect convergence towards a standard is the best explanation for the observed change. The younger dialect speakers' behavior could also be explained in terms of an increase in code mixing compared to the older group. This might, however, be indistinguishable from contact-induced change even at the theoretical level.

An interesting next step would be to observe whether the change now also spreads to words like *Vater* and *Kater* that are established in the Bavarian lexicon without the V:C: sequence.

We will now turn to what the ABM approach was able to add to that. Most of the conclusions related to the ABM_{IPS} were based on the assumption of closure duration change suggested by the previous chapter: a population resembling the older dialect group should develop stop closure durations in VC words towards values typical of the younger dialect speakers, when this population is in contact with a population resembling the standard speakers.

In comparing linear-scaled models, where agents were initialized with raw closure durations, to logarithmic-scaled models, where agents were initialized with the natural logarithm of these raw closure durations, we found a result contrary to our expectations. We expected the logarithmic-scaled models to predict the data better, but they performed worse. We offer three ideas to take away from this.

First, it would be interesting to conduct a non-simulated perception experiment with dialect listeners to see whether their categorization of stop durations is indeed better predicted by linear than logarithmic values; that would also be interesting for the further development of the ABM_{IPS}.

[2] Although these stops were not quite as long as the fortis stops in the word group that we labeled VC: a priori.

Second, it would be interesting to implement a version of the ABM_{IPS} that is based on logarithmic values, but not as simplistically as in the present study. In the present study, logarithms were calculated before initializing the agents and then used throughout the simulations. One improvement might be to calculate the logarithm in the agent-listeners during the individual interactions, before the decision of accepting or not accepting a token into memory. This way forward would be especially advisable if a perception experiment as described above were conducted and showed that logarithmic durations are a better predictor of stimulus categorization than linear durations.

And third, this suggests one possible failure condition for models of sound change that, like Ohala's (1993a), assume that rare instances of listener errors lead to sound change. These rare events might not be a complete failure of recognizing the speaker's intended production but rather a change from the typical logarithmic to an unusual linear perception. Another way simulation might add to Ohala's model would be to find out how many listener errors, as a fraction of total interaction, would be required to predict a change.

Another factor we explored was the amount of language contact between the agent populations. The results broadly suggested that modeling mixed-variety input is a better predictor of the observed change than only other-variety input (obviously, mixed-variety input is also a more realistic model of the contact situation). The results further suggested that the speed of sound change may be severely affected by small changes in the amount of other-variety input. A change from 50% standard input for the dialect agents to 40% drastically slowed down the population sound change. Based on this, it might be interesting to examine observed language contact situations more closely and analyze whether the speed of change is in fact coupled with amount of input, and with the symmetry of input between language community, in such a way.

Our analyses have also highlighted the issue of data scarcity. Generally, the field of phonetics should work towards larger speaker groups that can more reasonably be deemed representative of an entire language community; possibly by finding ethical and efficient ways of utilizing the resources generated by current personal speech assistants such as those sold by Google, Apple, and Amazon. In the present case, however, we have seen that simulation studies would benefit from a larger number of non-simulated experiments on the same phenomenon, in order to increase the options of validation. More experiments on the same phenomenon, at the expense of the number of phonetic phenomena investigated, would generally benefit the reliability of results. It is a challenge for fundamental researchers to strike a balance here; exploring many phenomena on the one hand and reliably testing theory drawn from available observations on the other hand.

Using the ABM_{IPS} together with our laboratory data has indeed yielded a number of unexpected results. We used these results to suggest hypotheses to be tested and ideas to be explored in future studies. The development of a simulation model is pretty costly, but it has proven a valuable tool in shaping theory. This theory needs to be validated against empirical data, as this remains something that a simulation study

cannot do. In Chapter 5, we have given a case study of how simulation can aid in theorizing, based on a pre-existing model that saw only minor modifications to facilitate the conducted comparisons. Researchers can use this case study as well as those provided by the model's developers (Harrington, Gubian, et al., 2019; Harrington et al., 2018; Harrington & Schiel, 2017; Stevens et al., 2019) to decide whether it is worth the cost with their specific research endeavors. It depends on how well-developed their theory is and whether there are details or parameters to it that can only be speculated about. In such cases, simulation can suggest parameters that are more likely than others.

Zusammenfassung auf Deutsch

Die vorliegende Arbeit befasst sich mit verschiedenen Methoden, Lautwandel zu untersuchen – insbesondere anhand des Westmittelbairischen, einem Dialekt des Deutschen. Westmittelbairisch wird im südöstlichen Landesteil der Bundesrepublik Deutschland (Oberbayern, Niederbayern) sowie in weiten Teilen Österreichs gesprochen. In diesem Dialekt wird seit einiger Zeit ein Lautwandel vermutet (Kleber, 2017a, 2018; Moosmüller & Brandstätter, 2014; Schikowski, 2009). Dieser beeinflusst ein phonotaktisches Gesetz, nach dem wort-mediale Lenis-Plosive ausschließlich nach Langvokal und wort-mediale Fortis-Plosive ausschließlich nach Kurzvokal vorkommen; diese Regel ist in der Literatur unter verschiedenen Namen bekannt (komplementäre Länge, mittelbairische Quantitätsverhältnisse oder auch Pfalzsches Gesetz; siehe z. B. Bannert, 1976; Hinderling, 1980; Pfalz, 1913; Seiler, 2005). Die Hinweise auf die Veränderung des phonologischen Systems stammen aus akustischer Evidenz über den verwandten Wiener Dialekt (Moosmüller & Brandstätter, 2014), aus akustischer Evidenz sowie Perzeptionsexperimenten über die bairisch gefärbte Standardsprache (Kleber, 2017a, 2018) sowie aus dialektologisch-ohrenphonetisch geprägter Arbeit über den westmittelbairischen Dialekt selbst (Schikowski, 2009).

Angesiedelt in einem trinationalen (Deutschland, Österreich, Schweiz; D-A-CH) Forschungsprojekt wurde in einem Sprachproduktionsexperiment ein cross-linguistisches Korpus erhoben, das Sprecher*innen des südlichen Standarddeutsch, wesmittelbairische Dialektsprecher*innen sowie Sprecher*innen verschiedener österreichischer und Schweizer Varietäten umfasst. In dem Sprachproduktionsexperiment wurden mit laborphonologischen Methoden Einzeläußerungen elizitiert, wobei in zwei verschiedenen experimentellen Bedingungen die Sprechgeschwindigkeit variiert wurde. Die sprechertypische normale Sprechgeschwindigkeit auf der einen Seite, eine sprechertypische schnelle Sprechgeschwindigkeit auf der anderen Seite. Für die schnelle Bedingung wurden die Proband*innen angewiesen, so schnell zu sprechen, wie es ihnen möglich sei ohne dabei in Hast zu verfallen. Auf diese Weise konnten wir eine schnelle Sprechgeschwindigkeit modellieren, die nicht nach Labormanier *so schnell wie möglich* war, sondern einer natürlich vorkommenden erhöhten Sprechgeschwindigkeit entsprach.

Der Zweck dieser Modellierung war es, die auf Kohler (1984), Lindblom (1990) und Ohala (1993a) basierende Hypothese zu testen, dass Hypoartikulation, verursacht durch schnelle Sprechgeschwindigkeit, ein Auslöser für diachrone Lautkürzung sein könnte.

Die Analysen in der vorliegenden Dissertation umfassen drei Sprechergruppen: eine standarddeutsch sprechende Kontrollgruppe, eine jüngere Dialektgruppe (Geburtsjahre 1995-1997) und eine ältere Dialektgruppe (Geburtsjahre 1950-1971). Unsere akustis-

chen Analysen haben insbesondere gezeigt, dass Wörter wie *Pudding* oder *Rabbi*, also solche mit wort-medialem Kurzvokal gefolgt von Lenis-Plosiv, von der älteren Dialektgruppe entsprechend der komplementären Länge produziert worden ist: Der Plosiv wurde mit fortis-ähnlichen Verschlussdauern produziert, der Vokal in Kurzvokallänge. Die jüngeren Dialektsprecher sind diesem Muster nicht gefolgt; ihre Plosivverschlussdauern waren genau wie die Vokale kurz. Die standarddeutsche Kontrollgruppe hat solche Wörter erwartungsgemäß mit Kurzvokal und kurzem Plosiv produziert. Wörter wie *Lupe*, also solche mit wort-medialem Langvokal gefolgt von Fortis-Plosiv, die gemäß der komplementären Länge im Westmittelbairischen ebenfalls für phonotaktisch illegal gelten, wurden von *beiden* Dialektgruppen, also sowohl von den jüngeren wie auch von den älteren Dialektsprechern, mit Langvokal gefolgt von langem Plosiv produziert; solche Wörter scheinen also selbst für die ältere von uns getestete Generation ins phonologische System zu passen. Die Ergebnisse liefern damit weitere Evidenz dafür, dass die komplementäre Länge im Dialekt als phonotaktische Gesetzmäßigkeit schwächer wird.

Desweiteren haben die Analysen gezeigt, dass Voice Onset Time (VOT) von den jüngeren Dialektsprechern als akustischer Cue für Fortis-Plosive verwendet wird. Dies war gemäß der dialektologischen Literatur (z. B. Bannert, 1976; Wiesinger, 1990) nicht zu erwarten. Der Cue ist in dieser Gruppe weniger stark ausgeprägt als in der standarddeutschen Kontrollgruppe, für die VOT als Fortis-Cue zu erwarten war (Jessen, 1998; Wiese, 1996); jedoch merklich stärker ausgeprägt als in der älteren Dialektgruppe.

Außerdem konnten wir mittels eines statistischen Maßes, der *Fortis-Lenis-Überlappung*, zeigen, dass nicht alle Wörter in einer Kategorie gleichermaßen von dem Wandel betroffen sind oder aber im Wandelprozess unterschiedlich weit fortgeschritten sind. Die Wörter *bieten*, *Bieter* und *Pute* zeigten im Gegensatz zu den fünf anderen Wörter in der Gruppe Langvokal-plus-Fortisplosiv deutlich erhöhte inter- und intraindividuelle Variation.

Die Modellierung der Sprechgeschwindigkeit führte in verschiedenen Analysen nicht zu einer Erhärtung der Hypothese, wonach durch erhöhte Sprechgeschwindigkeit verursachte Hypoartikulation Auslöser für den beobachteten Wandel sei. Weder konnten wir in der älteren Dialektgruppe beobachten, dass die von der Konsonantenkürzung betroffene Wortgruppe stärker von der sprechgeschwindikgeitsbedingten Kürzung betroffen war als die nicht betroffenen Wortgruppen. Noch trat eine stärkere statistische Dispersion in der betroffenen Wortgruppe auf; eine solche hätte als Zeichen einer instabilen Kategorie gewertet werden können, die wiederum in der älteren Gruppe ein Anzeichen für einen den Wandel begünstigenden phonetischen Bias (Garrett & Johnson, 2013) gewesen wäre, in der jüngeren Gruppen ein Anzeichen für eine Kategorie, die sich noch nicht stabilisiert hat.

Mit der Gegenüberstellung zweier Generationen im Sprachproduktionsexperiment sind wir dem Apparent-Time-Ansatz gefolgt (Bailey et al., 1991; Labov, 1963). In einer anderen Studie im Rahmen dieser Dissertation haben wir den Apparent-Time-Ansatz auch auf Finnisch angewandt. Finnisch gilt als eine Sprache, die starken Gebrauch von phonologischer Quantität sowohl in Vokalen als auch in Kosonanten macht (Bannert, 1976; Crystal, 2008; Dasinger, 1997). Als solche wird sie sowohl in der Literatur als auch von

uns als Referenzsprache genannt. Wir konnten mithilfe des Apparent-Time-Ansatzes zeigen, dass Quantität im Finnischen zwischen der älteren und der jüngeren Generation unverändert geblieben ist. Außerdem konnten wir mithilfe der in dieser Studie erhobenen Daten verifizieren, dass konsonantenquantität-bedingte Verschlussdauerunterschiede im Westmittelbairischen in der selben Größenordnung liegen wie in derjenigen Sprache, die häufig als Referenzsprache für Konsonantenquantität genannt wird. Außerdem haben wir in unserer Finnisch-Studie es unternommen, die wortmediale Vokal-Konsonanten-Sequenz in einem gemeinsamen, eindimensionalen Maß zu beschreiben: der proportionalen Vokaldauer (PVD), die in der Literatur auch häufig VC-Ratio genannt worden ist (z. B. Kleber, 2017a; Kohler, 1979). Mit diesem Maß ist es besser als mit Einzelmaßen gelungen, phonologische Kategorien selbst über Sprachstiländerungen hinweg zu beschreiben; ähnliche Ergebnisse haben Pickett et al. (1999) für das Italienische berichtet.

Im Einfühungskapitel und in Kapitel 5 dieser Dissertation haben wir uns dann mit der agentenbasierten Modellierung (Cioffi-Revilla, 2017b; Manzo, 2014) unserer laborphonologischen Daten befasst. Das Einführungskapitel befasst sich aus erkenntnistheoretischer Perspektive mit den Vor- und Nachteilen der agentenbasierten Modellierung, einer Simulationsmethode. Daraus hervor geht eine Aufstellung möglicher Erkenntnistypen, die die Methode rein theoretisch leisten kann, und zwar (a) tatsächliche Vorhersagen über Lautwandel treffen (ähnlich eines Wetterberichts), (b) eine komplexe Theorie über dynamische Systeme „durchzudenken“, deren Konsequenzen ohne Simulationsmethoden nicht zu überblicken sind, (c) das Falsifizieren von Hypothesen und (d) das Generieren von Hypothesen. Die Typen (b) und (d) wurden dabei als diejenigen identifiziert, die für phonetische Forschungsfragen die vielversprechendsten zu sein scheinen. In Kapitel 5 haben wir dann ein bestimmtes Modell, das ABM_{IPS}, verwendet, um mit unseren laborphonologischen westmittelbairischen Daten als Ausgangslage einen Lautwandel zu simulieren. Das ABM_{IPS} wird erst seit einigen Jahren entwickelt, vorgestellt und getestet (Harrington, Gubian, et al., 2019; Harrington et al., 2018; Harrington & Schiel, 2017; Stevens et al., 2019). Wir haben das Modell dahingehend erweitert, dass in einer simulierten Sprachkontaktsituation (zwischen Standarddeutsch und Westmittelbairisch) die Agenten nicht ausschließlich mit Agenten aus der anderen Sprachgruppe interagieren, sondern auch mit Agenten, die „die eigene Varietät sprechen“, also mit Daten eines Sprecher-Hörers aus der eigenen Sprechergruppe initialisiert wurden. Dabei konnten wir zeigen, dass sich insbesondere diese gemischten Sprachkontaktsituationen in der Simulation eignen, die beobachteten Daten vorherzusagen. Außerdem wurden in diesem Kapitel das ABM_{IPS}, intialisiert mit linear-skalierten Verschlussdauerdaten, mit dem ABM_{IPS}, initialisiert mit log-skalierten Verschlussdauerdaten, verglichen. Die Erwartung hierbei war, dass log-skalierte Daten näher am menschlichen Perzeptionsmechanismus sind (Rosen, 2005; Varshney & Sun, 2013) und sich daher besser für die Simulation eignen. Diese Erwartung hat sich nicht bestätigt; möglicherweise war unsere Implementierung einer logarithmus-basierten Verarbeitung akustischer Lautdauer allzu simplistisch.

Ein methodisches Kapitel dieser Dissertation hat sich mit der Entwicklung des *emuDB-Managers* befasst, einer Erweiterung des *EMU Speech Database Management System* (Winkelmann et al., 2017). Diese methodische Weiterentwicklung war notwendig, um im trinationalen Forschungsverbund die Annotation des aufzubauenden phonetischen Sprachkorpus zu organisieren. Es handelt sich dabei um ein Tool, das Cloud-Speicher für die phonetische Datenannotation nutzbar macht. Sämtliche für diese Dissertation empirisch erhobenen Daten wurden mit seiner Hilfe annotiert; darüber hinaus ist das Tool auch in anderen Forschungsprojekten eingesetzt worden, sowohl mit Fokus auf Sprachproduktion (Franzke et al., 2019; Kleber, 2017b; Klingler & Moosmüller, 2017; Wolfswinkler & Harrington, 2020) als auch auf Sprachperzeption (Jochim et al., 2018) oder sogar Beidem (Klingler et al., 2019).

References

Allegrini, J., Dorer, V., & Carmeliet, J. (2013). Wind tunnel measurements of buoyant flows in street canyons. *Building and Environment*, *59*, 315–326. https://doi.org/10.1016/j.buildenv.2012.08.029

Bailey, G., Wikle, T., Tillery, J., & Sand, L. (1991). The apparent time construct. *Language Variation and Change*, *3*(3), 241–264. https://doi.org/10.1017/S0954394500000569

Baker, A., Archangeli, D., & Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change*, *23*(3), 347–374. https://doi.org/10.1017/S0954394511000135

Bannert, R. (1976). *Mittelbairische Phonologie auf akustischer und perzeptorischer Grundlage* (B. Malmberg & K. Hadding, Eds.; Vol. 10) [Zweite ISBN: 978-3-7705-1452-6]. CWK Gleerup.

Bates, D., Maechler, M., Bolker, B., Walker, S., Bojesen Christensen, R. H., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., & Green, P. (2018). Lme4: Linear Mixed-Effects Models using 'Eigen' and S4 [R Package].

Beddor, P. S. (2006). Weighted multiple cues and phonological features. *The Journal of the Acoustical Society of America*, *119*(5), 3269–3269. https://doi.org/10.1121/1.4786134

Beddor, P. S. (2009). A Coarticulatory Path to Sound Change. *Language*, *85*(4), 785–821. https://doi.org/10.1353/lan.0.0165

Bradlow, A. R. (1995). A comparative acoustic study of English and Spanish vowels. *The Journal of the Acoustical Society of America*, *97*(3), 1916–1924. https://doi.org/10.1121/1.412064

Braun, A. (1988). *Zum merkmal "Fortis/Lenis". Phonologische betrachtungen und instrumental - phonetische untersuchungen an einem mittelhessischen dialekt*. Stuttgart, Steiner.

Braunschweiler, N. (1997). Integrated Cues of Voicing and Vowel Length in German: A Production Study. *Language and Speech*, *40*(4), 353–376. https://doi.org/10.1177/002383099704000403

Britain, D. (2010). Supralocal Regional Dialect Levelling (C. Llamas & D. Watt, Eds.). In C. Llamas & D. Watt (Eds.), *Language and Identities*. Edinburgh, Edinburgh University Press.

Bukmaier, V., & Harrington, J. (2016). The articulatory and acoustic characteristics of Polish sibilants and their consequences for diachronic change. *Journal of the*

International Phonetic Association, *46*(3), 311–329. https://doi.org/10.1017/S0025100316000062

Bukmaier, V., Harrington, J., & Kleber, F. (2014). An analysis of post-vocalic /s-ʃ/ neutralization in Augsburg German: Evidence for a gradient sound change. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00828

Cedergren, H. (1973). *The Interplay of Social and Linguistic Factors in Panama* (Doctoral dissertation). Cornell University.

Chen, M. (1970). Vowel Length Variation as a Function of the Voicing of the Consonant Environment. *Phonetica*, *22*(3), 129–159. https://doi.org/10.1159/000259312

Cioffi-Revilla, C. (2017a). *Introduction to Computational Social Science: Principles and Applications* (2nd ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-50131-4

Cioffi-Revilla, C. (2017b). Simulations I: Methodology (C. Cioffi-Revilla, Ed.). In C. Cioffi-Revilla (Ed.), *Introduction to Computational Social Science: Principles and Applications*. Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-50131-4_8

Crystal, D. (1987). *The Cambridge encyclopedia of language*. Cambridge, New York, Melbourne, Cambridge University Press.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics* (Sixth Edition). Malden, MA u. a., Blackwell.

Dasinger, L. (1997). Issues in the acquisition of Estonian, Finnish, and Hungarian: A crosslinguistic comparison (D. Slobin, Ed.). In D. Slobin (Ed.), *The crosslinguistic study of language acquisition*. Lawrence Erlbaum.

de Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics*, *28*(4), 441–465. https://doi.org/10.1006/jpho.2000.0125

"In order to make the interactions between the agents more interesting, noise can be added to the formant frequencies" (p.446)

de Boer, B. (2006). Computer modelling as a tool for understanding language evolution (N. Gontier, J. P. Van Bendegem, & D. Aerts, Eds.). In N. Gontier, J. P. Van Bendegem, & D. Aerts (Eds.), *Evolutionary Epistemology, Language and Culture: A Non-Adaptationist, Systems Theoretical Approach*. Dordrecht, Springer Netherlands. https://doi.org/10.1007/1-4020-3395-8_17

"As in many models randomness plays an important role, this needs to be modeled using the computer's pseudo random number generator." (p. 14/Section 3.4)

Devenport, W. J., Burdisso, R. A., Borgoltz, A., Ravetta, P. A., Barone, M. F., Brown, K. A., & Morton, M. A. (2013). The Kevlar-walled anechoic wind tunnel. *Journal of Sound and Vibration*, *332*(17), 3971–3991. https://doi.org/10.1016/j.jsv.2013.02.043

Doty, C. S., Idemaru, K., & Guion, S. G. (2007). Singleton and geminate stops in Finnish - acoustic correlates., In *INTERSPEECH*, ISCA.

Draxler, C., & Jänsch, K. (2004). SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software, In *Proc. of the IV. International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Draxler, C. (2008). *Korpusbasierte Sprachverarbeitung* [OCLC: 904822949]. Tübingen, Gunter Narr Verlag.

Dromey, C., & Ramig, L. O. (1998). Intentional Changes in Sound Pressure Level and RateTheir Impact on Measures of Respiration, Phonation, and Articulation. *Journal of Speech, Language, and Hearing Research*, *41*(5), 1003–1018. https://doi.org/10.1044/jslhr.4105.1003

Eames, I., & Flor, J. B. (2011). New developments in understanding interfacial processes in turbulent flows. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1937), 702–705. https://doi.org/10.1098/rsta.2010.0332

Engstrand, O., & Krull, D. (1994). Durational Correlates of Quantity in Swedish, Finnish and Estonian: Cross-Language Evidence for a Theory of Adaptive Dispersion. *Phonetica*, *51*(1-3), 80–91. https://doi.org/10.1159/000261960
Volltext als Print in der Phonetik-Bib

Epstein, J. M. (2006). *Generative Social Science: Studies in Agent-Based Computational Modeling* (STU - Student edition). Princeton University Press.

Ettlinger, M. (2007). An exemplar-based model of chain shifts (J. Trouvain & W. J. Barry, Eds.). In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2016). *Statistik. Der Weg zur Datenanalyse.* (8., überarbeitete und ergänzte Auflage). Berlin, Heidelberg, Springer.

Farnetani, E., & Recasens, D. (2010). Coarticulation and Connected Speech Processes (W. J. Hardcastle, J. Laver, & F. E. Gibbon, Eds.; 2nd ed). In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (2nd ed). Chichester, West Sussex, U.K. Malden, Ma, Wiley-Blackwell.

Fowler, J. (1986). *The social stratification of (r) in New York City department stores, 24 years after Labov* [New York University unpublished manuscript]. New York University unpublished manuscript.

Franzke, R., Wolfswinkler, K., & Harrington, J. (2019). Gender-based studies on the spoken language of children [Abstract only], In *15. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, Düsseldorf. Abstract only.

Garrett, A., & Johnson, K. (2013). Phonetic bias in sound change (A. C. L. Yu, Ed.). In A. C. L. Yu (Ed.), *Origins of Sound Change: Approaches to Phonologization.* Oxford, Oxford University Press.

Gay, T. (1981). Mechanisms in the Control of Speech Rate. *Phonetica*, *38*(1-3), 148–158. https://doi.org/10.1159/000260020

Gilbert, N., & Abbott, A. (2005). Introduction. *American Journal of Sociology*, *110*(4), 859–863. https://doi.org/10.1086/430413

Hardcastle, W. J., & Hewlett, N. (Eds.). (1999). *Coarticulation: Theory, Data and Techniques*. Cambridge, Cambridge University Press. https://doi.org/10.1017/CBO9780511486395

Hardcastle, W. J., Laver, J., & Gibbon, F. E. (Eds.). (2010). *The handbook of phonetic sciences* (2nd ed). Chichester, West Sussex, U.K. Malden, Ma, Wiley-Blackwell.

Harrington, J. (2007). Evidence for a relationship between synchronic variability and diachronic change in the Queen's annual Christmas broadcasts (J. Cole & J. Hualde, Eds.). In J. Cole & J. Hualde (Eds.), *Laboratory Phonology 9*. Berlin, Mouton.

Harrington, J. (2010). *Phonetic analysis of speech corpora* [OCLC: 123375074]. Chichester, U.K.; Malden, MA, Wiley-Blackwell.

Harrington, J., Gubian, M., Stevens, M., & Schiel, F. (2019). Phonetic change in an Antarctic winter. *The Journal of the Acoustical Society of America*, *146*(5), 3327–3332. https://doi.org/10.1121/1.5130709

Harrington, J., Kleber, F., & Reubold, U. (2012). The production and perception of coarticulation in two types of sound changes in progress (S. Fuchs, M. Weirich, D. Pape, & P. Perrier, Eds.). In S. Fuchs, M. Weirich, D. Pape, & P. Perrier (Eds.), *Speech Planning and Dynamics*. Frankfurt, Peter Lang.

Harrington, J., Kleber, F., Reubold, U., Schiel, F., & Stevens, M. (2018). Linking Cognitive and Social Aspects of Sound Change Using Agent-Based Modeling. *Topics in Cognitive Science*, *10*(4), 707–728. https://doi.org/10.1111/tops.12329

Harrington, J., Kleber, F., Reubold, U., Schiel, F., & Stevens, M. (2019). The phonetic basis of the origin and spread of sound change (W. F. Katz & P. F. Assmann, Eds.). In W. F. Katz & P. F. Assmann (Eds.), *The Routledge Handbook of Phonetics*. Routledge.

Harrington, J., & Schiel, F. (2017). /u/-fronting and agent-based modeling: The relationship between the origin and spread of sound change. *Language*, *93*(2), 414–445. https://doi.org/10.1353/lan.2017.0019

Hermann, E. (1929). Lautveränderungen in der Individualsprache einer Mundart. *Nachrichten der Gesellschaft der Wissenschaften zu Göttingen. Philologisch-historische Klasse.*, *11*, 195–214.

Hewlett, N., Matthews, B., & Scobbie, J. M. (1999). Vowel Duration In Scottish English Speaking Children, In *Proceedings of the 14th International Conference on Phonetic Sciences*, San Francisco.

Hinderling, R. (1980). Lenis und Fortis im Bairischen. Versuch einer morphophonemischen Interpretation. *Zeitschrift für Dialektologie und Linguistik*, *47*(1), 25–51.

Hinskens, F. (1998). Dialect Levelling: A Two-dimensional Process. *Folia Linguistica*, *32*(1-2), 35–52. https://doi.org/10.1515/flin.1998.32.1-2.35

Hoole, P., & Mooshammer, C. (2002). Articulatory analysis of the German vowel system (P. Auer, P. Gilles, & H. Spiekermann, Eds.). In P. Auer, P. Gilles, & H. Spiekermann (Eds.), *Silbenschnitt und Tonakzente*. Tübingen, Niemeyer.

Howell, R., Qin, N., Edwards, J., & Durrani, N. (2010). Wind tunnel and numerical study of a small vertical axis wind turbine. *Renewable Energy*, *35*(2), 412–422. https://doi.org/10.1016/j.renene.2009.07.025

Jessen, M. (1998). *Phonetics and Phonology of Tense and Lax Obstruents in German*. John Benjamins Publishing.

Jochim, M. (2017). Extending the EMU Speech Database Management System: Cloud Hosting, Team Collaboration, Automatic Revision Control, In *Proceedings of Interspeech 2017, Stockholm, Sweden*, Stockholm, Sweden. https://doi.org/10.21437/Interspeech.2017

Jochim, M., & Kleber, F. (2017a). What do Finnish and Central Bavarian have in common? Towards an acoustically based quantity typology. *Open Data LMU*. https://doi.org/10.5282/ubm/data.106

Jochim, M., & Kleber, F. (2017b). What do Finnish and Central Bavarian have in common? Towards an acoustically based quantity typology, In *Proceedings of Interspeech 2017, Stockholm, Sweden*, Stockholm, Sweden. https://doi.org/10.21437/Interspeech.2017-1285

Jochim, M., & Kleber, F. (2022a). Dataset and analysis for Chapter 4 of Internal and external factors in a Bavarian sound change: Agent-based simulations and measurements in apparent time. *Open Data LMU*. https://doi.org/10.5282/ubm/data.303

Jochim, M., & Kleber, F. (2022b). Dataset and analysis for Chapter 5 of Internal and external factors in a Bavarian sound change: Agent-based simulations and measurements in apparent time. *Open Data LMU*. https://doi.org/10.5282/ubm/data.304

Jochim, M., Kleber, F., Klingler, N., Pucher, M., Schmid, S., & Zihlmann, U. (2018). Measuring the Role of Hypoarticulation in a Sound Change in Progress in Southern German. [Abstract only], In *14. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, Wien. Abstract only.

Johnson, K. (1997). Speech Perception without Speaker Normalization: An Exemplar Model (K. Johnson & J. W. Mullenix, Eds.). In K. Johnson & J. W. Mullenix (Eds.), *Talker Variability in Speech Processing*. San Diego, Academic Press.

Keller, R. (1990). *Sprachwandel. Von der unsichtbaren Hand in der Sprache*. Tübingen, Francke.

Kerswill, P. (2002). Models of linguistic change and diffusion: New evidence from dialect levelling in British English. *Reading Working Papers in Linguistics*, *6*, 187–216.

Kerswill, P. (2003). Dialect levelling and geographical diffusion in British English (D. Britain & J. Cheshire, Eds.). In D. Britain & J. Cheshire (Eds.), *Social dialectology. In honour of Peter Trudgill,* Amsterdam, Benjamins.

Kingston, J. (2011). Tonogenesis, In *The Blackwell Companion to Phonology*. American Cancer Society. https://doi.org/10.1002/9781444335262.wbctp0097

Kirby, J. P. (2014). Incipient tonogenesis in Phnom Penh Khmer: Computational studies. *Laboratory Phonology*, *5*(1), 195–230. https://doi.org/10.1515/lp-2014-0008

Kisler, T., & Kleber, F. (2019). Zur Validität automatisch segmentierter Daten: Eine akustische Analyse der mittelbairischen Lenisierung im Deutsch Heute-Korpus.

(S. Kürschner, M. Habermann, & P. O. Müller, Eds.). In S. Kürschner, M. Habermann, & P. O. Müller (Eds.), *Methodik moderner Dialektforschung: Erhebung, Aufbereitung und Auswertung von Daten am Beispiel des Oberdeutschen.* Hildesheim/Zürich/New York, Olms.

Kisler, T., Reichel, U., Schiel, F., Draxler, C., Jackl, B., & Pörner, N. (2016). BAS Speech Science Web Services - an Update of Current Developments, In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, European Language Resources Association (ELRA).

Kleber, F. (2014). Partielle Neutralisierung des Stimmhaftigkeitskontrasts in zwei Varietäten des Deutschen (T. Krefeld & E. Pustka, Eds.). In T. Krefeld & E. Pustka (Eds.), *Perzeptive Linguistik: Phonetik, Semantik, Varietäten.* Stuttgart, Steiner.

Kleber, F. (2017a). Complementary length in vowel-consonant sequences: Acoustic and perceptual evidence for a sound change in progress in Bavarian German. *Journal of the International Phonetic Association.*

Kleber, F. (2017b). On the role of temporal variability in the acquisition of the German vowel length contrast, In *Proceedings of Interspeech 2017, Stockholm, Sweden*, Stockholm, Sweden. https://doi.org/10.21437/Interspeech.2017-1282

Kleber, F. (2018). VOT or quantity: What matters more for the voicing contrast in German regional varieties? Results from apparent-time analyses. *Journal of Phonetics*, *71*, 468–486. https://doi.org/10.1016/j.wocn.2018.10.004

Klingler, N., Kleber, F., Jochim, M., Pucher, M., Schmid, S., & Zihlmann, U. (2019). Temporal organization of vowel plus stop sequences in production and perception: Evidence from the three major varieties of German (S. Calhoun, P. Escudero, M. Tabain, & P. Warren, Eds.). In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, Canberra, Australia, Australasian Speech Science and Technology Association Inc.

Klingler, N., & Moosmüller, S. (2017). The different pronunciation of the homonyms lauter-lauter as a distinctive feature for two Viennese varieties, In *43. Österreichische Linguistiktagung*, Klagenfurt.

Knospe, W., Santen, L., Schadschneider, A., & Schreckenberg, M. (2002). A realistic two-lane traffic model for highway traffic. *Journal of Physics A: Mathematical and General*, *35*(15), 3369–3388. https://doi.org/10.1088/0305-4470/35/15/302

Kohler, K. J. (1977). The production of plosives. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel*, *8*, 30–110.

Kohler, K. J. (1979). Dimensions in the Perception of Fortis and Lenis Plosives. *Phonetica*, *36*(4-5), 332–343. https://doi.org/10.1159/000259970

Kohler, K. J. (1984). Phonetic Explanation in Phonology: The Feature Fortis/Lenis. *Phonetica*, *41*(3), 150–174. https://doi.org/10.1159/000261721

Kümmel, M. (2007). *Konsonantenwandel: Bausteine zu einer Typologie des Lautwandels und ihre Konsequenzen für die vergleichende Rekonstruktion* [OCLC: 778238999]. Wiesbaden, Reichert Verlag.

Kuznetsova, A., Brockhoff, P. B., & Bojesen Christensen, R. H. (2018). lmerTest: Tests in Linear Mixed Effects Models [R Package].

Labov, W. (1963). The Social Motivation of a Sound Change. *WORD*, *19*(3), 273–309. https://doi.org/10.1080/00437956.1963.11659799

Labov, W. (1994). *Principles of Linguistic Change. Volume 1: Internal Factors*. Malden, Oxford, Wiley-Blackwell.

Ladefoged, P. (2003). *Phonetic Data Analysis. An Introduction to Fieldwork and Instrumental Techniques.* Malden, MA u. a., Blackwell.

Lehtonen, J. (1970). *Aspects of Quantity in Standard Finnish*. Jyväskylä, K. J. Gummerus.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated Marginal Means, aka Least-Squares Means [R Package].

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, *48*(4), 839–862.

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory (W. J. Hardcastle & A. Marchal, Eds.). In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling*. Dordrecht, Springer Netherlands. https://doi.org/10.1007/978-94-009-2037-8_16

Manzo, G. (2014). Potentialities and Limitations of Agent-based Simulations (T. Matthews, Trans.). *Revue française de sociologie*, *Vol. 55*(4), 653–688.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking Rate and Segments: A Look at the Relation between Speech Production and Speech Perception for the Voicing Contrast. *Phonetica*, *43*(1-3), 106–115. https://doi.org/10.1159/000261764

Milroy, L. (2003). Social and linguistic dimensions of phonological change, In *Social Dialectology. In honour of Peter Trudgill.* Amsterdam/Philadelphia, John Benjamins Publishing Company.

Mitterer, H. (2018). The singleton-geminate distinction can be rate dependent: Evidence from Maltese. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *9*(1), Art. 6, pp. 1–16. https://doi.org/10.5334/labphon.66

Mooshammer, C., & Geng, C. (2008). Acoustic and articulatory manifestations of vowel reduction in German. *Journal of the International Phonetic Association*, *38*(02), 117–136. https://doi.org/10.1017/S0025100308003435

Moosmüller, S., & Brandstätter, J. (2014). Phonotactic information in the temporal organization of Standard Austrian German and Viennese dialect. *Language Sciencies*, *46*, 84–95.

Nagel, K., & Schreckenberg, M. (1992). A cellular automaton model for freeway traffic. *Journal de Physique I*, *2*(12), 2221–2229. https://doi.org/10.1051/jp1:1992277

Ohala, J. J. (1981). The listener as a source of sound change (C. S. Masek, R. A. Hendrick, & M. F. Miller, Eds.). In C. S. Masek, R. A. Hendrick, & M. F. Miller (Eds.), *Papers from the Parasession on Language and Behavior*. Chicago, Chicago Ling. Soc.

Ohala, J. J. (1993a). The phonetics of sound change (C. Jones, Ed.). In C. Jones (Ed.), *Historical Linguistics: Problems and Perspectives*. London, Longman.

Ohala, J. J. (1993b). Sound change as nature's speech perception experiment. *Speech Communication*, *13*(1), 155–161. https://doi.org/10.1016/0167-6393(93)90067-U

Pfalz, A. (1911). Phonetische Beobachtungen an der Mundart des Marchfeldes in Nieder-Österreich. *Zeitschrift für Deutsche Mundarten*, *6*, 244–260.

Pfalz, A. (1913). *XXVII. Mitteilung der Phonogramm-Archivs-Kommision. Deutsche Mundarten IV. Die Mundart des Marchfeldes.* (Vol. 6. Abhandlung). Wien, Hölder.

Pfeifer, W. (2005). *Etymologisches Wörterbuch des Deutschen* (8. Auflage). München, Deutscher Taschenbuchverlag.

Pickett, E. R., Blumstein, S. E., & Burton, M. W. (1999). Effects of Speaking Rate on the Singleton/Geminate Consonant Contrast in Italian. *Phonetica*, *56*(3-4), 135–157. https://doi.org/10.1159/000028448

Pierrehumbert, J. B. (2003). Phonetic Diversity, Statistical Learning, and Acquisition of Phonology: *Language and Speech*, *46*(2-3), 115–154. https://doi.org/10.1177/00238309030460020501

R Core Team. (2016). R: A Language and Environment for Statistical Computing. *Vienna, Austria*.

R Core Team. (2018). R: A Language and Environment for Statistical Computing. *Vienna, Austria*.

Raberto, M., Cincotti, S., Focardi, S. M., & Marchesi, M. (2001). Agent-based simulation of a financial market. *Physica A: Statistical Mechanics and its Applications*, *299*(1), 319–327. https://doi.org/10.1016/S0378-4371(01)00312-0

Raney, B., Voellmy, A., Cetin, N., Vrtic, M., & Nagel, K. (2002). Towards a Microscopic Traffic Simulation of All of Switzerland (P. M. A. Sloot, A. G. Hoekstra, C. J. K. Tan, & J. J. Dongarra, Eds.). In P. M. A. Sloot, A. G. Hoekstra, C. J. K. Tan, & J. J. Dongarra (Eds.), *Computational Science — ICCS 2002*, Springer Berlin Heidelberg.

Rathcke, T. V., & Stuart-Smith, J. H. (2016). On the Tail of the Scottish Vowel Length Rule in Glasgow: *Language and Speech*, *59*(3), 404–430. https://doi.org/10.1177/0023830915611428

Reinisch, E. (2016). Speaker-specific processing and local context information: The case of speaking rate. *Applied Psycholinguistics*, *37*(6), 1397–1415. https://doi.org/10.1017/S0142716415000612

Reubold, U., & Harrington, J. (2015). Disassociating the effects of age from phonetic change: A longitudinal study of formant frequencies (A. Gerstenberg & A. Voeste, Eds.). In A. Gerstenberg & A. Voeste (Eds.), *Language Development: The lifespan perspective*. Amsterdam/Philadelphia, John Benjamins Publishing Company.

Reubold, U., Harrington, J., & Kleber, F. (2010). Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Communication*, *52*(7), 638–651. https://doi.org/10.1016/j.specom.2010.02.012

Rosen, K. M. (2005). Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison. *Journal of Phonetics*, *33*(4), 411–426. https://doi.org/10.1016/j.wocn.2005.02.001

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English - ScienceDirect. *Journal of Phonetics*, *25*(4), 421–436.

Sankoff, G., & Blondeau, H. (2007). Language Change across the Lifespan: /r/ in Montreal French. *Language*, *83*(3), 560–588.

Schelling, T. C. (1971). Dynamic Models of Segregation. *Journal of Mathematical Sociology*, *1*(2), 143–186. https://doi.org/10.1080/0022250X.1971.9989794

Schikowski, R. (2009). Die Phonologie des Westmittelbairischen. https://doi.org/https://doi.org/10.5282/ubm/epub.10991

Scobbie, J. M. (2005). Interspeaker variation among Shetland Islanders as the long term outcome of dialectally varied input : Speech production evidence for fine-grained linguistic plasticity. *QMU Speech Science Research Centre Working Papers*, *WP-2*.

Seidelmann, E. (2013). Vokaldauer, Konsonantenschwächung und die Sprachrhythmik des Mittelbairischen (R. Harnisch, Ed.). In R. Harnisch (Ed.), *Strömungen in der Entwicklung der Dialekte und ihrer Erforschung. Beiträge zur 11. Bayerisch-Österreichischen Diaektologentagung in Passau, September 2010.* Regensburg, Edition Vulpes.

Seiler, G. (2005). On the development of the Bavarian quantity system. *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis*, *10*(1), 103–129.

Solé, M.-J., & Recasens, D. (Eds.). (2012). *The Initiation of Sound Change*. John Benjamins Publishing Company.

Stevens, M., Harrington, J., & Schiel, F. (2019). Associating the origin and spread of sound change using agent-based modelling applied to /s/-retraction in English. *Glossa: a journal of general linguistics*, *4*(1), 8. https://doi.org/10.5334/gjgl.620

Suomi, K., Meister, E., Ylitalo, R., & Meister, L. (2013). Durational patterns in Northern Estonian and Northern Finnish. *Journal of Phonetics*, *41*(1), 1–16. https://doi.org/10.1016/j.wocn.2012.09.001

Todd, S., Pierrehumbert, J. B., & Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, *185*, 1–20. https://doi.org/10.1016/j.cognition.2019.01.004

Torgersen, E., & Kerswill, P. (2004). Internal and external motivation in phonetic change: Dialect levelling outcomes for an English vowel shift. *Journal of Sociolinguistics*, *8*(1), 23–53. https://doi.org/10.1111/j.1467-9841.2004.00250.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9841.2004.00250.x

Treiber, M., & Kesting, A. (2013). Introduction (M. Treiber & A. Kesting, Eds.). In M. Treiber & A. Kesting (Eds.), *Traffic Flow Dynamics: Data, Models and Simulation*. Berlin, Heidelberg, Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-32460-4_1

Trudgill, P. (1986). *Dialects in contact*. Oxford, Basil Blackwell.

Trudgill, P. (1988). Norwich Revisited: Recent Linguistic Changes in an English Urban Dialect. *English World-Wide*, *9*(1), 33–49. https://doi.org/10.1075/eww.9.1.03tru

Varshney, L. R., & Sun, J. Z. (2013). Why do we perceive logarithmically? *Significance*, *10*(1), 28–31. https://doi.org/10.1111/j.1740-9713.2013.00636.x

Wiese, R. (1996). *The Phonology of German* (J. Durand, Ed.). Oxford, Oxford University Press.

Wiesinger, P. (1990). The Central and Southern Bavarian Dialects in Bavaria and Austria. (C. V. J. Russ, Ed.). In C. V. J. Russ (Ed.), *The dialects of Modern German*. London, Routledge.

Winkelmann, R. (2015). Managing Speech Databases with emuR and the EMU-webApp, In *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*.

Winkelmann, R., Harrington, J., & Jänsch, K. (2017). EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech \& Language*, *45*, 392–410. https://doi.org/10.1016/j.csl.2017.01.002

Wolfswinkler, K., & Harrington, J. (2020). Quality and quantity in the West-Central-Bavarian dialect – a comparison between children and adults [Abstract only], In *LabPhon*, Vancouver, BC, Canada. Abstract only.

Worth, N. A., & Nickels, T. B. (2011). Some characteristics of thin shear layers in homogeneous turbulent flow. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1937), 709–722. https://doi.org/10.1098/rsta.2010.0297