The transmission of unnatural phonology

Testing final voicing and final devoicing through iterated artificial language learning

Maria Evjen

Master’s thesis in Linguistics

LING4190

Department of Linguistics and Scandinavian Studies

University of Oslo

Spring 2021

Abstract

This thesis investigates why some phonological processes are typologically common, while others are typologically uncommon. One such typological asymmetry is that between final devoicing, a common, phonetically natural process, and final voicing, an uncommon, phonetically unnatural process. Two explanations have been proposed to account for this asymmetry: the channel bias account and the substantive bias account. Whereas proponents of the channel bias account hold that phonetic naturalness alone can explain typological asymmetries, proponents of the substantive bias account argue that phonetic naturalness is incorporated into grammatical constraints disfavouring the learning of unnatural processes.

In this thesis, the asymmetry between final devoicing and final voicing is investigated using an artificial language learning experiment. Participants were assigned to either of two conditions, DEVOICING or VOICING, and trained on artificial miniature languages displaying these processes. Moreover, the experiment made use of the iterated learning paradigm, in which participants formed diffusion chains and the languages were transmitted from generation to generation in the chains. This design was used with the intention of investigating whether final voicing was learned to a lesser extent than final devoicing, and whether such an effect was amplified through language transmission. Such effects would support a substantive bias view.

The results indicate that participants overall did not learn final voicing to a lesser extent than final devoicing. Participants learned the phonotactic restriction in their language, disallowing final voiceless obstruents in VOICING and final voiced obstruents in DEVOICING, equally well in the two conditions. A lower proportion of the items conforming to this phonotactic restriction were alternating in VOICING than in DEVOICING, but this difference was small, and it cannot be straightforwardly interpreted as a substantive bias effect.

As for the iterated learning design, there was no indication that a potential substantive bias effect was amplified throughout the generations. In most diffusion chains in both VOICING and DEVOICING, the relevant process was lost, and this experiment thus indicates that larger scale iterated learning experiments must be conducted to determine whether this design can provide new insights to the study of typological asymmetries in phonology.

Acknowledgements

There are a number of people I want to thank for helping me complete this project. First, I want to thank my two supervisors, Sverre Stausland Johnsen and Julian Kirkeby Lysvik, for their guidance at every step of this process. From the initial discussions about experimental design to the final draft, they have always been ready to discuss any issues and give me feedback. I am very grateful to Sverre that his door is always open when I have questions. I also want to thank him for always making me explain my arguments and my line of thinking, it undoubtedly made this thesis better. I am very grateful to Julian for sharing his valuable experiences with artificial language learning experiments, from stimuli and design to interpreting the results and reporting on the findings. I have also greatly appreciated all his encouragement.

I want to thank Jeremy Zehr, the developer of PennController for IBEX (PCIbex), for all his help with programming the experiment. When I began programming, I soon realised that the design of my experiment required programming skills far beyond my own. Through the support page for PCIbex, Jeremy replied to countless questions and came up with solutions I would never have been able to find myself. It is no exaggeration to say that conducting this experiment would not have been possible without Jeremy's help.

I would also like to thank Timo Roettger for his help with statistics and the analysis of the results.

I am very grateful to Henrik, Hedda, and Tove for making the long days in the office a lot more fun. I have really enjoyed all our discussions about our theses, linguistics in general, and everything else. Thank you also to Linn Iren, Kinjal, Eirik, and Alexander for sharing your experiences with conducting a research project and for your words of encouragement.

Thank you to my family for pilot testing the experiment and for listening to me going on about this project for the past year. Thanks also to my friends for encouraging me and for distracting me with lots of great days spent outside.

Last but not least, I want to thank everyone who took the time to participate in the experiment.

Contents

List of Figures

List of Tables

1 Introduction

It is widely recognised in typological research in phonology that certain structures, patterns, and processes are cross-linguistically more common than others. Such typological asymmetries can be found in the organisation of phonological inventories, in phonotactic patterns, and in phonological processes. A well-studied example of such an asymmetry is that between the processes *final devoicing* and *final voicing*. Final devoicing, the process by which voiced obstruents are devoiced word- or syllable-finally, is cross-linguistically very common, found across many languages and language families. Final voicing, on the other hand, the opposite process by which voiceless obstruents are voiced word- or syllable-finally, is extremely rare or even unattested. This asymmetry is found despite the two processes being structurally equally complex, and it thus raises the question of *why* certain phonological processes are more common than others.

In recent years, this question has been subject to debate, and two types of explanations are typically posited. On the one hand, we find those who hold that the underlying causes of typological asymmetries are cognition-external, owing to factors affecting the transmission of language across generations. This position is referred to as the *channel bias* account. On the other hand, there are those who ascribe typological asymmetries to cognitive mechanisms specific to language learning, referred to as the *substantive bias* account. In both accounts, *naturalness* underlies the explanations. A natural process is one which adheres to phonetic tendencies in articulation and perception, while an unnatural process works in opposition to such tendencies. By this definition, final devoicing is a natural process, while final voicing is unnatural.

Under the channel bias account, it is argued that phonetic naturalness alone can explain the observed typological patterns, as unnatural processes are less likely to be innovated and less likely to persist over time. According to this account, natural and unnatural processes should be learned equally well. Proponents of the substantive bias account, on the other hand, argue that naturalness must be incorporated in a learning bias specific to language. That is, speakers have implicit knowledge of phonetic pressures of articulation and perception, and they are biased against processes that work in opposition to these pressures, making unnatural processes more difficult to learn than their natural counterparts. The role of channel bias, then, is recognised by proponents of both accounts, but the role of substantive bias is disputed. The two accounts both predict that natural processes should be more common cross-linguistically than unnatural processes, but the underlying causes of this distribution are different.

Whether it is the substantive bias account or the channel bias account that best explains typological asymmetries is often investigated using artificial language learning (ALL) experiments. This method involves designing miniature artificial languages which exhibit the processes or structures under investigation. The languages are then taught to participants in an experimental setting, making it possible to study how well the languages are learned. ALL experiments thus offer insights not attainable using natural language stimuli, as it is possible to construct languages with structures or processes not found in natural language.

The ALL experimental method has been used to study a range of typological asymmetries, including that between final devoicing and final voicing. Most notable among the experiments investigating this asymmetry are Greenwood (2016), Glewwe (2019), and Lysvik (2020), who have found variable results with regard to the role of substantive bias and channel bias. While (Greenwood, 2016) argues for a channel bias view and against a substantive bias view, (Glewwe, 2019) and (Lysvik, 2020) conclude that their results indicate that the role of substantive bias cannot be excluded. The results in Lysvik (2020) are particularly relevant to the current study. In his Experiment 1, Lysvik found no evidence that word-final voiceless stops were preferred to word-final voiced stops. However, he did find indications that a language with final voicing was learned to a lesser degree than a language with final devoicing, in that participants learning final voicing to a larger extent avoided the alternation in their language. Lysvik argues that this result is best explained by a substantive bias against final voicing.

Lysvik (2020) contends that to further investigate whether this finding was in fact a substantive bias effect, a follow-up experiment should be run. He outlines an ALL experiment making use of the *iterated learning paradigm*, where participants form diffusion chains simulating the transmission of language across generations. By conducting such an experiment, it is possible to investigate whether the observed effect is replicated and whether the final devoicing and final voicing processes are equally stable throughout language transmission.

The aim of this thesis is to conduct such an iterated learning experiment isolating possible substantive bias effects. Because the substantive bias account and the channel bias account make similar predictions, it is necessary to separate the effects of the two types of biases to be able to attribute any observed effects to either explanation. As the effect found in Lysvik's (2020) Experiment 1 is hypothesised to be a substantive bias effect, it is this bias which is under investigation in the current experiment.

The experiment tested participants' learning of two artificial languages using a forced-choice task. Participants were divided into a VOICING condition, learning a language with final voicing, and a DEVOICING condition, learning a language with final devoicing. In each condition, participants formed diffusion chains consisting of five generations. Generation 1 in a diffusion chain was trained on the original stimuli, and the output of Generation 1, that is, the results of the forced-choice task, was used as the input to Generation 2, and so on. The development of the two languages over the course of the diffusion chains has implications for the role of substantive bias in shaping the typological asymmetry between final devoicing and final voicing. If there is a difference in learnability and stability between the VOICING condition and the DEVOICING condition when channel bias effects are excluded, it would provide evidence that a substantive bias does play a role in shaping the observed distribution in natural language.

The results in this thesis do not provide clear evidence for substantive bias. Participants did not avoid word-final voiced stops more than word-final voiceless stops, which would be the strongest indication of a substantive bias effect. There was a small difference between the conditions in that participants in VOICING alternated items to a lesser degree than those in DEVOICING. However, this difference diminished when different learning strategies among participants were taken into account, and I argue that it therefore does not constitute compelling evidence for the substantive bias account. As for the stability of the two processes throughout the diffusion chains, there was no difference between conditions. It rather appeared that both the final voicing process and the final devoicing process disappeared towards the end of the chains, and I argue that more extensive iterated learning experiments must be conducted in order to determine the usefulness of this paradigm in the study of typological asymmetries in phonology.

1.1 Outline of the thesis

The thesis is structured as follows. Chapter 2 introduces the theoretical background for the hypotheses tested in the current experiment. The term *naturalness* is defined, and the two approaches to accounting for typological asymmetries in phonology, the channel bias account and the substantive bias account, are presented. Subsequently, the phonetic underpinnings and the typological distribution of the final devoicing and final voicing processes are discussed. Chapter 3 introduces the artificial language learning method along with previous research on substantive bias, channel bias, and the asymmetry between final devoicing and final voicing. Furthermore, an outline of the current experiment is presented. Chapter 4 contains the methods section with a description of the participants, stimuli, procedure, and analysis used in the experiment. Chapter 5 presents the results, and Chapter 6 is a discussion of these results and their implications for the relevant typological asymmetry, the substantive bias account, and the experimental design. Chapter 7 concludes the thesis.

2 Theoretical background

In this chapter, I will start in Section 2.1 by discussing the term *naturalness* and why it is a central notion in research on typological asymmetries in phonology. Section 2.2 presents an overview of the biases which have been argued to shape phonological asymmetries, most notably channel bias and substantive bias. In Section 2.3, the two processes under investigation in this thesis, final devoicing and final voicing, are defined, and their distribution in natural language is discussed. Section 2.4 summarises the chapter.

2.1 Naturalness

When discussing phonological processes such as final devoicing and final voicing, the term *naturalness* is often used. Final devoicing is deemed a *natural* process, while final voicing is deemed *unnatural*. But the term *naturalness* has been used in several different ways in the literature, both in ALL research and more broadly, and so it is necessary to make explicit in which sense the term is being used.

Carpenter (2010), for example, takes natural processes to be those processes which are typologically common, while (Hayes & White, 2013) hold that natural processes are those which are either typologically common or facilitate perception or articulation. However, Greenwood (2016), Glewwe (2019), and Lysvik (2020) argue against including typological frequency in a definition of naturalness in phonology. In the approaches taken in these studies, *phonetic naturalness* alone determines whether a pattern is natural or unnatural. Greenwood (2016, p. 39) points out that while it is often the case that natural processes are in fact more typologically frequent than their unnatural counterparts, this pattern arises as a result of the phonetic naturalness of the process. That is, because phonetically unnatural processes are disfavoured, either by the physics of speech or by cognitive mechanisms, these are less likely to occur cross-linguistically. Nevertheless, it is not the case that all typologically frequent processes are phonetically natural, and in such cases, other factors than naturalness must underlie the observed distribution (Greenwood, 2016, pp. 38-39). Typological frequency is therefore not a satisfactory measurement of naturalness.

Similarly to Greenwood (2016), Glewwe (2019), and Lysvik (2020), this thesis uses a definition of naturalness as *phonetic* naturalness. That is, a natural process is one that is phonetically motivated by perceptual or articulatory pressures. Post-nasal voicing, for example, is a natural process because it arises from an articulatory tendency for stops following a nasal to be voiced. An unnatural process, then, is one that works in opposition to such phonetic pressures. Post-nasal devoicing would be an example of such a process. A third category is also possible under this definition, namely *unmotivated* processes (Beguš, 2018). These are processes that, while not directly motivated by phonetic pressures, also do not work in opposition to them. Beguš (2018, pp. 13-14) uses the Eastern Ojibwe process whereby /n/ becomes [ʃ] before front vowels as an example of an unmotivated process. This process is not motivated by phonetic tendencies, but it also does not work against them, as illustrated by the fact that the opposite process, /ʃ/ becoming [n] before front vowels, is not a natural process. In the case of post-nasal devoicing, on the other hand, this process must be unnatural because the opposite process has clear phonetic motivations.[1] In this thesis, only those processes which actively oppose phonetic pressures of articulation or perception are classified as unnatural.

With this definition of naturalness in place, it is possible to consider the distribution of natural and unnatural processes. As stated above, it is not necessarily the case that natural processes are cross-linguistically common while unnatural processes are uncommon, but there is a general trend in this direction. The underlying reason for this asymmetry is a subject of debate. While some argue that unnatural processes are impossible in natural language (Kiparsky, 2006), others argue that they are more or less strongly dispreferred (Lysvik, 2020), and others again argue that they are unlikely, but possible (Blevins, 2004).

It can appear relatively straightforward to determine whether a pattern is impossible in natural language by investigating whether it is attested or not. However, for two main reasons, this is not the case. Firstly, a pattern being unattested does not necessarily mean that it does not exist or has not existed in human language. Far from all languages existing today have been described and analysed, and so it is not possible to exclude the possibility that unnatural processes can be found in a language which as of yet remains undescribed. In addition, we cannot exclude the possibility that unnatural processes could be found in languages which are now extinct and have been so for thousands of years.

Secondly, whether or not a given process is found in a language is not an undisputable fact. As we will see below in Section 2.3, different analyses of the same data can result in widely discrepant conclusions. As a result, Blevins (2004, 2006) considers final voicing to be

---

[1] See Stanton (2017, pp. 44-48) for an account of post-nasal devoicing as perceptual enhancement, and Beguš (2018, pp. 22-28) for arguments against this account.

rare, but attested in Somali and Lezgian, while Kiparsky (2006, 2008), analysing the same data in a different way, concludes that it is in fact unattested. Such conflicting analyses illustrate the need to test hypotheses regarding typological asymmetries experimentally.

The view of unnatural processes as either impossible, disfavoured, or possible typically fall into one of the two main explanations for typological asymmetries, the substantive bias approach and the channel bias approach. While the first and the second viewpoints are held by proponents of the substantive bias approach, the third viewpoint is found with proponents of the channel bias approach. In the next section, these two approaches will be presented in more detail.

2.2 Biases in typological asymmetries

The biases argued to play a role in shaping typological asymmetries can be broadly divided into cognition-external biases and cognition-internal biases. Cognition-external biases are separate from our cognitive system and result from the physical pressures of articulation and perception. These biases, often referred to as channel biases, affect the transmission of language from generation to generation, resulting in the observed typological patterns. Importantly, channel biases do not affect the learnability of a pattern. As Greenwood (2016, p. 4) points out, if the differences in perception or articulation favouring one pattern over another were somehow reduced or removed, language users should under the channel bias account learn the two patterns equally well. The channel bias account is one of the two main approaches to typological asymmetries in phonology.

Cognition-internal biases, on the other hand, are those which, as the name suggests, are mechanisms of our cognitive system. Such biases, often referred to as analytic biases (Moreton, 2008), affect how humans learn patterns and rules. Analytic biases can be further divided into domain-general biases and biases which are specific to language learning, called substantive bias. Domain-general mechanisms, such as a bias favouring formally simpler patterns over formally more complex patterns, are found in for example the visual domain in addition to the linguistic domain (Moreton et al., 2017). The role of such biases in pattern learning has been documented in a number of studies and is largely accepted in the literature (Moreton & Pater, 2012a). The role of substantive bias, however, is disputed, the evidence for it being much more variable (Moreton & Pater, 2012b). Nevertheless, the substantive bias account constitutes the other main explanation for typological asymmetries in phonology.

Both channel bias and substantive bias are based on phonetic naturalness as discussed in Section 2.1, but this notion is incorporated into the two types of biases in different ways. Under the channel bias account, unnatural processes are rare because they are phonetically unmotivated and work in opposition to phonetic tendencies, making them less likely to occur and less likely to be stable throughout language transmission. In the substantive bias approach, phonetic tendencies also underlie the rarity of unnatural processes, but these tendencies are encoded in the phonological grammar. Unnatural processes are under this view rare because they violate grammatical constraints. Importantly, while the channel bias account rules out the role of a substantive bias, considering the postulation of such a bias a doubling of explanatory effort, the substantive bias account does not exclude the role of channel biases, but holds that these are not in themselves sufficient to account for typological asymmetries. Below, the two explanations are presented in detail; Section 2.2.1 discusses channel bias, while Section 2.2.2 discusses substantive bias. In Section 2.2.3, two cognition-general analytic biases, complexity bias and paradigm uniformity bias, are presented, and their relation to the two main approaches is discussed.

2.2.1 Channel bias

The channel bias approach holds that phonetic motivation is sufficient in explaining why natural phonological processes are cross-linguistically more common than unnatural processes. Natural processes have clear phonetic precursors causing systematic misperceptions and consequently misproductions throughout language transmission (Ohala, 1981, 1993). These misperceptions and misproductions can become *phonologised*, that is, gradient phonetic tendencies can be interpreted as categorical and incorporated into the phonological grammar. In this way, phonetic pressures are the driving force of sound change (Blevins, 2004).

Moreton (2008, p. 86) uses vowel harmony as an example of a process which can undergo phonologisation. Coarticulation of vowels in proximity to each other has a phonetic motivation, as it decreases the articulatory effort. This coarticulation is often compensated for in perception, so that the listener is able to recover the intended quality of the vowel. However, these perceptual mechanisms can fail, leading the listener to interpret the vowel harmony as a phonological process, intended by the speaker. When the listener turns speaker, this misperception will affect their own production of the same form. If this misperception and following misproduction occurs in the interaction between a range of speakers, vowel harmony can be phonologised, from a gradient phonetic tendency to a categorical rule in the relevant language (Kiparsky, 1995).

For the same reason that natural processes are common, unnatural processes are uncommon. These do not have phonetic precursors leading to the misperceptions and misproductions discussed above. On the contrary, they work in opposition to such precursors. Taking post-nasal devoicing as an example, this process works in opposition to the articulatory tendency to voice post-nasal stops (Beguš, 2018). There is no articulatory tendency to devoice stops after nasals, and so there is no variability in speakers' productions leading the listener to interpret post-nasal voiced stops as anything other than voiced. For this reason, post-nasal devoicing and other unnatural processes are rarely innovated. However, it is not impossible for such processes to arise, for example through telescoping, that is, through a series of independent natural processes. But if they are in fact innovated, they are not likely to remain stable throughout language transmission, as the phonetic pressures work against them. In the case of post-nasal devoicing, then, the pressures of articulation will lead speakers to variably produce more or less voiced obstruents post-nasally, leading to the type of misperceptions and misproductions discussed above and eventually to the loss of the post-nasal devoicing process over time.

A consequence of the theory of typological asymmetries as solely deriving from channel bias is that unnatural phonological processes are not impossible to learn. Certain processes are more unlikely to arise, but they are not dispreferred by our cognitive system on naturalness grounds, in contrast to what is contended by proponents of the substantive bias approach. As will be discussed below in Section 2.2.3, certain processes are thought to be dispreferred by cognitive mechanisms also within the channel bias approach, but such mechanisms are not subsumed by substantive bias.

Furthermore, proponents of the channel bias approach argue strongly against the doubling of explanatory effort in the substantive bias approach. Within this latter approach, the constraints that make up the grammar are grounded in phonetic pressures, thus recognising the role of channel bias. Yet, as Greenwood (2016, p. 30) points out, the substantive bias approach nevertheless posits grammatical constraints which are tasked with suppressing unnatural structures, the same structures already recognised to be suppressed by channel bias. The phonetic principles are thus stated a second time in the definition of substantive bias (Moreton, 2008, pp. 87-88). It appears, then, that if channel bias can in fact successfully predict typological asymmetries, simplicity favours this approach over the substantive bias approach.

2.2.2 Substantive bias

As stated above, substantive bias is a cognition-internal bias specific to language. This bias pertains to the learning of linguistic structures, and proponents of the substantive bias approach hold that humans have a cognitive predisposition to favour the learning of natural patterns and impede the learning of unnatural patterns. These predispositions arise from language users' implicit knowledge of the physical properties of speech, and this knowledge is encoded in grammatical constraints. Whether these constraints are innate, as in the typical understanding of Universal Grammar (UG), or learned in first language acquisition is a subject of debate (Hayes, Kirchner, & Steriade, 2004; Bermúdez-Otero & Börjars, 2006), but it is not essential to the definition of substantive bias. Regardless of whether constraints are innate or acquired, they are shared by all humans and shape the distribution of natural and unnatural processes in typology.

As an example of how substantive bias constrains phonological patterns, we can take weight-sensitive stress, as discussed by Greenwood (2016). A natural pattern in the domain of weight-sensitive stress is one in which syllables with long rhymes attract stress, while an unnatural pattern would be the opposite, that only syllables with short rhymes can attract stress (Greenwood, 2016, p. 4). Underlying this asymmetry is the phonetic tendency for longer syllables to have more perceptual energy than their shorter counterparts, making them perceptually more prominent regardless of stress assignment. In a bisyllabic word where the first syllable has a long rhyme, while the second has a short rhyme, then, the first syllable will be more prominent. If the syllable with the long rhyme also receives phonological stress, its role as the most prominent syllable will be enhanced and perceptually clear to the listener. If the syllable with the short rhyme receives phonological stress, on the other hand, the prominence of the two syllables will be ambiguous to the listener, as both syllables are prominent (Greenwood, 2016, p. 12).

A theoretical framework incorporating the naturalness effects of weight-sensitive stress as grammatical constraints is Optimality Theory (Prince & Smolensky, 1993/2004). In this theory, the phonetic motivation for stressing syllables with long rhymes and against stressing syllables with short rhymes is encoded in the grammar with the markedness constraint

WEIGHT-TO-STRESS (Greenwood, 2016, p. 6). This constraint prefers the natural pattern where the syllable with the long rhyme is stressed and disprefers the opposite, unnatural pattern. Supposing that the grammar also contains constraints determining the language's default stress pattern, for example that feet are left-aligned and trochaic, WEIGHT-TO-STRESS must be ranked above these constraints in order for the natural pattern to emerge in cases where the syllable with the long rhyme would not receive default stress. If the other constraints were ranked higher than the WEIGHT-TO-STRESS constraint, syllables with long rhymes would not be able to attract stress away from the default stressed syllable. The unnatural pattern, however, would not arise because there is no constraint in the grammar favouring stress on a syllable with a short rhyme. This prediction is borne out in typology, where the unnatural pattern is unattested (Greenwood, 2016, p. 4).

In Section 2.2.1, we saw that the channel bias account does not predict unnatural patterns to be impossible in natural language. In the substantive bias approach, researchers differ as to whether they consider unnatural patterns to be impossible or (more or less strongly) disfavoured. The former group are proponents of a strong substantive bias, where unnatural patterns cannot be learned (Kiparsky, 2006, 2008). The latter group favours a weak substantive bias explanation, where unnatural patterns can be learned, but not as well as their natural counterparts (e.g. Lysvik, 2020).

Regardless of whether unnatural processes are taken to be impossible or dispreferred, however, both proponents of a hard substantive bias and proponents of a soft substantive bias hold that an explanation based solely on channel bias fails to account for the observed typological asymmetries. In the case of final voicing, for example, Blevins (2004, p. 67), a proponent of the channel bias approach, lists a number of ways in which this process could arise through rule telescoping or analogical levelling. Kiparsky (2006, 2008), however, supporting a substantive bias approach, argues that despite these scenarios, final voicing does in fact not arise. The explanation, he contends, must be that humans are cognitively biased against this process, blocking it even in those cases where it could arise. Accordingly, the channel bias approach fails to explain typological asymmetries in their entirety.

2.2.3 Analytic biases

As mentioned in Section 2.2, substantive bias is one type of, but not the only, cognition-internal bias, often called analytic bias (Moreton, 2008). These are biases that pertain to general pattern learning, both linguistic and non-linguistic, and they are not grounded in phonetic naturalness. One such bias is a *complexity* bias, causing learners to prefer structurally simpler patterns over structurally more complex patterns. Complexity is determined by two things: the number of dimensions necessary to characterise a pattern, and the relationship between these dimensions (Greenwood, 2016, p. 33). In terms of phonological features, a process which refers to fewer features is simpler than a process which refers to more features.

Effects of complexity bias have been found both in phonology (e.g. Moreton, 2008) and in non-linguistic domains such as the visual domain (Moreton et al., 2017). Importantly, in not pertaining to naturalness, complexity bias is not considered a doubling of explanatory effort by proponents of the channel bias view in the same way as substantive bias. A complexity bias affecting all types of pattern learning is thus recognised by adherers of the channel bias account as well, for instance by Greenwood (2016).

Furthermore, with the evidenced effect of complexity bias in pattern learning, a point of caution arises when discussing and comparing natural and unnatural processes. Multiple studies aiming to uncover effects of substantive bias by comparing natural and unnatural processes do not take the complexity of the processes under comparison into account. Glewwe (2019) uses the study in (Wilson, 2003) as an example. Wilson (2003) found that participants in an artificial language learning experiment learned a nasal assimilation process where a suffix surfaced as [-na] after a nasal consonant and [-la] otherwise better than a process where the suffix surfaced as [-na] after dorsal consonants and as [-la] otherwise. Here, Wilson (2003) takes the former process to be more natural than the latter. However, the former process is also simpler than the latter process, relating two instances of [nasal] compared to one instance of [nasal] and one instance of [dorsal].

When investigating hypotheses regarding naturalness, then, it is important to control for complexity. In the current experiment, the two processes under investigation are final devoicing and final voicing. These two processes are equally complex, both of them relating two instances of the feature [voice], and thus they allow for an investigation of naturalness effects where complexity is not a confounding factor.

Another analytic bias which can influence pattern learning is a paradigm uniformity bias. This bias leads learners to prefer non-alternating paradigms where all forms have the same realisation, and it is typically found in children who say for instance *singed* as the preterite form of *sing* instead of *sang*. This bias can affect spontaneous productions as in the previous example, and it can lead to diachronic change, as seen in the long-term tendency to regularise irregular verbs in English (Lieberman et al., 2007).

Taking final devoicing as a phonological example, a paradigm uniformity bias can lead language users to prefer a system in which consonants never alternate, and the stem consistently either has /p/ or /b/ as the final consonant. This bias can thus compete with pressures of phonetic naturalness, which on their part favour final devoicing. As was the case with complexity bias, then, it is important to take paradigm uniformity bias into account as a possible confounding factor when investigating naturalness in phonological processes.

In the literature, this bias has mainly been discussed in relation to linguistic patterns (Do, 2013, 2018; Lysvik, 2020), but it is not given that this bias is not also domain-general. As was the case with the number of phonological features, intra-paradigm variation also adds to the complexity of a pattern. A system in which there is no alternation is necessarily formally simpler than a system with alternations, and this is expected to be the case also in for example the visual domain.

2.3 The voicing asymmetry

In previous sections, I have mentioned that the typological asymmetry under investigation in this thesis is that between final devoicing and final voicing, from here on referred to as *the voicing asymmetry*. I have stated that final devoicing is a natural process, while final voicing is an unnatural process, and in Section 2.1, I argued that typological data alone is not sufficient in determining whether final voicing is in fact attested or not in natural language. In this section, both of these issues will be further discussed. Section 2.3.1 presents the phonetic underpinnings of the final devoicing process as opposed to its unnatural counterpart, final voicing. Section 2.3.2 discusses the evidence for final voicing in natural language.

2.3.1 Naturalness in the voicing asymmetry

In Section 2.1, I mentioned that the two opposite processes final devoicing and final voicing differ with regard to phonetic naturalness. While final devoicing is phonetically natural, that is, it adheres to phonetic pressures of articulation and perception, final voicing is phonetically unnatural, which means that it is not only phonetically unmotivated, it also works in opposition to phonetic pressures.

Final devoicing, on the one hand, can be defined as a process in which voiced obstruents are devoiced in the syllable-final or word-final position. This process is found in languages such as Dutch (Grijzenhout & Krämer, 2000) and Turkish (Kopkalli, 1993), and an example of this process as found in Dutch is shown in Table 2.1.

Table 2.1 From Harris (2009, p. 11).

Singular	Plural	Gloss
ba[t]	ba[d]en	‘bath’
dui[f]	dui[v]en	‘dove’
hui[s]	hui[z]en	‘house’

Final voicing, on the other hand, is the process in which voiceless obstruents are voiced in the syllable- or word-final position. Table 2.2 illustrates the final voicing process in Lezgian, as described by Yu (2004).

Table 2.2 From Yu (2004, p. 76). No glosses of the grammatical category of suffixes are given by Yu.

pa[b]	pa[p]a	‘wife’
ga[d]	ga[t]u	‘summer’
le[g^w]	le[k^w]e	‘tub’

Both of these processes thus impose a phonotactic restriction on which segments are allowed in the final position. A language with a final devoicing process only permits voiceless obstruents in this position, while a language with a final voicing process only permits voiced obstruents in this position. The alternations found in Table 2.1 and Table 2.2 arise when there is a contrast between voiced and voiceless obstruents in one part of the paradigm, as in the plural forms in the Dutch examples, while the neutralisation of this contrast is found in another part of the paradigm, in this case in the singular form.

Final devoicing is, as stated, phonetically motivated. First of all, there are clear phonetic pressures favouring voiceless obstruents over voiced obstruents, both in the final position and in general. In order for voicing to occur, the vocal folds must vibrate, and to obtain this vibration, the subglottal pressure must be sufficiently high, and the supraglottal pressure must be sufficiently low. In the articulation of stops, the obstruction of the oral cavity can lead to the supraglottal pressure, that is, the pressure between the glottis and the oral constriction, being too high. If this is the case, the difference between the subglottal and the supraglottal pressure will be too small, and voicing will fail (Blevins, 2004, p. 104).

Moreover, the subglottal pressure decreases throughout the utterance, and so it is at its lowest utterance-finally, inhibiting voicing. In this position, speakers also have a tendency to spread their vocal folds in anticipation of the following pause with the same effect of preventing voicing. As Blevins (2004, pp. 104-105) points out, there is also a phonetic tendency for segments in the utterance-final position to be lengthened. As it is more difficult to sustain voicing in obstruents that have a longer duration, Blevins argues, this lengthening can result in a voiceless production of voiced obstruents. It is important to note that these effects are found utterance-finally and not necessarily word-finally, and the devoicing process is thus to a greater degree motivated in the utterance-final position. However, Myers & Padgett (2014) find that speakers generalise utterance-final devoicing to word-final devoicing, which can explain why final devoicing often applies word-finally.

This last point is also true of the perceptual motivation for final devoicing. Blevins (2004, pp. 104-105) argues that utterance-final lengthening not only inhibits the production of voiced obstruents, but also facilitates the perception of voiced obstruents as voiceless. That is, voiced obstruents are generally shorter than voiceless obstruents, and so lengthening can contribute to utterance-final obstruents being perceived as voiceless. Again, this utterance-final effect can be generalised to the syllable-final or word-final position, giving rise to the final devoicing process as found in for example Dutch.

Voiced stops in the final position are thus both difficult to articulate and likely to be misperceived as their voiceless counterpart, making final devoicing a natural process. For these same reasons, final voicing is an unnatural process. While phonetic pressures can lead to the devoicing of final obstruents even as a gradient phonetic tendency in languages without a phonological devoicing process, final voicing actively works in opposition to these pressures. By voicing a final voiceless obstruent, final voicing increases the articulatory effort, and there is also no tendency for final voiceless obstruents to be perceived as voiced. Final voicing is thus not simply unmotivated by phonetics, it is phonetically unnatural.

In a channel bias account, these phonetic pressures are themselves sufficient to explain why final voicing is extremely rare. While final devoicing is subject to the type of misperceptions and misproductions which can lead to phonologisation, as discussed in Section 2.2.1, final voicing is not subject to phonologisation because there is no phonetic tendency to voice final obstruents. In a substantive bias account, on the other hand, language users' knowledge of the phonetic unnaturalness of voicing final obstruents is encoded as a grammatical constraint, inhibiting the learning of such as process.

2.3.2 Final voicing in natural language

Due to the phonetic unnaturalness of final voicing, it is expected to be a rare process in natural language, both within a channel bias approach and a substantive bias approach. This prediction is borne out. While final devoicing is typologically well-attested, final voicing is at best very uncommon (Blevins, 2004, pp. 108-109). This latter process has been argued to be found only in a handful of languages, most notably Somali (Blevins, 2006) and Lezgian (Yu, 2004). The details of the final voicing process as found in these languages will not be discussed here, and the aforementioned references are recommended for the full analyses. The important point is that several researchers have proposed analyses including a final voicing process. However, these analyses have been criticised, most notably by Kiparsky (2006, 2008), who argues that the processes found in these languages are not best explained as final voicing. It is argued that the process in Lezgian, for example, is better described as initial degemination and devoicing (Kiparsky, 2006, pp. 232-234). This example illustrates that the analysis of a single process can differ greatly depending on the researcher.

Turning to the different approaches to explaining typological asymmetries, the analysis of processes such as the one found in Lezgian has implications for the understanding of the role of channel bias and substantive bias in phonology. Proponents of the channel bias approach take the evidence from these languages as true examples of final voicing, albeit in somewhat restricted environments (Blevins, 2004, p. 110). Proponents of the substantive bias approach, on the other hand, argue against analysing the processes found in these languages as final voicing (Kiparsky, 2006; Lysvik, 2020). In this latter view, competing analyses involving other phonological processes are proposed, and it is argued that simplicity favours these analyses.

A strong substantive bias approach in which unnatural processes are argued to be impossible in natural language must necessarily reject the analyses of the processes in Somali, Lezgian, and other languages as final voicing. This is the case in Kiparsky (2006). In a weak substantive bias view, however, whether or not these processes are in fact examples of final voicing is not essential to the explanation. In this approach, unnatural processes are not necessarily taken to be impossible, but highly improbable. Therefore, accepting these examples as processes of final voicing does not contradict the predictions of a weak substantive bias account.

Here, I do not argue for or against the analysis of these processes as true examples of final voicing. If a strong substantive bias view were to be taken, this question would be of importance in the current study. However, as will be discussed below in Chapter 3, it appears based on evidence from artificial language learning research that a strong substantive bias account is too restrictive. Therefore, I hold that either a channel bias view or a weak substantive bias view best explains the observed typological asymmetries. As these both predict final voicing to be very rare, but not impossible, leaving the question of whether final voicing is attested open does not limit the explanation to either approach.

From the discussion of final voicing in natural language in Kiparsky (2006, 2008), Blevins (2004, 2006), Yu (2004), and others, it appears that the currently available data does not allow for a single, indisputable analysis. These data and these analyses thus cannot successfully settle the question of whether final voicing exists or not. What the controversy regarding final voicing demonstrates is that any analysis based on natural language data is at the mercy of both the initial description of the language under investigation and the subsequent analysis. This issue therefore illustrates the shortcomings of using only natural language data when studying typological asymmetries and the necessity of conducting experimental research to complement natural language analyses.

2.4 Summary

In this chapter, I have presented the theoretical background for the experiment conducted in this thesis. I discussed the importance of defining the notion of naturalness when studying typological asymmetries regardless of theoretical stance, and I presented the leading explanations for such asymmetries, the channel bias account and the substantive bias account. Furthermore, the phonetic pressures underlying the typological asymmetry investigated in the current experiment, that between final devoicing and final voicing, were presented, and I argued that based on natural language data alone, it is not possible to reach a conclusion as to which bias gives rise to this asymmetry. In the next chapter, I will present an experimental approach to the investigation of biases in phonology, which I argue can be a valuable supplement to natural language analyses.

3 Artificial language learning

This chapter presents the artificial language learning (ALL) method and previous research which has made use of this method. Section 3.1 introduces ALL and the most common experimental designs used in ALL experiments, followed by a discussion of both the advantages and the disadvantages with this method. In Section 3.2, previous ALL experiments are examined, and their implications for the substantive bias account, the channel bias account, and the voicing asymmetry are discussed. Section 3.3 presents an outline of the current experiment, and finally, Section 3.4 is a summary of the chapter.

3.1 The artificial language learning method

The discussion of the typology of final voicing in the previous section demonstrates the need to test hypotheses regarding typological asymmetries empirically, and increasingly, such experimental research has made use of the artificial language learning (ALL) method. In ALL experiments, participants learn artificial miniature languages designed to test for a specific linguistic feature. The artificial languages tend to be small in size and designed in such a way that the linguistic feature of interest is the only or one of few variable elements. Such experiments have been conducted to investigate typological asymmetries in different linguistic subfields, including syntax, morphology, and phonology, and common to the experimental design is a training phase where participants are exposed to the pattern to be learned and a testing phase where they are tested on how well they learned this pattern.

Lysvik's (2020) experiments investigating the voicing asymmetry serve as an example of a typical ALL experiment in phonology. In these experiments, participants were assigned to either of two conditions, one learning final devoicing and one learning final voicing. In the training phase, participants were trained on the relevant language for their condition through the use of auditory stimuli and images. In the testing phase, participants completed both a forced-choice task and a production task with the aim of determining whether there was a difference between the two conditions.

This template is found in most ALL experiments, although the details can vary. For example, while experiments pertaining to phonology tend to include auditory stimuli, this is not necessarily the case for experiments investigating other subfields in linguistics. Moreover, most experiments only make use of a forced-choice task, rather than both a forced-choice task and a production task in the testing phase. The forced-choice task has the advantage that it is easier to analyse, but the disadvantage that participants can only choose between the given alternatives. The production task has the advantage of not constraining participants' responses, but it is both practically more difficult to conduct and more difficult to analyse because of a high number of confounds.

Four main designs are typically found in ALL experiments, as sketched out by Culbertson (2015). Firstly, the *ease-of-learning paradigm* involves two or more groups that are taught minimally differing languages. The performances of the different groups are then compared to each other and to chance to find out which language was learned better. In the *poverty-of-the-stimulus paradigm* and the *mixture-shift paradigm*, participants learn a language with a certain linguistic structure, but there is not enough information in the input to infer how the language deals with this structure. It is then possible to test participants' strategy for handling this lack of information. In the poverty-of-the-stimulus design, this information is simply lacking, whereas in the mixture-shift paradigm, the information is ambiguous, showing different patterns within the same language. Finally, the *iterated learning paradigm* can make use of any of the three designs above, but crucially, this paradigm also simulates language change. Participants are divided into generations, and the results from the first generation of participants are used as input to the next generation and so on, resulting in a diffusion chain similar to the one found in natural language.

Another factor which must be considered when designing an ALL experiment is whether the learning of the relevant patterns is implicit or explicit. That is, participants are either instructed on the pattern they will be taught (explicit learning), or they are exposed to the pattern without explanation, in which case they have to infer the pattern from the input (implicit learning). Because natural language learning is implicit in nature, most ALL experiments involve this type of learning.

What is more, it is important to consider whether to teach participants the languages up to a set threshold or not. This would be implemented by including two separate testing phases in the experiment, one in which participants are tested on how well they have learned the items in the artificial language, and one in which they are tested on the patterns under investigation. If participants do not perform at or above a set limit in the first test, they will have to complete more training before moving on to the main testing phase. The disadvantage with doing this is that in natural language learning, there is no feedback (Baker, 1979, p. 536). It is also important that this threshold is not too high, in which case participants might have received enough training to completely master the patterns. Skoruppa & Peperkamp (2011) found that, with enough training, participants can learn completely arbitrary patterns, and so it is important to withhold training enough for participants' underlying biases to emerge. Yet, if participants do not learn the items to a set threshold, this could impede the testing of the typological pattern. As will be discussed below, this was the case in Lysvik's (2020) Experiment 2.

Having presented an overview of the ALL method and the most common design choices in ALL experiments, arguments both in favour of and against this method become apparent. Let us first discuss the advantages of using the ALL method to investigate research questions concerning typological patterns, as opposed to relying on typological data and natural language experiments alone.

First of all, ALL experiments allow us to find the effects of minimal differences in the input while all other factors are controlled for. Typically, artificial languages are designed in such a way that the linguistic feature of interest is the only variable element, and so it is possible to isolate this specific feature. In this way, other linguistic processes, of which there are plenty in natural language, can be excluded as possible confounding factors (Fedzechkina et al., 2016). In the case of the voicing asymmetry, then, it is possible to design experiments such as the ones conducted by Lysvik (2020), where the final devoicing process and the final voicing process are the only variable elements in the artificial languages. Moreover, as Fedzechkina et al. (2016, pp. 213-214) point out, ALL experiments differ from natural language studies in that it is possible to have control over the amount of exposure to the pattern under investigation participants receive. Differences in the amount of input received can therefore also be excluded as a possible confound.

Secondly, ALL experiments offer a method for investigating the learning of patterns that are not found in natural language. When using typological data, it is not possible to find negative evidence excluding the existence of a process or feature. As discussed in Chapter 2, typological data showing that a process such as final voicing is not found or at best very rare could be a symptom of cognitive biases against this process, or it can be due to channel bias or simply to coincidences of history. ALL experiments, on the other hand, can to a larger degree test specific hypotheses about the reasons why rare patterns are rare, because it is possible to design artificial languages in which these patterns are found. The behavioural data resulting from these experiments provide insights into how learners handle rare or unattested patterns such as final voicing, which cannot be obtained using typological data or natural language experiments alone.

There are, however, certain disadvantages to the ALL method which must be taken into account when conducting ALL experiments and analysing the results. A central concern regarding ALL experiments is whether it is actually possible to tap into language acquisition in this type of experiment. Compared to natural language, artificial languages are much simpler, and learners are exposed to much less input (Fedzechkina et al., 2016, pp. 221-222). The learning strategy can also be different from natural language acquisition in those experiments which employ an explicit learning strategy (Culbertson et al., 2012). As mentioned above, speakers do not receive explicit explanations in natural language learning, especially in first language acquisition, but rather have to learn linguistic patterns implicitly by deduction.

Yet, a number of studies have shown that the learning of artificial languages is actually comparable to the learning of natural languages. Sanders et al. (2015) and Avcu et al. (2019) have used EEG to show that artificial language processes activated the same ERP (event related potential) effect as natural language processes, although only when learned implicitly rather than explicitly. Furthermore, Ettlinger et al. (2015) found a positive correlation between artificial language learning and second language learning, which further supports the claim that ALL experiments can tap into language learning mechanisms.

If ALL experiments are in fact able to access the same mechanisms as natural language learning, then, it appears that it more closely resembles second language (L2) learning than first language (L1) learning. If the experiment is conducted on adults, this is necessarily the case, and for ALL experiments to access L1 acquisition mechanisms, it would thus be necessary to conduct experiments with infants and child learners. However, such experiments are more time-consuming and therefore less common than experiments with adult learners. This remains a point of caution when interpreting the results from ALL research.

Another concern when conducting ALL experiments is the possibility of transfer from participants’ L1. First language influence has been found in the learning of artificial languages in that participants can show a preference for structures which are dissimilar to those found in their L1, thus suppressing knowledge about their L1. Baer-Henney (2019), for example, found that German and Mandarin L1 speakers were more likely to accept items in an artificial language if they were less similar to their L1. Other studies, such as Greenwood (2016), have found a possible influence from L1 in that participants use knowledge about their L1 when interpreting the pattern in the artificial language (see Section 3.2.2). These studies illustrate that participants' L1 is an important variable to take into account, and that the same ALL experiments should be conducted on speakers with different first languages. Another way of mitigating transfer effects from the L1 is to design the artificial language so that it does not test for processes that are found in the L1 of the participants, although this limits the scope of possible linguistic features that can be investigated (Fedzechkina et al., 2016, p. 224).

A final concern, presented by Greenwood (2016), is that language learning in an experimental setting often does not accurately represent language learning in a naturalistic setting. While natural language is often fast and casual, Greenwood argues, the stimuli used in ALL experiments tend to be hyperarticulated and slow. To mitigate this, Greenwood (2016) included two speech rates in her final devoicing experiment, a CAREFUL condition and a CASUAL condition, meant to resemble these differences in natural and artificial languages. Greenwood contends that only by controlling for these differences can true effects of substantive bias or channel bias be detected.

Based on this discussion of the advantages and disadvantages to the ALL method, I argue that although there are shortcomings to the method which must be taken into account, it nevertheless provides a way of investigating typological patterns while controlling for confounding factors. It is especially useful because it allows the researcher to study speakers' behaviour when faced with unnatural processes. This insight cannot be obtained using natural language data alone. Furthermore, if ALL research does in fact access L2 rather than L1 learning mechanisms, the resulting behaviour should still be of interest to the study of language learning biases to the same extent as L2 learning (Glewwe, 2019, p. 6)

3.2 Previous ALL research

In Section 3.1, the ALL method and its possibilities and limitations were presented, and it was argued that ALL experiments, when interpreted with caution, can be a valuable complement to the typological method in the investigation of the biases underlying typological asymmetries. In this section, an overview of previous ALL research is presented. Section 3.2.1 reviews the evidence for analytic biases, substantive bias, and channel bias found in ALL experiments so far. Section 3.2.2 discusses the experiments testing for biases in the voicing asymmetry specifically, while in Section 3.2.3, previous experiments making use of the iterated learning design are presented.

3.2.1 Evidence for biases in phonology

Starting with analytic biases which are not specific to language learning, ALL experiments have obtained strong evidence for the role of such biases. Complexity bias in particular, that is, the preference for featurally simpler patterns over featurally more complex ones, has been found to affect participants' behaviour when learning artificial languages. As stated above, however, this bias is not under investigation in the current experiment, and so it will not be discussed in further detail. For a general discussion and overview of complexity bias in ALL experiments, see Moreton & Pater (2012a).

As for the paradigm uniformity bias discussed in Section 2.2.3, this has not been as extensively researched. However, this bias could potentially influence performance in the current experiment, as it is possible that participants prefer to make paradigms uniform by avoiding the voicing alternations in the final voicing and final devoicing processes.

The effect of a paradigm uniformity bias was found in Do (2018), where Korean 6-7 year old children dispreferred a verbal morphology alternation when learning their native language. Do (2018) argues that this behaviour is not rooted in an ignorance of the alternations, but in a preference for making paradigms uniform. Moreover, in his Experiment 1, Lysvik (2020) found indications that a paradigm uniformity bias can interact with a substantive bias. Lysvik's (2020) results show that participants learning a final devoicing language chose to alternate items to a larger extent than participants learning a final voicing language. That is, while participants in the DEVOICING condition chose the form *rusubu* when exposed to the trigger *rusup*, participants in the VOICING condition chose the form *rusubu* when exposed to the trigger *rusub*. Lysvik (2020) argues that this difference between conditions, which he refers to as *the alternation asymmetry*, could be the result of a substantive bias disfavouring the unnatural voicing process, and that participants' strategy for avoiding the unnatural alternation is influenced by their paradigm uniformity bias.

Based on these findings, then, it is important for the current experiment to account for potential effects of a paradigm uniformity bias both when designing the stimuli and procedure (Chapter 4) and when analysing and discussing the results (Chapter 5 and Chapter 6).

Turning now to substantive bias, the evidence for the role of this bias has been very variable in previous ALL experiments. While some experiments find robust effects of substantive bias (e.g. Finley, 2012; Shapp, 2012), others find inconsistent evidence, where there is some effect of substantive bias, but the evidence is not clear-cut. The alternation asymmetry in Lysvik's (2020) Experiment 1 is an example of such an experiment. There are also a number of ALL experiments which fail to find any effect of substantive bias (e.g. Do et al., 2016; Peperkamp & Dupoux, 2007). Moreton & Pater (2012b) present an overview of the evidence gathered from ALL experiments aiming to find effects of substantive bias and conclude that judging by the number of studies finding inconsistent or null results, the effect of substantive bias must be weak, if it exists at all.

The results from ALL experiments thus provide a messy picture of substantive bias effects. Greenwood (2016, p. 42) argues that a possible explanation for these discrepancies is that the circumstances under which naturalness comes to shape typology are not accurately recreated in ALL experiments. Contrary to most ALL studies, Greenwood (2016) explicitly takes a channel bias approach to investigating typological asymmetries, aiming at incorporating the effects of the phonetic pressures of speech in ALL experiments.

Greenwood contends that most ALL experiments make the mistake of looking for effects of naturalness with stimuli recorded in hyperarticulated speech, thus failing to properly recreate the circumstances under which the effects of channel bias would emerge. Greenwood (2016) therefore included two registers in her final devoicing experiment, a CAREFUL and a CASUAL register. Within each register, participants were further divided into a DEVOICING and a VOICING condition. According to the channel bias approach, any effects of naturalness should only be found in the CASUAL register, whereas in the CAREFUL register, where the phonetic pressures argued to underlie the voicing asymmetry were reduced or non-existent, participants should perform equally well in the VOICING condition and the DEVOICING condition.

Greenwood's (2016) result show that participants performed better in the CAREFUL register than in the CASUAL register in the VOICING condition, while in the DEVOICING condition, there was no difference between registers. Greenwood argues that these findings support the channel bias account. Yet, the results do not directly adhere to the predictions of channel bias when DEVOICING and VOICING are compared in the CASUAL register, as there was no difference between the two language conditions. That is, under the channel bias view, participants are expected to perform better in DEVOICING than in VOICING in this register, where the cues to final voicing are reduced, but this was not the case. Furthermore, in the CAREFUL register, participants in VOICING performed significantly better than those in DEVOICING. Similar results have been found in other experiments testing for differences between final voicing and final devoicing (e.g. Glewwe, 2019), and this finding contradicts both the substantive bias view that final voicing is dispreferred and the channel bias view that there should be no difference between conditions in the CAREFUL register. Greenwood's (2016) final devoicing experiment, one of very few which explicitly tested for channel bias effects, thus found no support for substantive bias and unclear results with regard to channel bias.

Based on the discussion above, it is possible to draw a few conclusions. First of all, there is strong evidence that complexity bias influences learning of phonological patterns in ALL experiments. The evidence also indicates that a paradigm uniformity bias is at work, but too few studies have investigated this bias to be able to determine its role in typological asymmetries in phonology. Moreover, it appears that a strong substantive bias account, in which unnatural processes are predicted to be unlearnable, is not supported. Greenwood (2016), along with a number of other experiments such as Do et al. (2016) and Glewwe (2019), found evidence that participants are able to learn unnatural patterns, sometimes to the same extent or even better than their natural counterparts.

As for the role of a weak substantive bias, however, it is so far not possible to determine to what extent it plays a role in the learning of phonological patterns. While several experiments have failed to find evidence for a substantive bias effect, the findings in among others Lysvik's (2020) Experiment 1 indicate that such an effect cannot be written off entirely. Additionally, as seen in Greenwood's (2016) final devoicing experiment, the predictions of the channel bias account do not always manifest themselves in a straightforward way either. The debate thus cannot be settled based on the evidence reviewed so far.

In the next section, further ALL evidence will be reviewed, this time pertaining to the voicing asymmetry specifically. The experiments in Myers & Padgett (2014), Greenwood (2016), Glewwe (2019), and Lysvik (2020), making use of this typological asymmetry to investigate the effects of substantive bias and channel bias, will be discussed. The findings in these experiments provide the specific research background for the current experiment.

3.2.2 ALL research on the voicing asymmetry

A number of ALL experiments have investigated the typological asymmetry between final devoicing and final voicing. There are several reasons for this asymmetry being particularly relevant. Firstly, the two processes are entirely parallel, and thus of equal complexity, excluding complexity bias as a confounding factor. Secondly, the phonetic motivations underlying the final devoicing process, and therefore disfavouring the final voicing process, are relatively well-understood. In contrast, asymmetries pertaining to stress patterns, for example, have not been investigated as often in ALL research because the phonetic motivation underlying such asymmetries are more difficult to define and control (Greenwood, 2016, p. 92). The voicing asymmetry, then, serves as a suitable example when investigating effects of substantive bias and channel bias in ALL experiments.

The evidence for substantive bias effects obtained in ALL experiments investigating the voicing asymmetry is varying. Myers & Padgett (2014) is an example of a study which has found indications for the role of substantive bias, although this evidence is not clear-cut. In their Experiment 1, Myers & Padgett found that participants trained on a final devoicing language learned the relevant pattern to a greater extent than participants trained on a final voicing language. Participants in the final devoicing condition also more readily extended this pattern from the utterance-final position to the word-final position than participants in the final voicing condition.

However, this experiment only measured behaviour on the coronal fricatives [s] and [z], and participants were not exposed to any alternations. It is therefore a very limited study. As will be seen in the discussion of Lysvik’s (2020) Experiment 1 below, participants’ behaviour with regard to the alternation between a voiced final obstruent and a voiceless intervocalic obstruent can inform on the unnaturalness of the final voicing process. It is therefore possible that important effects are missed in a study where items never alternate.

Greenwood’s (2016) final devoicing experiment is similar to Myers & Padgett’s (2014) Experiment 1 in a number of ways. The main findings of this experiment were discussed in Section 3.2.1, and so these will not be laid out again here, but certain details of Greenwood’s design are worth discussing. As was the case in Myers & Padgett’s (2014) Experiment 1, items never alternated, and critical items could only have the coronal fricatives [s] and [z] and the post-alveolar affricates [t͡ʃ] and [d͡ʒ] in the final position. The investigation of final devoicing and final voicing was therefore restricted to a small subset of obstruents.

Furthermore, as discussed by both Myers & Padgett (2014) and Greenwood (2016), the choice of these obstruents can be problematic when conducting an ALL experiment with participants whose L1 is English. Specifically, Greenwood (2016) speculates that the preference for VOICING discussed above and a generally lower performance in DEVOICING than expected could arise from participants' English L1 influence. That is, because English has phonetic devoicing of /z/ to [s], speakers might reconstruct /z/ instead of /s/ when they are exposed to [s].

Moving on to Glewwe's (2019) Experiment 5[2], this experiment tested the voicing asymmetry using both alternating and non-alternating items, argued above to be missing in Myers & Padgett's (2014) Experiment 1 and Greenwood's (2016) final devoicing experiment. In Glewwe's (2019) Experiment 5, participants were assigned to one of three conditions: Devoicing, Voicing, or Exchange. The first two conditions learned a final devoicing process and a final voicing process, respectively, while the last condition learned a pattern in which voiced obstruents were devoiced and voiceless obstruents were voiced word-finally.

Participants were tested on a forced-choice task, and it was predicted that if a substantive bias affects the learning of phonological patterns, participants in the Devoicing condition would learn their instructed pattern better than those in the Voicing condition. Participants in the Exchange condition were expected to perform worse than both of the other conditions, as this pattern is phonologically circular and appears to be unattested (Moreton, 1999).

However, Glewwe conceded that there was a possibility that the Voicing process would be learned better than the Devoicing process. In Glewwe's (2019) Experiments 3 and 4, the nature of the filler items was found to play an important role. When fillers were sonorant-final only, participants showed a preference for voiced obstruents, but when fillers were voiceless fricative-final only, participants preferred voiceless obstruents. In Experiment 5, fillers were all sonorant-final, and so the Voicing rule was structurally less complex than the Devoicing rule. That is, while the pattern in the Voicing condition could be captured with the constraint *[–voice]#, referring to a single feature, the pattern found in the Devoicing condition required the constraint *[–son, +voice]#, referring to two features. If there was no effect of substantive bias, or if this effect was weak, then, effects of complexity bias might surface.

The results from Experiment 5 show that participants learned the Voicing pattern significantly better than the Devoicing pattern, and, surprisingly, they also learned the

---

[2] Glewwe's (2019) Experiment 5 was first reported in Glewwe et al. (2018).

Exchange pattern better than the Devoicing pattern. This experiment thus found no effects of substantive bias, and the last prediction above, that a complexity bias would prefer Voicing over Devoicing, appears to hold true. Whether this means that substantive bias played no role, or that effects of substantive bias were washed away by the stronger effects of complexity bias is uncertain. But this result does indicate that if substantive bias exists, it is considerably weaker than complexity bias.

Finally, I turn to Lysvik's (2020) experiments. Similarly to Glewwe's (2019) Experiment 5, Lysvik tested participants' learning of a voicing language (the VOICING condition) and a devoicing language (the DEVOICING condition) using alternating and non-alternating items. In Lysvik's experiments, however, participants were native speakers of Norwegian. As participants' L1 has been found to be a possible confounding factor in the experiments discussed above, this difference can have a considerable effect on the results. Whereas English has a phonetic tendency to devoice final obstruents, this is not the case in Norwegian (Kristoffersen, 2000, p. 74). For this reason, Norwegian speakers make suitable participants for the investigation of the voicing asymmetry in ALL experiments.

Furthermore, Lysvik's (2020) Experiment 1 did not include filler items, while Experiment 2 did. Fillers are often included in order to distract from the pattern under investigation, making it less obvious to participants what they are being tested on. However, as Glewwe (2019) discovered in her Experiments 3, 4, and 5, these can also constitute a confounding factor. In Lysvik's Experiment 2, this was in fact the case, although not in the same way as in Glewwe (2019). Whereas the filler items were argued to sway participants' learning of the relevant pattern in a certain direction in Glewwe's (2019) experiments, they seemingly made the stimuli too difficult to learn overall in Lysvik's Experiment 2. The results from this experiment therefore cannot inform on the underlying biases in the voicing asymmetry, but they do inform on experimental design in ALL research. Specifically, when including filler items, it is essential both to make sure that they do not favour one of the patterns under investigation, as in Glewwe (2019), and that they do not make the artificial language too complex to learn, as in Lysvik (2020).

Because of these learnability issues in Experiment 2, the results from Lysvik's Experiment 1 provide the most insight into the nature of the voicing asymmetry. In the training phase of this experiment, items were matched with pictures to represent objects, and the voicing alternations were found in pairs of singulars and plurals. In the testing phase, Lysvik included two different tasks, a production task and a forced-choice task. In both tasks, the target form could be either singular or plural, allowing Lysvik to investigate both participants' performance on the word-final restriction, found in the singulars, and whether they preferred to alternate items or not, which was apparent in the plurals.

The overall results from Lysvik's Experiment 1 do not indicate that the final devoicing rule was learned better than the final voicing rule. There was no difference between the conditions in the singular target trials, indicating that the phonotactic restrictions were learned equally well. The results from the plural target trials, however, show a significant effect of condition. That is, participants in DEVOICING chose the alternating form significantly more than those in VOICING, constituting the alternation asymmetry mentioned in Section 3.2.1. Lysvik explains this finding with a paradigm uniformity bias resulting from substantive bias effects. Because participants were substantively biased against the final voicing process, Lysvik contends, they found a way to avoid it. Instead of the predicted avoidance strategy, to devoice voiced final stops, participants opted for a strategy in which the unnatural process was avoided by making paradigms uniform.

However, Lysvik also recognises other possible explanations for this finding. It is possible that the avoidance of the final voicing process stems from a channel bias against voiceless intervocalic stops, which are also phonetically disfavoured (Westbury & Keating, 1986, p. 152). What is more, in Experiment 1, the stimuli were created synthetically, and Lysvik concedes that it is possible that the synthetic stimuli influenced participants to interpret intervocalic voiceless stops as voiced. The aim of his Experiment 2 was therefore to investigate whether the same effect was found with naturalistic stimuli. As mentioned above, however, the results from Experiment 2 were not able to reveal any effects of either substantive bias or channel bias.

Finally, Lysvik (2020) presents a suggestion for a future ALL experiment which can shed further light on the voicing asymmetry and phonological biases. The suggested experiment should make use of the iterated learning paradigm, in which language change is simulated through diffusion chains. Generation 1 in a chain would then learn either a final voicing language or a final devoicing language as in Experiments 1 and 2, and the output from this generation would subsequently be used as input for Generation 2 in the chain, and so on.

In conducting such an experiment, Lysvik argues, effects of substantive bias and channel bias can emerge. The predictions from channel bias hold that final voicing will become less stable over time as misperceptions and misproductions conspire against it. These same processes are not expected to affect final devoicing to the same degree. Substantive bias also predicts that final voicing, and not final devoicing, will become less stable throughout the diffusion chains, but that this difference would result from a learning bias disfavouring final voicing. In this experiment, both singular target trials and plural target trials should be included in order to investigate not only final stops but also the alternation asymmetry found in Lysvik's Experiment 1. The current experiment builds on this suggested iterated learning experiment, and an outline of the experiment will be presented in Section 3.3 below.

3.2.3 Iterated learning experiments

ALL experiments making use of the iterated learning paradigm are not as common as those making use of the other three paradigms discussed in Section 3.1, but a number of experiments have been conducted investigating among others compositionality (Griffiths et al., 2008; Kirby et al., 2008), morphosyntactic structures (Smith & Wonnacott, 2010), and word-meaning mappings (Reali & Griffiths, 2009). However, to my knowledge, all studies making use of iterated learning in phonology have used mathematical or machine learning simulations rather than ALL experiments (De Boer, 2000; Wedel, 2012). Yet, iterated learning experiments looking at structures in other linguistic domains have found interesting results, and there is no reason to believe that this will not also be the case in an iterated learning experiment examining phonological processes such as final devoicing and final voicing.

The most important finding in previous iterated learning experiments is that the behaviour of populations can differ from the behaviour of individuals (Smith & Wonnacott, 2010, p. 445). That is, while underlying biases in individual learners may not become apparent in just one generation, such biases can be augmented when language is transmitted between generations. This effect has been found for example in experiments investigating the regularisation of inconsistent patterns, such as Smith & Wonnacott (2010). In this experiment, participants formed ten diffusion chains, each with five generations, and were exposed to an artificial language with two plural markers and four nouns. Both plural markers were found with all nouns in the input to the first generation in all chains, and so the marking of plurals was inconsistent. These participants were then tested on short sentences containing a verb, a noun, and a plural marker, and the choice of plural marker with each noun was used in the training language for the next generation.

Smith & Wonnacott (2010) found that although individual participants did not greatly alter the language compared to the input, the predictability of plural marking cumulatively increased throughout the diffusion chains. In nine out of ten chains, the language in the final generation either had only one plural marker or consistently paired plural markers and nouns. This bias for regularisation could not be observed in the first generation in the diffusion chains, but became more apparent for each generation.

Returning to the voicing asymmetry, the experiences from previous iterated learning experiments indicate that the biases of individual learners such as the possible dispreference against final voicing may only become apparent on the population level. If this is the case, it could explain why the experiments conducted by Greenwood (2016), Glewwe (2019), and Lysvik (2020) largely did not find naturalness effects. If the bias underlying the asymmetry between final devoicing and final voicing is weak, as Glewwe (2019, p. 159) suggests, it is possible that its effects are not visible in individual-based experiments, but that they can be amplified and become apparent in an iterated learning experiment. That is not to say that such biases cannot be found in individual-based experiments, but it is worth investigating population-level effects as well. The findings in previous iterated learning experiments thus motivate the use of this paradigm in the current experiment.

3.3 Outline of the current experiment

In the preceding sections, I presented the ALL method and discussed previous experiments showing that the evidence for substantive bias and channel bias has been very variable. I argued that, based on the experiences made in previous research on the voicing asymmetry as well as previous research making use of the iterated learning paradigm, conducting an iterated learning experiment examining the voicing asymmetry can provide new insights into which type of bias can explain this and other typological asymmetries in phonology.

The current experiment thus builds on the research discussed in Section 3.2, in particular the experiments conducted by Lysvik (2020) and his suggested follow-up experiment. It is an iterated learning experiment making use of the ease-of-learning paradigm where participants were divided into ten diffusion chains consisting of five generations. Five of these chains were in the DEVOICING condition and five were in the VOICING condition. Generation 1 in each chain was trained on completely consistent final devoicing or final voicing languages, and the output obtained from this generation was used as the input in the training phase of Generation 2, and so on. This design is illustrated in Figure 3.1.

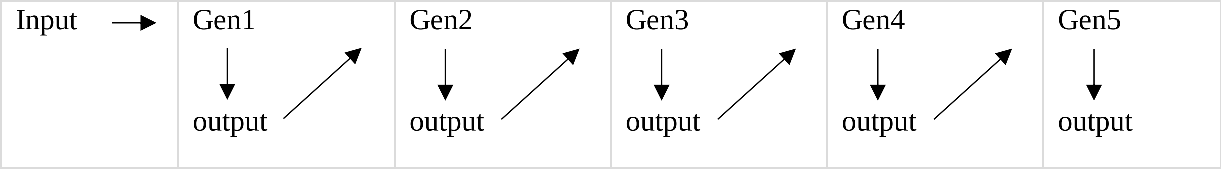


Figure 3.1 Example of a diffusion chain. There were in total ten such chains in the experiment, five in the VOICING condition and five in the DEVOICING condition.

This design was used with the aim of further investigating the alternation asymmetry found in Lysvik's (2020) Experiment 1. If the difference between the DEVOICING condition and the VOICING condition did in fact arise from a substantive bias, it is expected that participants in DEVOICING will alternate items to a larger degree than those in VOICING, and that this effect will be amplified throughout the diffusion chains in the iterated learning experiment. That is, in each generation in the VOICING condition, participants' dispreference against the unnatural process will cause them to alternate less than in the previous generation. Because participants are not biased against the final devoicing process, this should not be the case in the DEVOICING condition. Moreover, although previous experiments have not found a dispreference against voiced final stops, it is also possible that a difference in learnability and stability between conditions in the current experiment will target these segments. In fact, such a difference would be the clearest indication that participants are substantively biased against final voicing.

If the alternation asymmetry was not due to substantive bias, but to channel bias, a similar effect would be expected, but for a different reason. Specifically, due to misperceptions and misproductions in the transmission of the artificial language between generations, participants in VOICING are expected to gradually get rid of the unnatural process, either by devoicing voiced final stops or by voicing voiceless intervocalic stops, or both.

The two accounts thus make similar predictions, and consequently, it is necessary to isolate the effects of one type of bias in the iterated learning experiment. That is, if effects of both substantive bias and channel bias could influence participants' performance, it would not be possible to determine which of these is responsible for any observed behaviour. Because the aim of this experiment is to investigate the alternation asymmetry in Lysvik's (2020) Experiment 1, and because this effect is hypothesised to arise from a substantive bias, only potential substantive bias effects are investigated. In the current experiment, then, effects of channel bias are excluded to the extent that it is possible. The implementation of this separation will be discussed further below.

Turning now to the experimental design, this experiment used the stimuli from Lysvik's Experiment 2, and it thus included both critical items containing stem-final stops and filler items containing stem-final nasals and voiceless fricatives. Based on Glewwe's (2019) experience that the inherent voicing of the filler items can affect performance on critical items, the current experiment included both inherently voiced and inherently voiceless stem-final consonants in the fillers. Because Norwegian does not have either voiceless nasals or voiced fricatives, there are no confounds for these segments in the participants' native language (Lysvik, 2020, p. 66). Additionally, to mitigate a lack of learning of the critical items due to influence from the filler items, only 7 filler items were included in the input to each participant.

As for the critical items, these could be both alternating and non-alternating in the input to each participant. Both the training phase and the testing phase consisted of 46 items, where 7 items were fillers and there was a 2/3 to 1/3 distribution between alternating and non-alternating critical items. If a majority of items were non-alternating, it is possible that participants would regularise paradigms instead of learning the voicing alternation rules. Therefore, a majority of critical items were alternating in order to make sure that participants were exposed to enough evidence of the alternation.

Another choice regarding the experimental design is whether or not to include an interim test following the training phase to make sure participants learn the stimuli to a set threshold. Based on the lack of learning of the stimuli in Lysvik's (2020) Experiment 2, including such a threshold can seem advantageous. However, the implementation of this test proves difficult in an iterated learning experiment, as the stimuli changes from generation to generation. In Generation 1, where participants are trained on carefully designed stimuli, it would be possible to include such a test. After Generation 1, however, the researcher no longer controls the nature of the stimuli, and the threshold used for Generation 1 no longer applies. For this reason, I did not test whether participants had learned the stimuli to a set threshold in the current experiment.

Furthermore, the current experiment only included a forced-choice task in the training phase, rather than both a forced-choice task and a production task. This choice of task was made mainly due to practical concerns. I considered it easier to recruit participants to an online experiment, in addition to the actual completion being less time-consuming, and so I opted for conducting the experiment online rather than in the laboratory. Therefore, it would have been difficult to obtain sufficiently good audio recordings from participants, and I consequently decided against using a production task. In doing so, I excluded production effects, but as mentioned above, the aim of the experiment was to study the learning of the final devoicing and final voicing processes. The investigation of channel bias effects in production and perception are left for future research.

Turning to the actual transmission of the languages in an iterated learning experiment, this can be done in two different ways. The first alternative is to use a statistical measure such as the mean or the median of correct and incorrect or alternating and non-alternating items calculated from all the participants in Generation *n* and used as input to all the participants in Generation *n+1*. That is, each language condition would make up one chain, and there would be five participants in each generation. The second alternative is to have five diffusion chains in each language condition, where each generation is made up of one participant. The chosen form for each item in the output of Generation *n* would then be given as input to Generation *n+1*. That is, if Generation 1 in a given chain heard *rusub – rusupu* in the training phase, but chose *rusub – rusubu* in the testing phase, Generation 2 of this chain would be trained on the specific item *rusub – rusubu*.

In this experiment, I have chosen the latter approach. There are arguments in favour of the former approach, most importantly that learners of natural languages are not only exposed to the input from one other speaker. Yet, if I were to go with the former approach, there would be a risk of ending up with stimuli that does not accurately represent the output of any participant. For example, if all participants either have a very high or a very low proportion of alternating items, the mean would fail to capture the nature of the output from these participants. Using the median is not ideal either, as it would not represent participants whose performance is at the extremities. It could be the case that it is exactly these participants who drive language change, and so their data could provide essential insights into the development of the two voicing alternation processes over time.

For these reasons, each generation in this experiment consisted of one participant, and the exact same items used in the testing phase of Generation *n* were used in the training phase of Generation *n+1*. This approach is also the most common in iterated learning experiments, used in for example Reali & Griffiths (2009) and Smith & Wonnacott (2010). A point of caution, however, is that this approach makes each diffusion chain very sensitive to individual differences among participants. A likely scenario is that an entire chain is influenced by the performance of the participant in Generation 1 if this participant does not learn the relevant pattern. If this were to be the case, it would not be possible to know whether the following four participants, if given the input consistent with the relevant pattern, would have learned the voicing alternation under investigation. This issue is discussed further with regard to the results in Chapter 5 and Chapter 6.

A final important detail in the design of this experiment is the use of orthography in the testing phase. In the training phase, participants were only exposed to the audio recordings of items along with the pictures of the object they represented, but in the testing phase, the audio recordings and pictures were accompanied by orthographic representations of the two alternatives in the forced-choice task. The decision to include orthography was made to avoid participants having to keep the order of the two alternatives in memory, and the consequence of this decision is that perceptual channel biases were mostly excluded. When participants saw the two alternatives orthographically, any perceptual difficulties in distinguishing between voiced and voiceless obstruents was removed. There is, however, a possibility that perception played a role in the training phase, where orthography was not included. Yet because the native language of the participants, Norwegian, has a contrast between voiced and voiceless stops in the word-final position and no tendency to devoice the voiced series, there is reason to believe that participants were not influenced by such a channel bias effect to a great extent.

As was mentioned above, then, only a learning bias was tested for in the current experiment. Any differences found between VOICING and DEVOICING can thus not be attributed to transmission effects of production or perception but must be due to the two languages being learned to different degrees. By excluding a production task and including orthography, and thus separating out effects of channel bias from possible effects of substantive bias, it is possible to say with greater certainty that substantive bias must be responsible for any observed effects.

To sum up, the current experiment uses the iterated learning paradigm to investigate the role of substantive bias in shaping the voicing asymmetry. Specifically, by isolating the effects of a potential substantive bias, it tests the hypotheses that final voicing will be learned to a lesser degree than final devoicing, and that this effect will be amplified throughout the generations of the diffusion chains. Effects of a learning bias are predicted to manifest themselves either in an avoidance of voiced final stops relative to voiceless final stops, or in an avoidance of alternating items in the final voicing condition compared to the final devoicing condition, or both.

3.4 Summary

In this chapter, the ALL method was introduced. The most common design features were presented, and the advantages and disadvantages with this method were discussed. I argued that despite the shortcomings of the method, it has the advantage that it allows the researcher to test specific hypotheses regarding typological asymmetries while controlling for confounding factors. Furthermore, previous ALL research was presented, and it was pointed out that the evidence for substantive bias and channel bias has been varying, both in experiments pertaining to the voicing asymmetry specifically and in ALL experiments in general. I then discussed previous research showing that iterated learning experiments can provide novel insights into typological asymmetries, as they can reveal population-level biases. I argued that the previous research motivates conducting the current experiment, investigating the effects of substantive bias in the learning of final voicing and final devoicing in an iterated learning experiment. Finally, I presented an outline of the overarching choices made in the design of this experiment.

4 Methods

In this chapter, the details of the experimental design used to collect the data for the current experiment are presented. Section 4.1 describes how the iterated learning paradigm is implemented, and Section 4.2 presents the participants. In Section 4.3, the stimuli are described, while Section 4.4 presents the experimental procedure. Section 4.5 is a presentation of the analysis used to examine the data, and finally, Section 4.6 is a summary of the chapter.

4.1 The iterated learning paradigm

As discussed above, the experiment was conducted using the iterated learning paradigm. There were two language conditions, VOICING and DEVOICING, and in each of these conditions, there were five diffusion chains consisting of five generations of participants. Each participant made up one generation in the diffusion chain. The ten participants in Generation 1 underwent a training phase where they were exposed to fully consistent languages with either final voicing or final devoicing, and these participants were then tested on how well they had learned the languages using a forced-choice task (see Section 4.4). The results of this task, that is, the output from the participants in Generation 1, was subsequently used as input in the training phase for Generation 2. This same process was repeated for each generation in the chain, as illustrated in Figure 3.1 above, repeated in Figure 4.1.

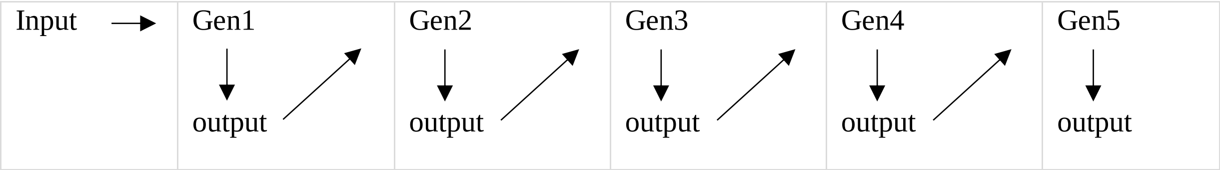


Figure 4.1 Example of a diffusion chain. There were in total ten such chains in the experiment, five in the VOICING condition and five in the DEVOICING condition.

4.2 Participants

There were 50 participants in the experiment. Participants were randomly assigned to one of the two language conditions and to a generation in one of the diffusion chains. There were 26 women and 24 men, and their age ranged from 19 to 58 years with a mean age of 27.

Participants were recruited from the University of Oslo and among friends and acquaintances through social media. They were informed that only native speakers of Norwegian could participate and were also asked to report any other languages they know well. If any participant had reported that they speak a language with final devoicing, such as German, Dutch, or Russian, they would have been excluded from participation, but this was not the case. Participants were informed that participation was voluntary and that they could withdraw at any point.

4.3 Stimuli

The stimuli used in the experiment were based on the stimuli constructed by Lysvik (2020) for his Experiment 2. The stimuli were produced by a phonetically trained native Norwegian speaker and were recorded with a Zoom H4n handheld recorder. All items were recorded in the carrier phrase *jeg sier __ først* 'I say __ first', and the order of the items was randomised. The items had two forms: a monosyllabic or bisyllabic stem and the stem combined with the suffix /u/. The stem represented the singular form of a noun and the suffix represented the plural form, but this was not known to the speaker.

Monosyllabic stems had the structure CVC, while bisyllabic stems had the structure CVCVC, and the distribution of monosyllabic and bisyllabic stems was 31% and 69%, respectively. The vowels could be either of /i e a o u/, while non-final consonants could be any of the consonants /p t k b d g m n l r f s/. The final consonant in both types of stems represented the *target consonant*. These could be either a voiceless stop /p t k/, a voiced stop /b d g/, a nasal /m n/ or a voiceless fricative /f s/ depending on the language condition and the type of the item. Items with a nasal or a voiceless fricative as the target consonant were *filler items*. In both the DEVOICING and the VOICING condition, these could have either of /m n s f/ as their target consonant, and these items did not undergo any alternations. These segments thus always surfaced as [m n s f].

Items which had a stop as the target consonant were *critical items*, that is, the items targeted by the voicing alternation processes. In both conditions, the target consonant in critical items could be either of the voiced or voiceless stops [p t k b d g] in the plural form, where the underlying form of the consonant was expressed. In the singular form, the voicing neutralisation took place, and only the voiceless stops [p t k] were allowed as the target consonant in DEVOICING, while only the voiced stops [b d g] were allowed as the target consonant in VOICING. Critical items could thus be either alternating or non-alternating. Examples of all types of items can be found in Table 4.1.

Table 4.1 Examples of the possible item types in the original input. Filler items were common to the two conditions.

	Item	Singular	Plural	Type
VOICING	banana	seseg	seseku	alternating
	flower	nad	nadu	non-alternating
DEVOICING	banana	sesek	sesegu	alternating
	flower	nat	natu	non-alternating
Filler	hammer	tipis	tipisu	n/a
	airplane	gum	gumu	n/a

There were 65 items in total, of which 39 were alternating critical items, 19 were non-alternating critical items, and 7 were fillers. As discussed in Chapter 2 and Chapter 3, the reason both alternating and non-alternating critical items were included was so that participants would learn the relevant alternation. On the one hand, if there had not been enough evidence of the alternation, there would have been a risk of participants opting to consistently not alternate, keeping paradigms uniform. If all items had been alternating, on the other hand, it is possible that participants would have learned a rule of intervocalic voicing or intervocalic devoicing. Therefore, it was necessary to include non-alternating critical items as well.

As for the filler items, making up 15% of the items in both the training phase and the testing phase, these were included in order to make the final voicing and final devoicing rules less obvious to the participants. The stimuli from Lysvik’s (2020) Experiment 2 only contained 7 suitable filler items, and because it would have been time-consuming to record more items, these were the only fillers included in the experiment. Yet, the reduction in the number of filler items in the input falls in line with Lysvik’s (2020) outline for a follow-up experiment, as he suspected that including too many fillers in his Experiment 2 inhibited learning of the critical items. Moreover, Glewwe (2019, p. 4) mentions that the proportion of filler items in previous ALL research ranges from two thirds to 5%, and so with a proportion of 15%, this experiment does not deviate from other experiments.

4.4 Procedure

The experiment was designed in PennController for IBEX (Zehr & Schwarz, 2018) and completed online. Because of the iterated learning design of the experiment, each participant received a separate link, leading to a separate version of the experiment depending on their place in a specific diffusion chain. Participants were not aware of this design of the experiment.

Participants were first shown an introductory text explaining the procedure of the experiment. They were told that they would be learning the unknown language *nugobisk* (in English, this name roughly translates to 'Nugobian') and that after the training phase, there would be a test to see how well they had learned the language. They were then asked to fill in a form stating their age, gender, and other known languages. In addition, they were informed that participation required wearing headphones. Following this introduction was a sound check, where participants could adjust the volume by listening to the word *mipidu*. This word was not included in the training or the testing phase of either condition, but it follows the same structure as the other stimuli. Importantly, as it is a plural form, it does not show evidence of either final voicing or final devoicing, and it is thus allowed in both language conditions.

The experiment consisted of two phases, a training phase and a testing phase, and each phase consisted of 46 items, the order of which was randomised. In the training phase, each trial had the following structure, here exemplified for the item *necklace*:

1. Picture of a necklace is shown.
2. Audio recording of the singular form of 'necklace' is played.
3. Picture is removed.
4. Picture of two necklaces is shown.
5. Audio recording of the plural form of 'necklace' is played.
6. Plural picture is removed.

Participants could not repeat items and thus only heard each form of each item one time.

After the training phase, there was a break in the experiment where participants were informed of their task in the testing phase. This phase consisted of a forced-choice task where participants were exposed to a *trigger*, either the singular or the plural form of an item, before they were shown a picture of the opposite number of the same item, the *target*. Two audio recordings were then played, one with a voiced stop as the target consonant and one with a voiceless stop as the target consonant. The orthographic forms of the words were also displayed in order to avoid participants having to keep the order of the two forms in their short-term memory (see Section 3.3 for discussion). Participants were asked to choose the correct form by pressing either F (first audio played and left orthographic form) or J (second audio played and right orthographic form). Figure 4.2 illustrates a trial from the testing phase.[3]

a) Plural trigger *fegu* ‘necklace-pl’.

b) Singular target.

Figure 4.2 Illustration of a singular target trial in the testing phase. The participant is asked to choose the correct form of ‘necklace’, either feg or fek, by pressing F or J, respectively.

---

[3] All pictures were license-free from Google Images.

In this phase, each trial had the following structure, again exemplified with the item *necklace* (here explained with a plural trigger and a singular target, but the structure is the same for trials with a singular trigger and a plural target):

1. Picture of two necklaces is shown.
2. Audio recording of the plural form of ‘necklace’ is played.
3. Plural picture is removed.
4. Picture of a single necklace is shown.
5. Singular VOICING audio recording of ‘necklace’ is played.
6. Singular DEVOICING audio recording of ‘necklace’ is played.
7. Both orthographic forms are displayed.
8. Key press, either F or J.

The order of the two target forms, that is, whether the VOICING or the DEVOICING form was played first and thus ended up on the left side of the screen, was randomised.

The selection of items was different in the first generation than in the following generations, so let us begin with the first generation. The ten participants in Generation 1 were trained on 26 alternating and 13 non-alternating critical items, in addition to the 7 filler items. The 39 critical items were, except for the number of alternating and non-alternating items, randomly selected from the in total 58 critical items.

In the testing phase, participants were tested on the 7 fillers and 39 critical items. Of these 39 critical items, 20 were old, that is, they had been introduced to the participant in the training phase, and 19 were new. The new items were the remaining 19 of the 58 critical items which were not selected in the training phase, while the 20 old items were randomly selected from the 39 critical items the participant had already heard in the training phase. This division between new and old items was included to determine whether participants extended the rule to items they had not heard before.

Furthermore, the number of alternating and non-alternating critical items was not chosen deliberately in the testing phase in Generation 1, as it would not be possible to control this in the following generations when participants will have altered the types of items and the proportion of alternating and non-alternating items in each language. That is, if the item *feg* – *feku* ‘necklace’ was introduced to a participant in Generation 1 in the VOICING condition, this item was of the alternating type. But if this participant in the testing phase was exposed to the singular trigger *feg*, and in instead of choosing the plural form *feku* chose the form *fegu*, this item would be of the non-alternating type in the input to Generation 2. From Generation 2 on, then, the proportion of alternating and non-alternating items in each chain varied, and so this proportion varied in the testing phase of Generation 1 as well.

As for the generations following Generation 1, their input was determined by the results of the testing phase in the previous generation. In order for the output of Generation 1 in a given chain to be used as the input to Generation 2 of this chain, the 46 items in the training phase of Generation 2 will have to be the same as the 46 items in the testing phase of Generation 1. As discussed above, the number of alternating or non-alternating items in the training phase was thus no longer controlled from Generation 2 on, as this was determined by the choices of Generation 1 in the testing phase. In the testing phase of Generations 2 to 5, the selection of items was the same as in Generation 1.

Moreover, in the testing phase, the trigger was plural in 74% of items, that is, in 27 critical items and the 7 fillers. Thus, 74% of the targets were singular, which is the most telling measure of whether participants have learned the relevant voicing or devoicing rules or not. In the remaining 12 items (26%), the trigger was singular. It is these trials which can reveal whether participants preferred to alternate or not, providing insight into the alternation asymmetry. In the singular target trials, there was a “right” and a “wrong” answer depending on the language condition, in that only voiceless stops adhere to the phonotactic constraint in the DEVOICING condition and only voiced stops adhere to the phonotactic constraint in the VOICING condition. In the plural target trials, on the other hand, both voiced and voiceless stops were allowed in both conditions, and so there was no right or wrong answer.

As for the fillers, these were tested in the same way as the critical items in the testing phase, but when participants were to choose between two forms, the correct form with a nasal or a voiceless fricative was given as one alternative, and another form with a confound as the target consonant was given as the other alternative. These confounds were stops of the same place of articulation as the nasal or voiceless fricative, and in the VOICING condition, these were voiced, while in the DEVOICING condition, they were voiceless. However, even if participants chose the confound, the filler items were never altered between generations to ensure the same presence of fillers for all participants. Table 4.2 illustrates the plural, singular target and singular confound form in fillers with the item *hammer*.

Table 4.2 The plural, singular target, and singular confound form of the filler item *hammer*.

	Pl.	**Sg. target**	**Sg. confound**
VOICING	tipisu	tipis	tipid
DEVOICING	tipisu	tipis	tipit

4.5 Analysis

The results of the experiment outlined above were analysed descriptively. If inferential statistics were to be conducted, the most suitable choice of a statistical model would be a mixed effects model using the *lme4* package in R (Bates & Sarkar, 2016). However, such a model would run into problems both with convergence and with detecting possible significant effects.

When running mixed effects models in the *lme4* package, the model is estimated through an optimisation algorithm which estimates the maximum likelihood estimate. Convergence issues arise when the model fails to find a stable estimate. Often, such issues are due to the attempted model being too complex for the amount of data (Winter, 2020, pp. 265-267). In the current study, a mixed effects model would have to be relatively complex due to the complex design of the experiment and the fact that there are a lot of dependencies in the data. Furthermore, although the experiment has 50 participants, the amount of data for each grouping variable is sparse because of there being only one participant in each generation in each diffusion chain. It is therefore very likely that a mixed effects model based on these data would run into convergence issues.

The sparsity of the data is also the reason it would be unlikely that any model would detect significant results, even if there was in fact a difference between conditions. The design of the experiment has the consequence that the statistical power is low, and the chances of getting a false negative (Type II error) are therefore high. In other words, if the estimated model did in fact converge, it would be difficult to interpret the results with much confidence.

Because of these issues, I have decided against conducting inferential statistics in the following analysis. The analysis will instead be a descriptive exploration of the data, which can provide useful insights into participants' behaviour in the two language conditions. Furthermore, it can be an important building block in further developing the iterated learning experimental design.

With the choice of analysis in place, then, it is necessary to discuss the measure used to evaluate participants' performance in the analysis. Initially, it appears that the relevant measure is accuracy, that is, the percentage of items in the output of each participant which correspond to the form found in the input. In an iterated learning experiment, however, this approach is not the most informative to the question of whether there is a difference in learnability and stability between the final voicing process and the final devoicing process. Looking at the results from the first generation only, this measure gives an indication as to how well the two alternation patterns were learned. Yet, from the second generation on, the accuracy measure depends on the output of the previous generation in each diffusion chain.

If Generation 1 in a chain in the VOICING condition was exposed to the plural trigger *putu* 'anvil-pl', for example, and chose the singular form *put* 'anvil' instead of the expected form *pud,* this item was presented to Generation 2 in this chain as *put – putu* in the training phase. This item, which had been classified as alternating in the input to Generation 1, would thus be classified as non-alternating in the input to Generation 2. But not only would it be non-alternating, it would be non-alternating of a different type than the non-alternating items in the original input, with a voiceless stop as the target consonant in both the singular and the plural form. Yet, if accuracy was used as the relevant measure for singulars, the correct response for this item in Generation 2 would be *put*, retaining its type from the output of Generation 1. This measure therefore does not inform about whether participants preserve the voicing alternation throughout generations, but it does tell us how well each participant learned the pattern from the previous generation. This measure is in itself also relevant and will be discussed in Section 5.7.

A measure that can in fact inform about the stability of the two voicing neutralisation processes throughout the diffusion chains is the percentage of items adhering to the phonotactic restriction in the word-final position in the relevant condition. If there is a high percentage of these items, which will be referred to as *conforming* items, in the output of a given participant, it indicates that the participant learned the voicing neutralisation rule. Yet, if participants learned the relevant patterns, it is also expected that a substantial proportion of the conforming items were alternating. That is, a participant with a high percentage conforming items, but with a very low proportion of these being alternating, cannot be said to have fully learned the relevant voicing alternation. In this case, it is more likely that the participant learned a pattern in which paradigms were uniform and only one series of stops was allowed, either voiced or voiceless. Therefore, it is necessary to measure both the total percentage conforming items and the percentage *alternating* conforming items to determine whether participants learned final voicing in the VOICING condition and final devoicing in the DEVOICING condition.

As stated above in Section 4.4, 67% of critical items were alternating in the input to Generation 1 in each diffusion chain. Yet, as was the case with non-alternating items, it is also possible that items could become alternating of the opposite type than would be expected for the relevant language condition. Take the item *terik – teriku* 'chocolate bar' in the DEVOICING condition, for example. If in the testing phase for Generation 1 in a given chain the participant was exposed to the plural trigger *teriku*, but instead of the expected form *terik* chose the form *terig*, this item would be classified as alternating in the input to the next generation. However, this alternation does not conform to the alternating items in the original input in the DEVOICING condition, but rather corresponds to the alternation expected to be found in the VOICING condition.

It is thus necessary to distinguish between four types of critical items in the analysis of the results: alternating and non-alternating items which conform to the rule in the relevant condition (conforming items), and alternating and non-alternating items which do not conform to this rule (non-conforming items). In what follows, these will be referred to as alt (conforming), non-alt (conforming), alt (non-conforming), and non-alt (non-conforming). Table 4.3 gives an overview of the four types with an example of each type. In examining the percentage conforming and non-conforming items and the proportion of these which were alternating, it is possible to get an overview of how well the voicing and devoicing rules were learned, whether participants prefer alternating or non-alternating items, and in which ways they change the two languages.

Table 4.3 The four possible types of critical items in the results.

a) VOICING

Item type	**Example 'chocolate bar'**
Alt (conforming)	terig – teriku
Non-alt (conforming)	terig – terigu
Alt (non-conforming)	terik – terigu
Non-alt (non-conforming)	terik – teriku

b) DEVOICING

Item type	**Example ‘chocolate bar’**
Alt (conforming)	terik – terigu
Non-alt (conforming)	terik – teriku
Alt (non-conforming)	terig – teriku
Non-alt (non-conforming)	terig – terigu

As for the filler items, these were not changed according to the responses of the previous generation in each chain, and it is therefore not relevant to study how these developed. For these items, the relevant measure is accuracy.

4.6 Summary

This chapter presented the methods for data collection in the current experiment. The organisation of participants into conditions and diffusion chains was presented, as was the participant pool. The nature of the stimuli, both critical items and filler items, was presented, and the experimental procedure, including the training phase and the testing phase, was described. Finally, I argued against using inferential statistics in the analysis of the data, and I presented the chosen analysis, a descriptive approach in which the percentages conforming and non-conforming as well as alternating and non-alternating items are the relevant measures.

5 Results

In this chapter, the results of the experiment outlined in Chapter 4 will be presented. In Section 5.1, the main measures of interest, the percentages conforming and alternating conforming items at each stage in the diffusion chains, are presented both by condition and separately. In addition, the distributions of all four possible item types in each chain are presented and the two conditions are compared. Subsequently, a range of variables which can have affected the distribution of item types are presented, namely alternation (Section 5.2), novelty (Section 5.3), length (Section 5.4), final stop (Section 5.5), number (Section 5.6), and accuracy (Section 5.7). The results of the filler item trials are laid out in Section 5.8, and finally, Section 5.9 summarises the chapter.

5.1 Distribution of item types

5.1.1 Conforming items

When analysing the results, it is useful to study the ten diffusion chains by condition as well as separately. Let us begin by looking at the percentage conforming items in each chain in the two conditions. Recall that this measure provides insight into the degree to which participants adhered to the restriction of having only voiced stops (VOICING) or only voiceless stops (DEVOICING) in the word-final position, but it does not inform about how many of these items were alternating. The ten chains are presented by condition in Figure 5.1.

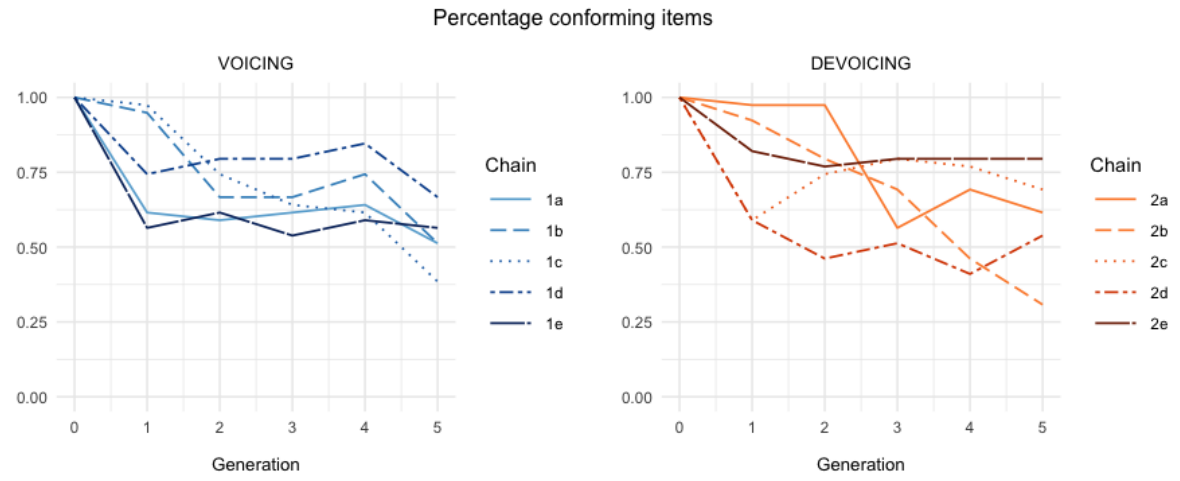


Figure 5.1 Percentage conforming items in each generation in VOICING and DEVOICING.

Figure 5.1 illustrates that in the input to Generation 1, represented as Generation 0 in the figure, all critical items were conforming, that is, either alt (conforming) or non-alt (conforming). It becomes clear that the ten diffusion chains diverged already from Generation 1, with the percentage conforming items spanning from 56% to 97% in the VOICING condition and from 59% to 97% in the DEVOICING condition. It thus appears that there were great individual differences between participants, but that the span in performances was quite similar in the two conditions in Generation 1.

Similarly, the following generations did not show dramatic differences between VOICING and DEVOICING either. In both conditions, certain diffusion chains retained a high percentage conforming items, between 70% and 80% in most or all generations (Chain 1d in VOICING and Chains 2c and 2e in DEVOICING). Most chains in both conditions, however, had a percentage conforming items around 60% in most generations and converged on 50% in the final generation. Two chains, 1c and 2b, had a percentage below 50% in Generation 5, with 38% and 31% respectively. Percentages around or below 50% indicate that participants do not adhere to the phonotactic restriction in their language, and so Figure 5.1 demonstrates that in a majority of diffusion chains, the final voicing and final devoicing processes were lost by Generation 5.

While Figure 5.1 illustrates to which degree participants adhered to the phonotactic restrictions in the word-final position in VOICING and DEVOICING, it does not inform us of how many of these items were alternating. It is possible that a majority of the conforming items were in fact non-alternating, which would indicate that participants avoided the voicing alternations. To investigate whether this was the case, the percentage alternating conforming items was calculated for each generation in each diffusion chain, as illustrated in Figure 5.2.

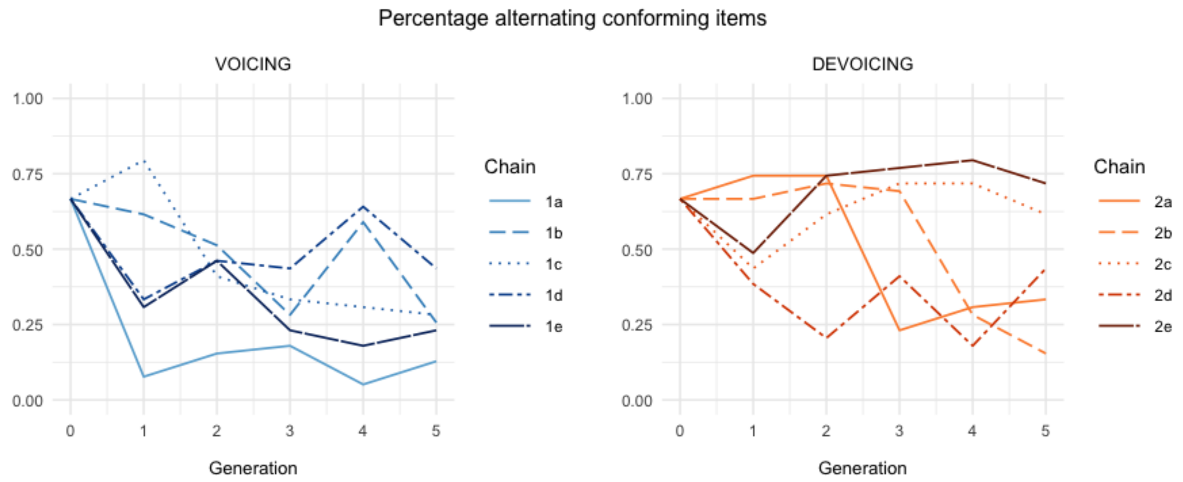


Figure 5.2 Percentage alternating conforming items in each generation in VOICING and DEVOICING.

The percentage alternating conforming items in the original input, 67%, was plotted as Generation 0. In Generation 1, the two conditions do not appear to be very different, with chains both above and below input levels in both VOICING and DEVOICING. The exception is Chain 1a in VOICING, which had a considerably lower percentage than the other chains, 8%. From Generation 2 on, however, the percentages alternating conforming items seem to drop to a larger extent in VOICING than in DEVOICING. Two chains in DEVOICING, 2a and 2b, had a percentage at or slightly above the original input level until Generations 2 and 3, respectively, while Chains 2c and 2e showed an initial drop in Generation 1 and a subsequent increase to approximately 70-75%, which remained stable throughout the generations. In VOICING, on the other hand, all chains dropped to around 50% or lower in Generation 2, and apart from an increase in Generation 4 in Chains 1b and 1d, the percentages remained below 50% until Generation 5.

Because of the great variability between participants, caution is necessary when interpreting these results and when making comparisons between chains and conditions. Taking this variability into account, it is possible to make a tentative comparison of VOICING and DEVOICING. In terms of the overall percentages conforming items, the two conditions show similar trends, indicating that there were no considerable differences between them. As for alternating conforming items, there appears to be a difference between VOICING and DEVOICING. The two chains showing the greatest amount of stability throughout the generations were 2c and 2e, both in the DEVOICING condition. In addition, Chains 2a and 2b had a high percentage alt (conforming) items until Generation 2 and Generation 3, respectively, before it dropped substantially in the following generations. In the VOICING condition, on the other hand, some individual generations had a high percentage alt (conforming) items, but only one chain, 1b, had a percentage above 50% for two or more consecutive generations. In contrast, this was the case in four chains in the DEVOICING condition. The diffusion chains in DEVOICING thus retained alternating conforming items to a larger degree than those in VOICING.

In addition to the plots in Figure 5.1 and Figure 5.2 illustrating individual diffusion chains, it is relevant to assess the results collapsed for condition. Here, I have decided to do so by calculating the mean percentages in each generation in VOICING and DEVOICING rather than the median percentages. The latter measure has the advantage that it is not as sensitive to outliers as the mean. However, with only five data points in each generation, it is difficult to determine whether a given value is an outlier or not. Therefore, the mean will be presented whenever overall percentages in VOICING and DEVOICING are discussed in the following analysis. I have also calculated the medians, and these did not alter the overall findings. Furthermore, following the reasoning for using the mean, the error bars in the following figures display the full range rather than the interquartile range within each generation.

Returning to the comparison of VOICING and DEVOICING, the difference hinted at in Figure 5.2 is further supported by the mean percentage alternating conforming items in each generation in the two conditions, illustrated in Figure 5.3.

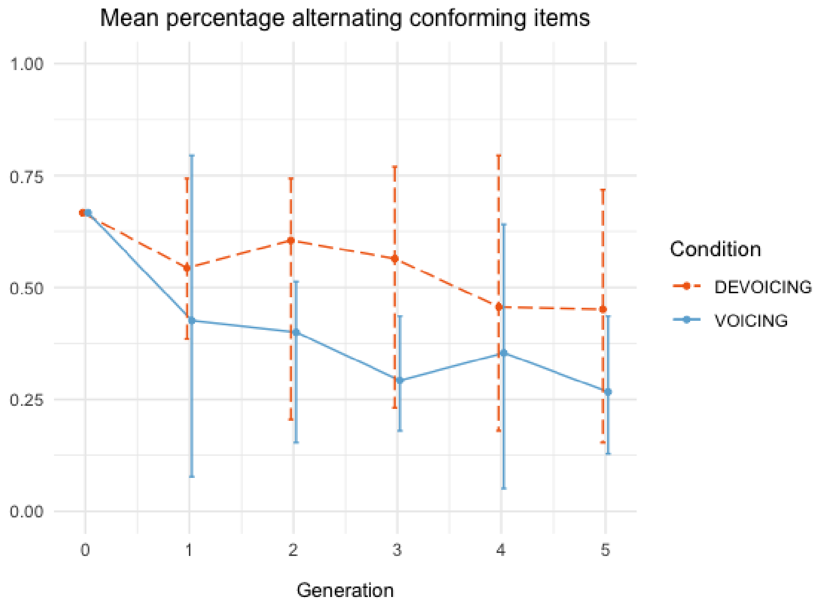


Figure 5.3 Mean percentage alt (conforming) items in each generation in VOICING and DEVOICING.

As indicated by the error bars, there is great variability between participants within each generation. However, taken with a grain of salt, Figure 5.3 shows that the mean percentage alternating conforming items was higher in the DEVOICING condition in all generations. Especially interesting here is Generation 1 and Generation 5, the beginning and the end of the chains. In Generation 1, participants' behaviour did not rely on the performance of the previous generations, and it can thus give insight into how well the two processes were learned when they were entirely consistent. In this generation, the mean percentage alt (conforming) items in the VOICING condition was 43% and corresponding percentage in the DEVOICING condition was 54%. Looking at Generation 5, the mean percentage in the VOICING condition was 35%, while the mean percentage in the DEVOICING condition was 45%. The means thus decreased in both conditions, but it was still higher in Generation 5 in DEVOICING than in Generation 1 in VOICING.

In Figure 5.4, the corresponding percentages for all conforming items are illustrated. As was the case in Figure 5.3, the error bars show that there was a large degree of variability within each generation. Overall, however, Figure 5.4 illustrates that the performances in the two conditions were largely overlapping. In Generation 1, the percentage conforming items was 77% in VOICING and 78% in DEVOICING. In Generation 5, the corresponding percentages were 53% and 59%, respectively. The impression from Figure 5.1 that there were no great differences between conditions in terms of conforming items is thus confirmed by the means in Figure 5.4.

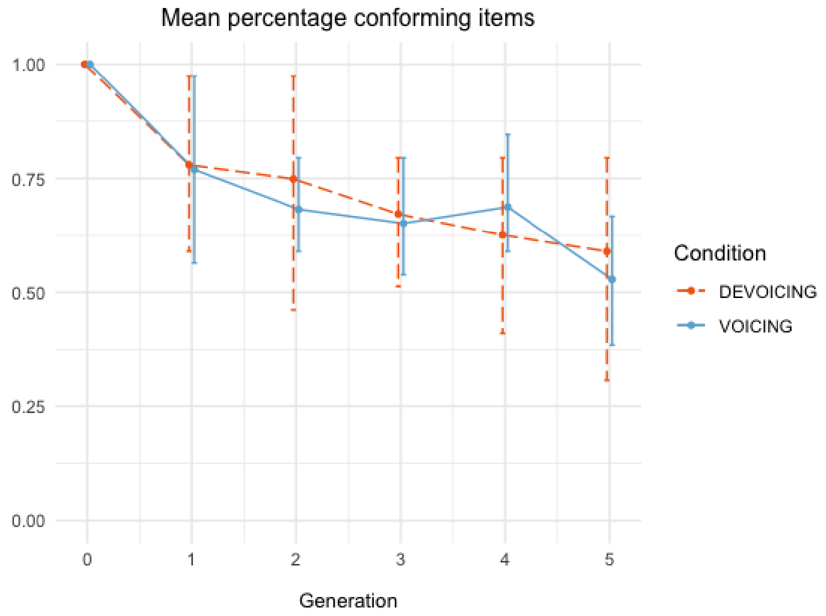


Figure 5.4 Mean percentage conforming items in each generation in VOICING and DEVOICING.

When comparing Figure 5.3 and Figure 5.4, an interesting picture emerges. While the mean percentage *alternating* conforming items was higher in the DEVOICING condition than in the VOICING condition in all generations, the mean percentage conforming items in each generation was fairly equal in the two conditions. Thus, participants learned the phonotactic constraint that only voiced or only voiceless stops were allowed in the final position in their language to the same extent in VOICING and DEVOICING, but participants in DEVOICING alternated conforming items to a greater extent than participants in VOICING. This finding resembles the alternation asymmetry found in Lysvik's (2020) Experiment 1, where participants in VOICING did not show a dispreference against voiced final stops, but rather a dispreference against the unnatural final voicing alternation. The results presented in this section suggest that a similar behaviour was found in the current experiment.

5.1.2 All item types

In addition to the distribution of the conforming items presented in Section 5.1.1, it is useful to look at the distribution of all four item types in each chain, illustrated in Figure 5.5 (see the

Appendix for tables containing all percentages in each diffusion chain). In doing so, it is possible to get an overview of the distribution of alternating and non-alternating conforming items, as well as the distribution of the non-conforming items. This latter point is important to get a fuller picture of participants' behaviour.

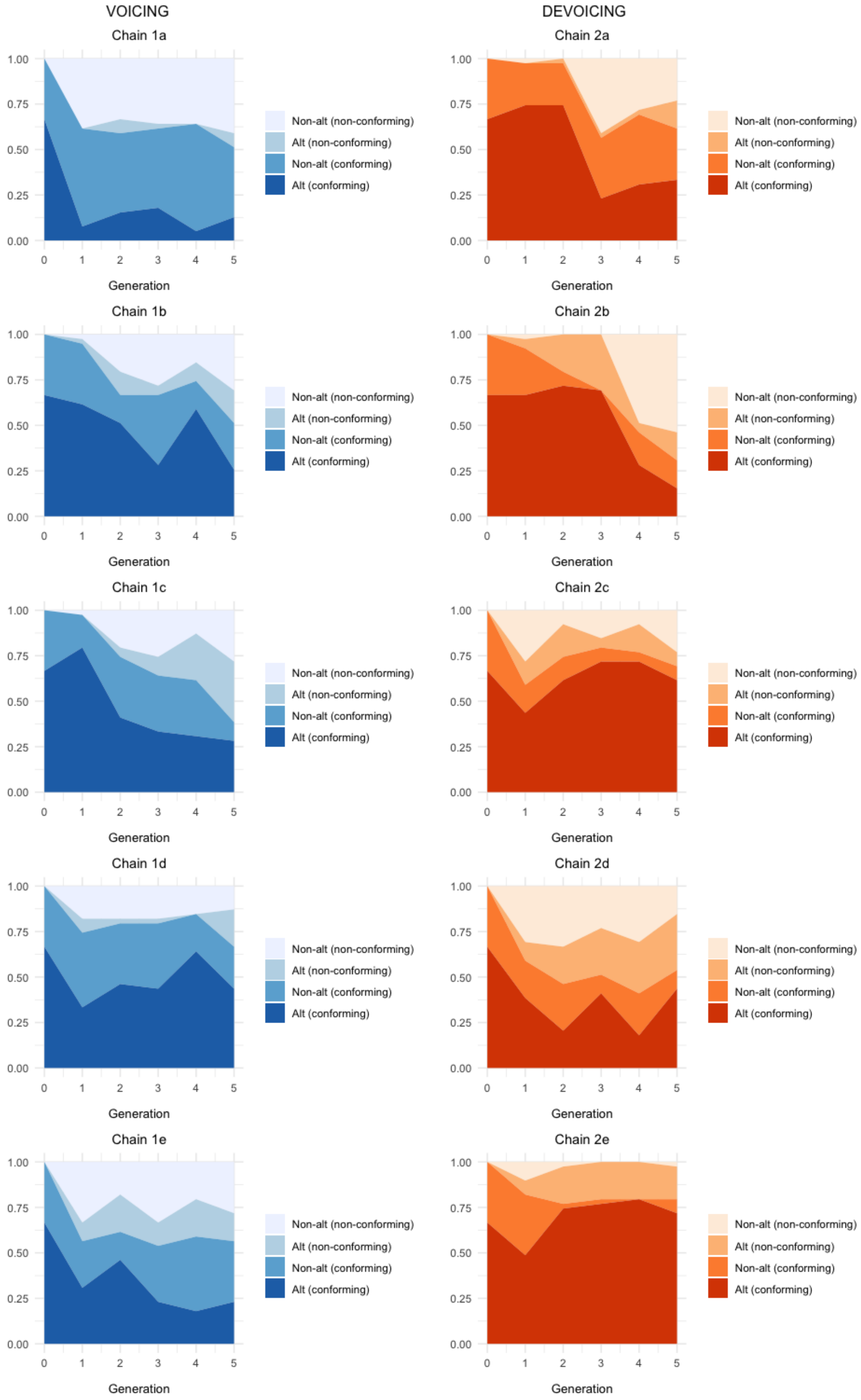


Figure 5.5 Distribution of all four possible item types in each generation in each chain.

In Figure 5.5, the two darkest colours (appearing at the bottom of each plot) represent the conforming items and thus indicate how well the relevant phonotactic restriction was learned. Starting with the five diffusion chains in the VOICING condition, it appears that many participants did not learn the final voicing rule, indicated by a low percentage conforming items. In Chain 1e, this was the case in all generations following a lack of learning in Generation 1. In this chain, the percentage conforming items was consistent around 50%, and it appears that participants chose more or less randomly between item types. Chain 1a stands out with the emergence of a different strategy than in the other chains, in which participants made the language almost exclusively non-alternating, as shown by the high percentage non-alt (conforming) and non-alt (non-conforming) items. These participants thus mainly chose the same voicing in the target as was seen in the trigger. In the remaining three chains, certain participants had a high percentage alt (conforming) items and a high total percentage conforming items, but these were mostly found in early generations, and the percentages declined towards the final generation. In these chains, a majority of non-conforming items were non-alternating.

Turning now to the DEVOICING condition, the plots in Figure 5.5 indicate that there was a large degree of variability between participants. Similarly to the VOICING condition, Generation 1 in one chain, 2d, did not learn the voicing neutralisation rule, and the development in this chain is parallel to the development in Chain 1e. In one chain, 2a, the final devoicing process was learned close to perfectly in Generation 1 and 2, but in Generation 3, the percentage conforming items dropped substantially, indicating that this participant did not learn the relevant pattern despite the consistent input. Chain 2c demonstrates the opposite development, where the percentages conforming items and alt (conforming) items dropped in Generation 1, only to increase and remain stable at a high level in the subsequent generations.

In Chain 2e and to some extent Chain 2b, an entirely different strategy emerged. At first glance, it appears that Generations 2-5 in Chain 2e and Generations 2-3 in Chain 2b learned the final devoicing process well, as the percentage alt (conforming) items is high, around 70-75%. However, the distribution within the conforming and non-conforming items suggests that this might not have been the case. Both conforming and non-conforming items were almost exclusively alternating in these generations. Instead of learning final devoicing, then, it appears that these participants learned an exchange rule where items were always alternating. Whatever the voicing of the trigger, these participants chose the opposite voicing in the target.

In Chain 2b, this rule was lost in Generation 4, but in Chain 2e, all participants from Generation 2 on learned this exchange rule.

Based on the distributions in the individual chains, then, the finding in Section 5.1.1 that the percentage conforming items is similar in the two conditions, but that participants in DEVOICING alternated more than those in VOICING, is corroborated. In the VOICING condition, four chains had a somewhat steady majority of conforming items, namely 1a, 1b, 1c, and 1d, but the percentage alt (conforming) items varied. Looking at the distribution in the generations which learned final devoicing well in the DEVOICING condition, namely Generations 1-2 in Chain 2a, Generation 1 in Chain 2b, and Generations 2-5 in Chain 2c, the majority of conforming items were alternating. So far, then, it appears that these findings point in the same direction as Lysvik's (2020) alternation asymmetry.

However, Figure 5.5 revealed that one chain in each condition, 1a in VOICING and 2e in DEVOICING, learned different rules than the final voicing and final devoicing rules in the input. In Chain 1a, participants learned that items were mostly non-alternating, while in Chain 2e, participants learned the opposite rule, that items were alternating. These two chains thus pull the mean percentage alternating conforming items, illustrated in Figure 5.3, in opposite directions in the two conditions, and their effect must be taken into account when interpreting the results. That is, before assessing the implications of the alternation asymmetry, it is necessary to investigate to what extent this asymmetry is found when Chains 1a and 2e are excluded. We will return to this issue in the discussion of the results in Chapter 6.

5.2 Alternation

In the previous sections, the distribution of alternating and non-alternating items has been discussed with regard to the proportion of conforming items which were alternating in each generation, but the overall distribution of alternating items in both conforming and non-conforming items has not been examined. Due to the potential effects of a paradigm uniformity bias, it was expected that participants would prefer non-alternating critical items over alternating critical items. However, as the original input was made up of 67% alternating items, it is possible that participants received enough evidence of the alternation to overcome this bias. Looking at the overall percentage alternating items, both conforming and non-conforming, these made up 56% of critical items when the results were collapsed for condition and generation. Across all participants, then, there were more alternating than non-alternating items, although this difference was small.

When separating by condition, a difference emerges. In the VOICING condition, 45% of critical items were alternating. In the DEVOICING condition, on the other hand, the corresponding percentage was 67%. Separating by conforming and non-conforming items, Figure 5.6 illustrates the percentage alternating items in VOICING and DEVOICING.

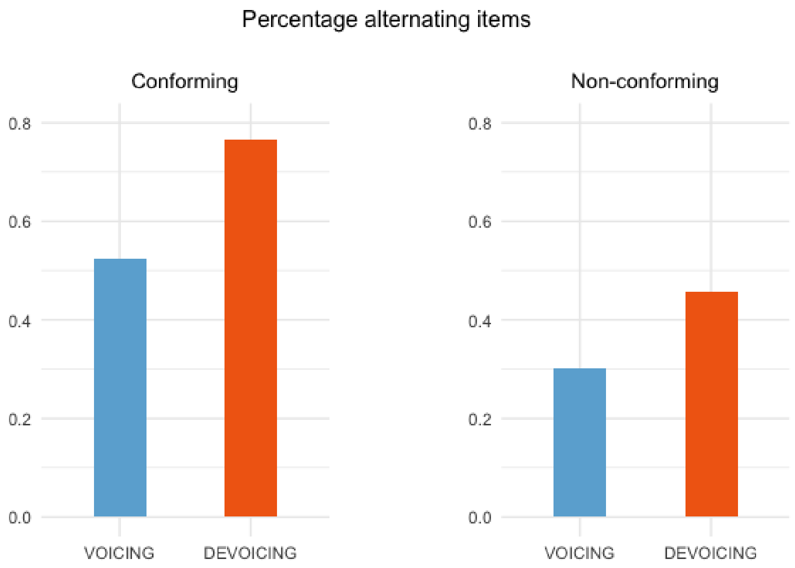


Figure 5.6 Mean percentage alternating conforming and non-conforming items in VOICING and DEVOICING.

It becomes clear that a higher proportion of conforming than non-conforming items were alternating in both VOICING and DEVOICING. Additionally, there is a difference between conditions both for conforming and non-conforming items. In VOICING, 52% of conforming items were alternating, while the percentage in DEVOICING was 77%. In DEVOICING, then, a majority of items were alternating, while in VOICING, just around half of conforming items were alternating. Among the non-conforming items, a higher percentage of items were alternating in DEVOICING than in VOICING, with 46% compared to 30%.

Thus, participants in the DEVOICING condition alternated more than those in the VOICING condition both among the conforming and among the non-conforming items. These findings indicate that participants in the DEVOICING condition to a larger extent learned that items alternated, while in the VOICING condition, participants chose more randomly between alternating and non-alternating items.

5.3 Novelty

In Section 5.1 and 5.2, an overview of participants' performances with regard to conforming items and alternating items in the VOICING and the DEVOICING conditions was presented. However, condition is not the only variable that might affect the percentage conforming and alternating conforming items. As mentioned above in Section 4.4, participants were tested on both old items, that is, items which were presented to them in the training phase, and new items in the testing phase. For each participant, 27 items, including fillers, were old, while 19 items were new. Excluding fillers for now, 20 critical items were thus old in each testing phase. It is possible that performance on old items was slightly better than on new items, but because participants could not repeat items in the training phase and had limited time to learn the items, this difference is not expected to be very big. The effect of novelty is expected to be the same for all generations in the two conditions, and these are therefore collapsed in the following calculations.

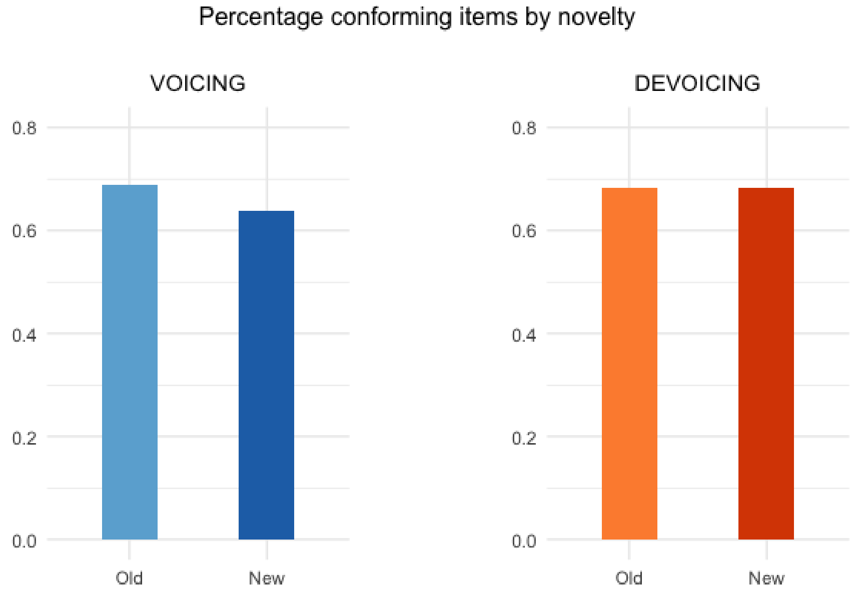


Figure 5.7 Percentage of old and new items which were conforming in VOICING and DEVOICING.

Figure 5.7 illustrates the percentage conforming items by novelty in the two conditions. In the VOICING condition, there was a very small difference between old and new items, with 69% and 64% conforming items respectively. In the DEVOICING condition, the corresponding percentage was 68% for both old and new items. In both conditions, then, it does not appear that novelty had any effect on the choice of conforming as opposed to non-conforming items. In addition, there was essentially no difference between the two conditions, which supports the finding in Section 5.1 that the mean percentage conforming items across generations in the VOICING and the DEVOICING conditions was by and large the same.

Figure 5.8 shows the corresponding plot for alternating conforming items in the two conditions.

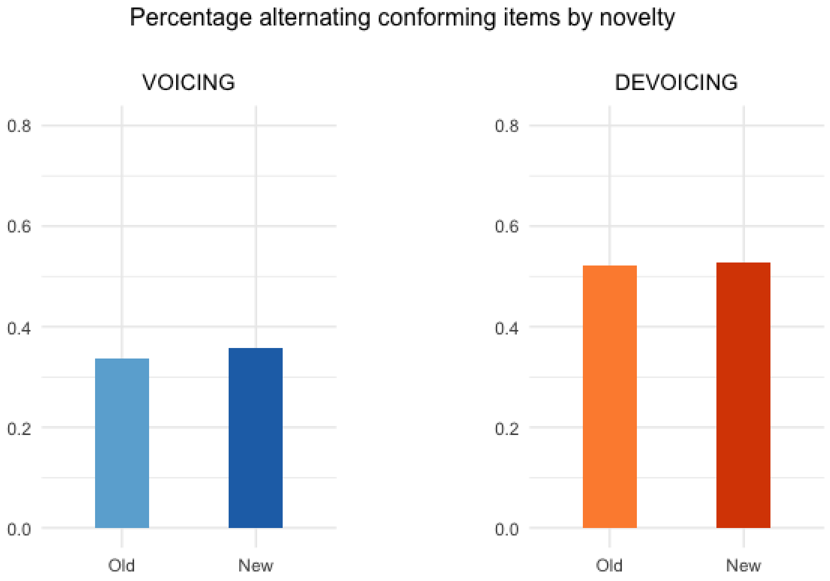


Figure 5.8 Percentage of old and new items which were alternating conforming in VOICING and DEVOICING.

As was the case for conforming items overall, there were no clear differences between old and new items within conditions in Figure 5.8. In summary then, there was no effect of novelty in either condition. However, while the percentages were parallel in the two conditions in Figure 5.7, there is a clear difference between VOICING and DEVOICING in Figure 5.8. That is, while the percentage conforming items was around 68% in both conditions, the percentage alternating conforming items was approximately 35% in VOICING and 53% in DEVOICING. This once again illustrates the alternation asymmetry found in Sections 5.1 and 5.2.

5.4 Length

Another variable element is found in the design of the stimuli, namely item length. Stems were either monosyllabic (short), such as *dud/dut* ‘fish’, or bisyllabic (long), such as *ligog/ligok* ‘cat’. It was not expected that this would influence the percentage conforming items, and this was in fact the case. Overall, 67% of short items and 68% of long items were conforming, while the corresponding percentages for alternating conforming items were 44% and 43%.

When separating by condition, these results still hold, as seen in Figure 5.9.

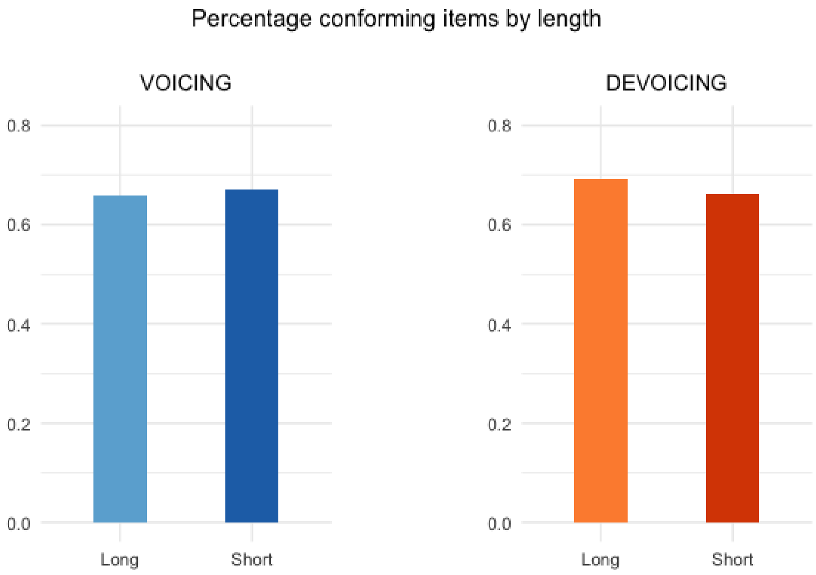


Figure 5.9 Percentage conforming items in VOICING and DEVOICING separated by length.

In both conditions, the percentage conforming items for both short and long items was around 67%, indicating that there was no effect of either length or condition. As for alternating conforming items, displayed in Figure 5.10, these once more demonstrate the alternation asymmetry between VOICING and DEVOICING, but there is no notable difference between long and short items in either condition.

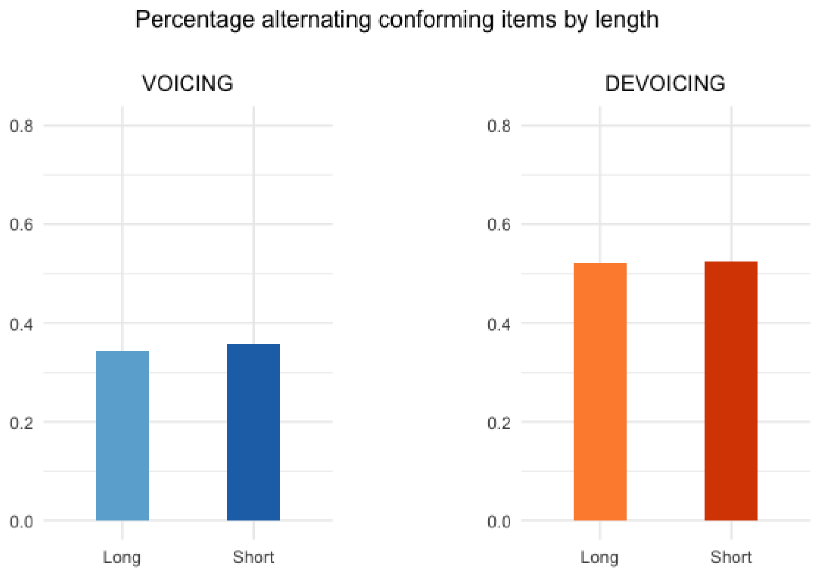


Figure 5.10 Percentage alternating conforming items in VOICING and DEVOICING separated by length.

5.5 Stop quality

The target stops in the stimuli could have three different places of articulation: labial, coronal, and velar. Each place of articulation had a voiced and a voiceless counterpart, so that items which had the voiced labial stop /b/ in the VOICING condition had the voiceless labial stop /p/ in the DEVOICING condition, as was the case with /d/ ~ /t/ and /g/ ~ /k/. It is possible that the place of articulation of the target stop had an effect on participants' behaviour, and so let us compare the overall percentage conforming items for the three consonant pairs. An important note is that in the VOICING condition, all labial stops were categorised as /b/, even if the item changed throughout the generations to contain only the voiceless stop /p/. Similarly, all coronal stops in the DEVOICING condition were classified as /t/, and so on. The voiced stops therefore represent the VOICING condition, while the voiceless stops represent the DEVOICING condition.

Looking at the percentage conforming items, these made up 66% of the labials, 69% of the coronals, and 67% of the velars. For both the labials and the velars, the percentage alternating conforming items was 43%, while for the coronals, it was 45%. It is thus safe to conclude that there was no difference between the three places of articulation when collapsed for condition. However, it is necessary to investigate the stops individually as well, as there might be effects of condition interacting with place of articulation. Figure 5.11 shows the percentage conforming items for each stop.

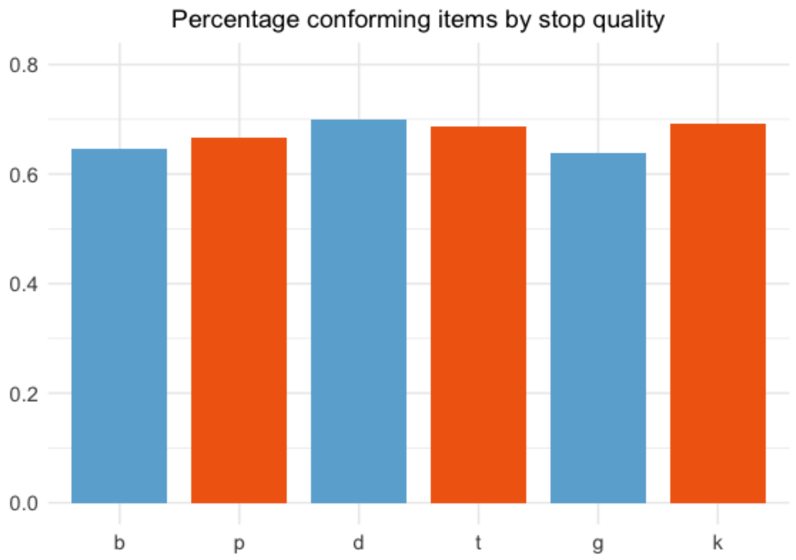


Figure 5.11 Mean percentage conforming items by the quality of the target stop consonant.

Looking at Figure 5.11, there were no substantial differences between places of articulation in VOICING (b, d, g) and DEVOICING (p, t, k), with percentages ranging from 65% to 70%. In Figure 5.12, the corresponding percentages for alternating conforming items are shown.

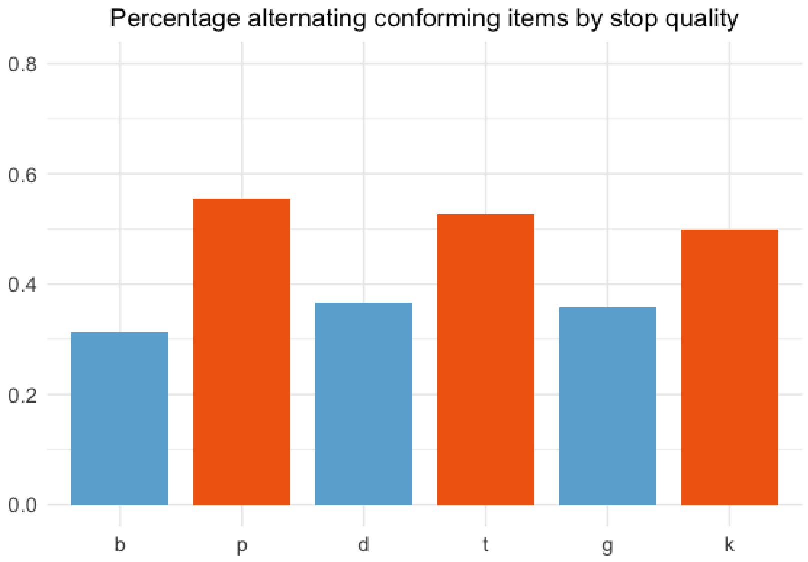


Figure 5.12 Mean percentage conforming items by the quality of the target stop consonant.

Again, there do not appear to be any great differences between places of articulation when separated for condition. /b/, /d/, and /g/ all have percentages between 31% and 37%, while the percentages for /p/, /t/, and /k/ range between 50% and 56%. Thus, the overall finding that there was no difference between labials, coronals, and velars still holds when separated for condition. As a note, Figure 5.11 and Figure 5.12 once again illustrate the alternation asymmetry.

5.6 Number

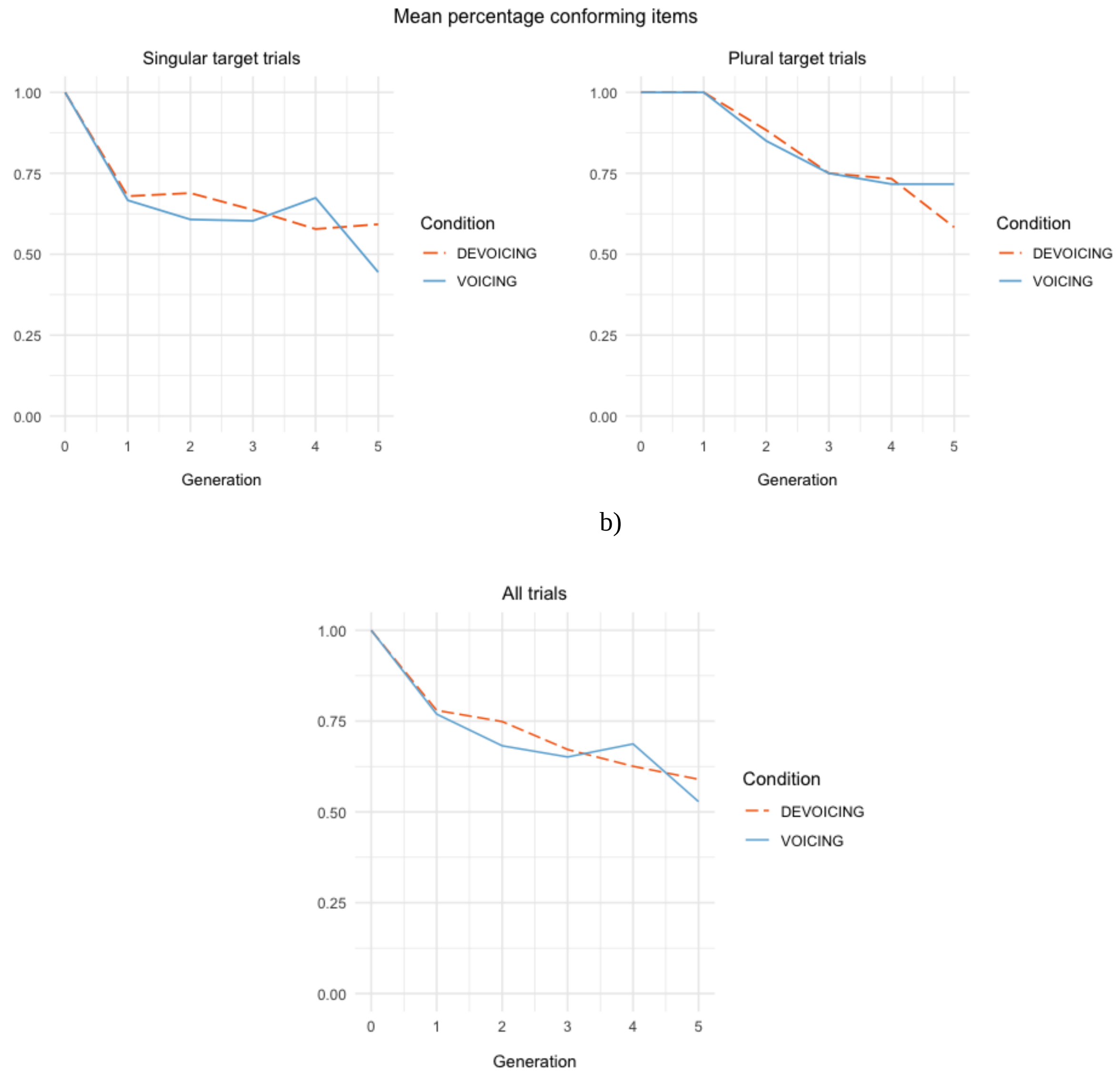
In the previous sections in this chapter, all critical item trials have been analysed together. Yet, as discussed in Section 4.4 above, 27 out of 39 critical items (69%) in the testing phase had a singular target, prompted by a plural trigger, while 12 items (31%) had a plural target, prompted by a singular trigger. When investigating participants' behaviour with regard to conforming and non-conforming as well as alternating and non-alternating items, it is also relevant to examine the results from these trials separately. As will be seen below, combining the iterated learning design with the inclusion of both singular and plural target trials has certain consequences for participants' choice in the forced-choice task, which become particularly clear when the measures of interest are the percentages conforming and alternating conforming items.

In the singular target trials, the plural trigger could have either a voiced or a voiceless target stop, and importantly, these were both allowed in both conditions. When participants chose the singular form, they could adhere to the voicing neutralisation rule in their language, or they could alter the type of the item if they chose the alternative with the non-conforming voicing. That is, if a participant in the VOICING condition was exposed to a plural trigger with the voiced stop [g] as the target stop, they could make the item non-alt (conforming) by choosing the singular form with the same voicing, [g], or they could make it alt (non-conforming) by choosing the singular form with the opposite voicing, [k]. If the plural trigger had the voiceless stop [k] as the target stop, the participant could make the item non-alt (non-conforming) by choosing the same voicing as in the trigger, [k], or they could make it alt (conforming) if they chose the form with the opposite voicing, [g]. Thus, for any given plural trigger, the participant had the choice to make the item either conforming or non-conforming. Whether those items which ended up as conforming were alternating or not depended on the voicing in the plural trigger, and so the most relevant measure in the singular target trials is the overall percentage conforming items rather than the percentage alternating conforming items.

In the plural target trials, however, participants did not have the choice of making an item conforming or non-conforming. If the singular trigger had a voiced final stop in the VOICING condition, the item would be classified as conforming regardless of whether the participant had in fact learned the phonotactic restriction that only voiced stops were allowed in the final position. Similarly, if the voiced singular trigger was presented in the DEVOICING condition, the item would be non-conforming regardless of the participant's choice of plural target. In Generation 1, the final stop in singular triggers was necessarily of the conforming voicing, as the original input only contained conforming items. In the following generations, previous participants in the chain could have altered items to have final stops of the non-conforming voicing, and if such an item was selected as a plural target trial, the item would necessarily be classified as non-conforming. The degree to which this was the case depended on how many non-conforming items had been introduced by previous participants and how many of these were randomly selected as plural target trials as opposed to singular target trials.

For this reason, calculating the percentage conforming items in the plural target trials only tells us whether the trigger had a voiced or a voiceless final stop, and this measure is thus not very informative. Yet, these trials can nonetheless inform us about participants' behaviour in the two conditions if we also investigate the percentage alternating conforming items. Specifically, the plural target trials reveal whether, given a conforming final stop in the singular trigger, participants preferred to alternate or not. This measure can provide further insight into the alternation asymmetry discussed in previous sections. The most interesting measure in the plural target trials is therefore how many of the conforming items were alternating in VOICING and DEVOICING.

The percentage conforming items in singular target trials, plural target trials, and all trials combined is shown in Figure 5.13 below. As stated above, this measure is most relevant in singular target trials, displayed in Figure 5.13a) but the plots in 5.13b) and 5.13c) are included for comparison.



a)

b)

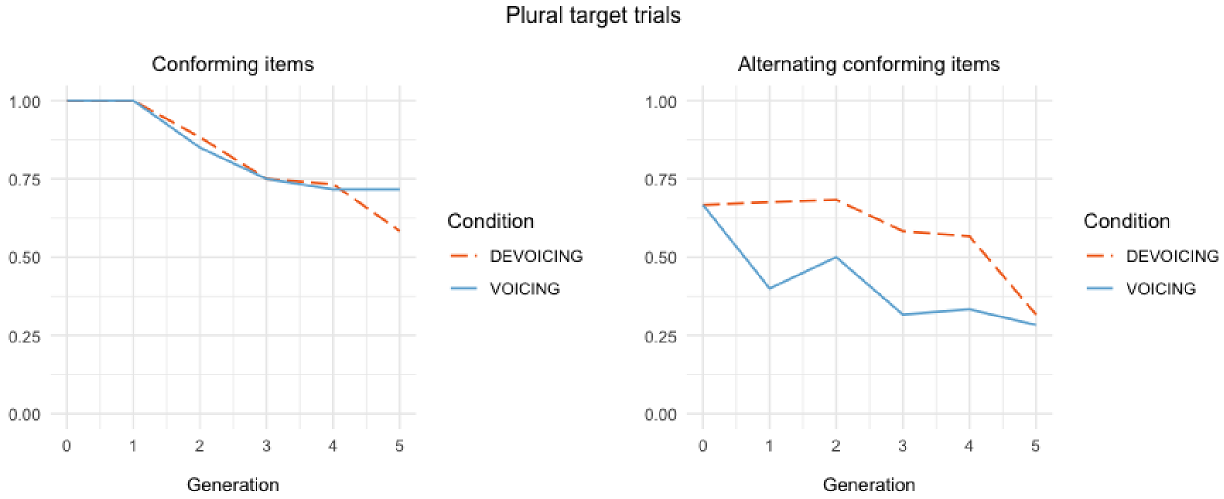
c)

Figure 5.13 Mean percentage conforming items in VOICING and DEVOICING in a) singular target trials, b) plural target trials, and c) all trials.

The three plots show similar trends, although the plural target trials in 5.13b) stand out from the remaining plots. The percentage conforming items was generally higher in the plural target trials than in the singular target trials, but this is not surprising given that these started out at 100% conforming in Generation 1. Although non-conforming items were introduced into the languages throughout the diffusion chains, these were still a minority in the plural target trials, as seen in Figure 5.13b).

The singular target trials in Figure 5.13a) show a parallel development to the combined trials in Figure 5.13c), indicating that including the plural target trials does not greatly alter the overall percentages. Figure 5.13c) demonstrates that the inclusion of plural target trials pulls the mean percentage conforming items up compared to the singular target trials alone, but because there were only 12 plural target trials for each participant, compared to 27 singular target trials, the patterns in the singular target trials affect the means to a larger extent. These trials show that the percentages in VOICING and DEVOICING were very similar, and that the findings discussed in Section 5.1 regarding the general trends in performance in VOICING and DEVOICING still hold when singular target trials are evaluated separately.

Turning now to the plural target trials, the mean percentages for conforming items are repeated in Figure 5.14a) along with the mean percentages for alternating conforming items in Figure 5.14b). It is worth keeping in mind that there were only 12 plural target trials for each participant, and so the percentages below are not based on a large amount of data.



a) b)

Figure 5.14 Mean percentage a) conforming items and b) alternating conforming items in VOICING and DEVOICING in plural target trials.

As mentioned above, the percentages conforming items in the plural target trials were high, but recall that this only means that a large number of these trials had a singular trigger which had the conforming voicing for the relevant condition. The most informative measure as to the learnability of the two processes in the plural target trials is therefore the comparison of the two plots in Figure 5.14. The percentages in plot b) illustrate to what extent the conforming items in a) were alternating, and thus whether participants chose an alternating plural form when exposed to a singular trigger with the conforming voicing. By comparing the plots in Figure 5.14, it is possible to assess in which generations the alternation asymmetry discussed in previous sections was found.

In Figure 5.14b), the percentages alternating conforming items were generally higher in DEVOICING than in VOICING. In Generation 1, participants in DEVOICING probability matched the input so that 67% of conforming items were alternating, while participants in VOICING alternated 40% of conforming items. In Generations 2, 3, and 4, the percentage conforming items in a) decreased to the same extent in both conditions, but the percentage alternating conforming items in b) continued to be higher in DEVOICING. In Generation 5 in a), the percentage conforming items was higher in VOICING than in DEVOICING, but the percentage alternating conforming items in b) was nevertheless higher in DEVOICING.

From Figure 5.14, then, it appears that a larger proportion of conforming items were alternating in DEVOICING than in VOICING. In Figure 5.15, the proportion of conforming items which were alternating in each generation in the two conditions is illustrated.

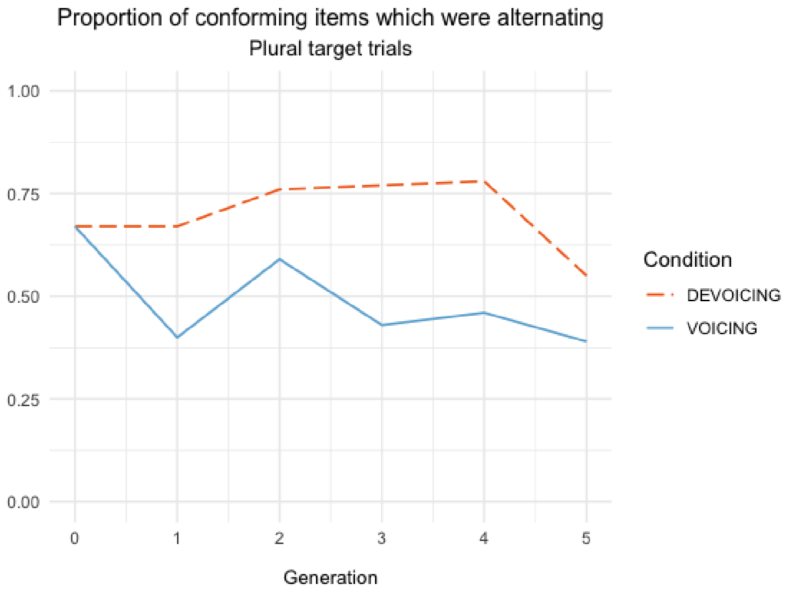


Figure 5.15 Proportion of conforming items which were alternating in plural target trials.

Figure 5.15 shows that the proportion alternating items was larger in DEVOICING in all generations. In VOICING, only one generation had a proportion above 50%, namely Generation 2. In the other four generations, participants chose the alternating form in approximately 40% of conforming plural target trials. In DEVOICING, on the other hand, all proportions were above 50%. The proportion was smallest in Generation 5, at 55%, but in all other generations, the proportions were at 67% or above. Figure 5.15 thus illustrates that the alternation asymmetry was found in all generations in the plural target trials.

Overall, then, the results obtained for plural and singular target trials combined are representative of the patterns found in singulars and plurals separately. In singular target trials, where the most relevant measure is conforming items, the performances in VOICING and DEVOICING were relatively similar. In plural target trials, where the proportion alternating items out of the conforming items is the most relevant measure, the proportions in VOICING were lower than in DEVOICING. Accordingly, the results presented in this section confirm the findings in Sections 5.1 and 5.2 that there is an alternation asymmetry between VOICING and DEVOICING.

5.7 Accuracy

In the previous sections, we have seen to what extent participants in VOICING and DEVOICING retained the final voicing and final devoicing rules. This section will investigate the extent to which participants learned the pattern they were exposed to in the input. In Generation 1, this pattern was the completely consistent final voicing and final devoicing rules, but in the following generations, the pattern in the input was determined by the output of the previous generation. By studying this measure, which I will refer to as *accuracy*, it is possible to gain an understanding of whether participants who did not learn the final voicing and final devoicing rules learned a different pattern, or whether they were choosing at random. The percentage accurate items, that is, items in which participants chose the same form as they were given in the input, for each generation in VOICING and DEVOICING is given in Figure 5.16. When calculating these percentages, the new items in the testing phase were excluded, as it is only possible to measure accuracy on the old items.

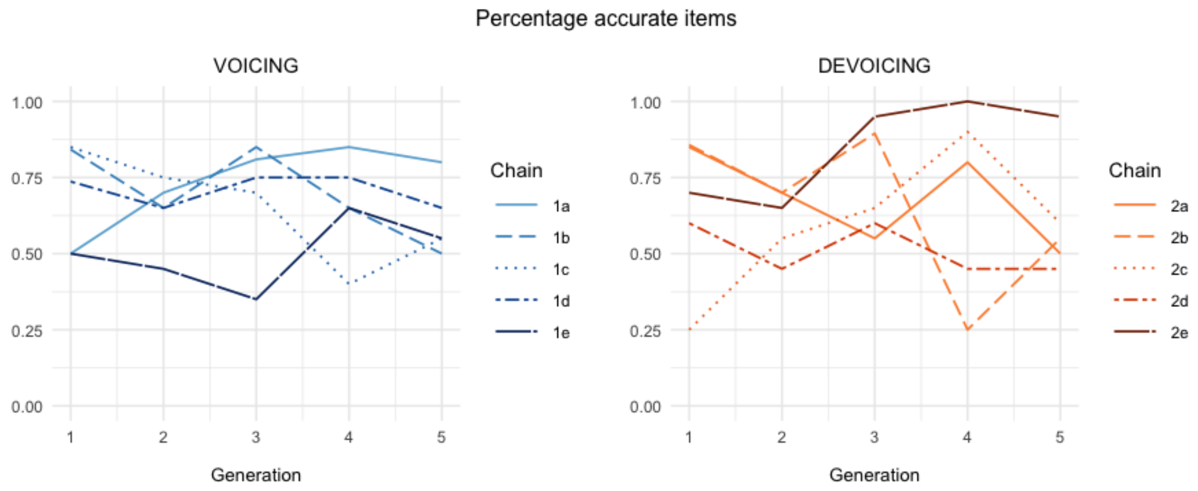


Figure 5.16 Percentage accurate items in each chain in VOICING and DEVOICING.

When accuracy is high in a given chain, this indicates that the participants either learned the relevant voicing alternation, or that they learned some other rule. Chain 1d on the one hand and Chains 1a and 2e on the other hand represent these two possibilities.

Chain 1d was the most stable chain in the VOICING condition, with participants retaining the final voicing rule throughout the generations. This stability can be seen in Figure 5.16 in that the percentage accurate items remained steady between 65% and 75%. In the two other chains, 1a and 2e, accuracy was even higher, but as discussed in Section 5.1, participants in these chains did not learn the final voicing and final devoicing processes. In Chain 1a, it rather appeared that Generation 1 chose non-alternating items, either voiced or voiceless, quite consistently. In Figure 5.16, this strategy can be seen in that the percentage accurate items was 50% in Generation 1, where the input still had the final voicing process. In the following generations, however, the percentage accurate items increased to above 75%, with 85% at the highest in Generation 4. Thus, the following generations chose the same form as in the input in the majority of items, indicating that they learned the pattern where items were generally non-alternating. The opposite pattern was found in Chain 2e, where Generation 2 learned a pattern in which items were mainly alternating. From Generation 3 on, participants learned this pattern well, with accuracy reaching 100% or just below.

An example of a chain in which participants chose more randomly is Chain 2d. The accuracy in all generations was between 45% and 60%, and so it seems like participants were not more accurate than chance. Other chains, in particular in the DEVOICING condition, show more fluctuations. It appears that accuracy and percentage conforming items can go hand in hand, as in Generation 4 in Chain 2b, where there was a steep drop in both of these measures, but this is not necessarily the case. In Chain 2c, for example, accuracy is relatively low in Generations 3 and 5 even though the previous participants in the chain and the relevant participants learned the final devoicing process. In these cases, it appears that participants matched the level of conforming items, but that the conforming items in their output were not the same as the conforming items in their input.

Overall, it appears from Figure 5.16 that the percentages for accuracy were slightly more variable in DEVOICING than in VOICING. The DEVOICING condition contained both the highest and the lowest percentages, and the diffusion chains appear to be fluctuating more than in the VOICING condition. When looking at the mean percentage accurate items in each generation in Figure 5.17 below, however, the means were very similar in the two conditions.

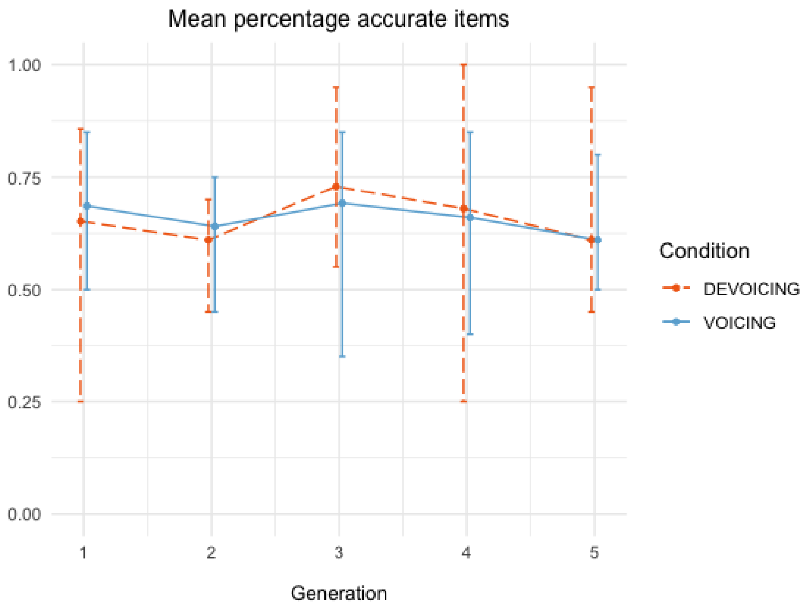


Figure 5.17 Mean percentage accurate items in VOICING and DEVOICING.

Judging by Figure 5.17, participants had on average approximately 67% accurate items in all generations in both conditions, indicating that participants learned the pattern they were exposed to in the input to the same extent in VOICING and in DEVOICING.

5.8 Fillers

In addition to the 39 critical items in the testing phase, there were 7 filler items. These were not altered between generations in order to ensure that all participants were exposed to the same fillers, and so accuracy is the relevant measure to determine how well participants learned these items. Recall that fillers were never alternating in the input, and that in the testing phase, the trigger was always the plural form and the target the singular form. In the forced-choice task, participants could choose between the form shown in the input, which contained either a nasal or a voiceless fricative, and a confound containing a stop at the same place of articulation as the nasal or fricative. Keep in mind that there were only 7 fillers in the input and in the testing phase of each participant, and so the results are based on very sparse data. The results are illustrated in Figure 5.18.

Figure 5.18 Percentage accurate filler items for each participant in VOICING and DEVOICING.

When inspecting the two plots, it appears that participants in the VOICING condition were more accurate than participants in the DEVOICING condition. This impression is confirmed when the mean accuracy in the two conditions is calculated, where the mean in VOICING was 76% and the mean in DEVOICING was 64%. Moreover, it is possible that participants performed differently on the nasal-final items and the voiceless fricative-final items, and that the segment type could interact with condition. Therefore, the mean accuracy in VOICING and DEVOICING separated by segment type is illustrated in Figure 5.19.

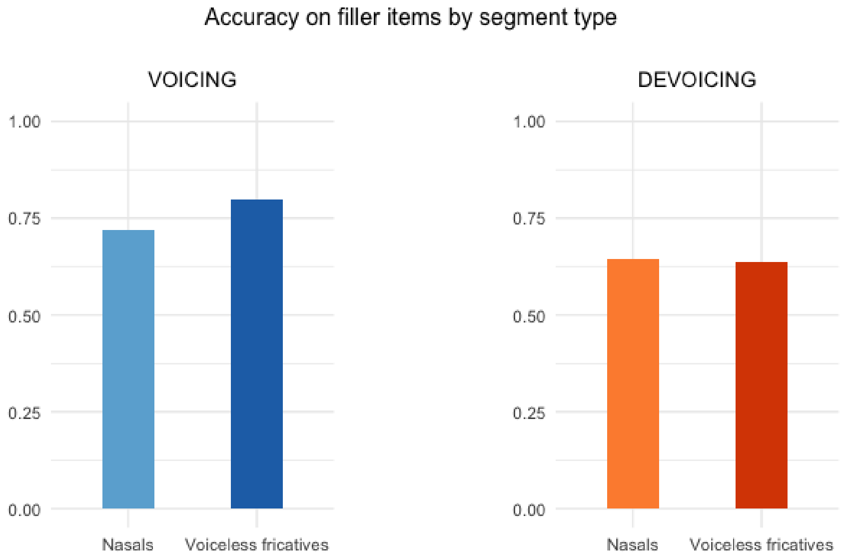


Figure 5.19 Mean accuracy on nasals and voiceless fricatives in VOICING and DEVOICING.

Figure 5.19 shows that there were no major differences in accuracy between nasals and voiceless fricatives. The difference was largest, but still small, in VOICING with 72% for nasals compared to 80% for voiceless fricatives. In DEVOICING, both segment types had a mean accuracy around 64%. Again, the percentages in VOICING were slightly higher than those in DEVOICING. Whereas participants in DEVOICING to some extent performed better on critical items, then, participants in VOICING seem to have learned the filler items to a greater extent. Possible explanations for this difference are discussed in Chapter 6.

5.9 Summary

In this chapter, the results of the current experiment have been presented through a descriptive analysis of the data. The results show that participants in VOICING and DEVOICING overall learned the relevant phonotactic restriction in their language to the same extent, although there was great individual variation within each condition. The most notable finding was that there was a difference in the percentage *alternating* conforming items, in that participants in DEVOICING alternated more than participants in VOICING. As for non-conforming items, a majority of these were non-alternating in both conditions, but a higher proportion of them were alternating in DEVOICING than in VOICING. Other variables which could possibly have affected participants' behaviour on the critical items were also analysed, but these did not appear to have had an impact on the main findings. Finally, the results from the filler item trials showed that participants in VOICING performed better on fillers than those in DEVOICING, and that this effect was not influenced by the segment type in the filler items.

6 Discussion

In this chapter, the results presented in Chapter 5 will be discussed and related to the hypotheses stated in Section 3.3 and previous ALL research. Section 6.1 discusses the patterns which emerged in the two language conditions. Section 6.2 discusses the main findings and how these relate to the relevant hypotheses. Furthermore, the implications of the results for the voicing asymmetry and the role of substantive bias in shaping this asymmetry are discussed. In Section 6.3, I discuss the implications of the results in Chapter 5 for the iterated learning paradigm in ALL experiments. The limitations of this design are presented along with an outline of the ways in which future research can mitigate these limitations. In Section 6.4, participants' performance on the filler items is discussed. I present a problematic aspect of the design of the fillers and argue that future research must take this issue into account. Finally, Section 6.5 is a summary of the chapter.

6.1 Patterns in learning

The results presented in Chapter 5 show that participants differed substantially in the extent to which they learned the pattern presented to them (see Figure 5.5 for the distribution of item types in all generations in all chains). In this section, the main groupings of participants will be discussed in order to understand to what extent participants learned the final voicing and final devoicing rules, and which patterns, if any, emerged among those who did not learn these rules.

First of all, one group of participants learned the relevant voicing neutralisation rule close to perfectly (Generation 1 in 1b, 1c, 2a, 2b, and Generation 2 in 2a). These participants had well over 90% conforming items and at least 60% alternating conforming items. All but one of these participants were in Generation 1, and the remaining participant was in Generation 2. This is not surprising, as these participants were exposed to completely consistent input, as opposed to most of the remaining generations. These participants are found in both VOICING and DEVOICING, indicating that both processes could be learned close to perfectly.

Another group of participants learned the relevant rule despite inconsistencies in the input. Chains 1d and 2c represent this category. In most generations in these chains, there were approximately 75% conforming items, and so even though 25% of items were non-conforming, these participants nevertheless learned the final voicing and final devoicing rules. Again, this pattern is found in both the VOICING condition (1d) and the DEVOICING condition (2c). However, while most conforming items in Chain 2c were alternating, the distribution of alternating and non-alternating items among conforming items was more even in Chain 1d (see Section 6.2 for discussion).

In both VOICING and DEVOICING, there were also a number of participants who had approximately 60-65% conforming items, with at least half of these being alternating (Generations 2-4 in 1b, Generations 3-4 in 1c, Generations 3-5 in 2a). In these cases, it is difficult to determine whether the participants learned the relevant rule or not. In other cases, the percentage conforming items was around 50%, and it is therefore fairly certain that the participant did not learn the final voicing or the final devoicing rule (Chains 1e and 2d). Common to all of these participants was a relatively even distribution of alternating and non-alternating items, both among the conforming and the non-conforming items. Most participants who did not learn the relevant rule, then, did not show a preference for non-alternating items, but rather chose between the four item types at random. It thus appears that a potential paradigm uniformity bias did not greatly affect their behaviour. This finding indicates that if such a bias does in fact play a role in the learning of phonological processes, there was sufficient evidence of the alternation in the original input to overcome its effects, as was the intention when designing the stimuli. Whether participants learned the relevant voicing neutralisation rule or not, then, they generally learned that items could be both alternating and non-alternating.

Certain participants stand out from the patterns above in that they did not learn the relevant voicing neutralisation rule, nor did they choose at random. Instead, these participants, found in Chain 1a and Chain 2e, learned a new rule. Two generations in Chain 2b, Generations 2 and 3, showed a similar pattern as in Chain 2e, but as this rule was lost in Chain 2b after two generations, the following discussion focuses on Chains 1a and 2e, illustrated in Figure 6.1.

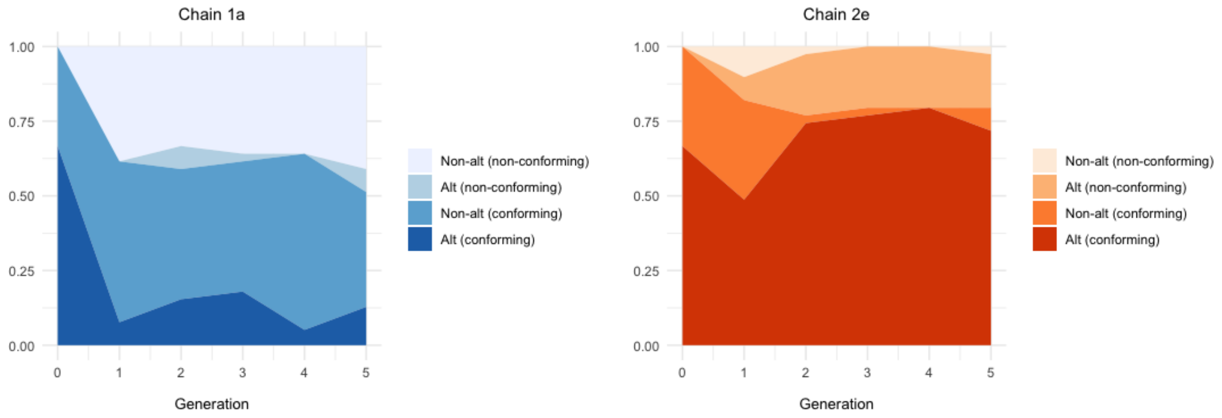


Figure 6.1 Distribution of item types in Chain 1a in VOICING and Chain 2e in DEVOICING.

In these chains, two different strategies were chosen. In Chain 1a, Generation 1 appeared to learn a system in which items generally did not alternate, as seen by the high percentage non-alt (conforming) and non-alt (non-conforming) items in Figure 6.1. Generation 1 thus introduced a rule stating that the voicing in the target should be the same as the voicing in the trigger, and the following generations maintained this pattern. As a result, a majority of items in this chain were non-alternating. The pattern found in Chain 1a can, contrary to the patterns found in the remaining chains, be understood as a paradigm uniformity bias effect. That is, a possible explanation for this finding is that the evidence of the alternation in the input to Generation 1 in this chain was not sufficient to overcome this bias, separating this participant from those in most other diffusion chains. When this pattern was introduced early in the chain, it is not surprising that it remained stable throughout the generations, as it is the simplest pattern out of all the patterns found in the experiment. It thus adheres to the pressures of both a paradigm uniformity bias and a complexity bias.

The opposite pattern was found in Chain 2e. Generation 1 to a large extent learned the final devoicing process, but from Generation 2 on, a different strategy was chosen. Generation 2 learned a pattern in which items were almost exclusively alternating, as seen by the high percentage alternating items among both the conforming and the non-conforming items. The rule introduced by Generation 2 was thus that whatever the voicing in the trigger, the opposite voicing should be chosen in the target. As Figure 6.1 shows, this pattern was learned by the remaining generations in the chain as well. At first glance, it appears that the final four generations in this chain successfully learned the final devoicing process, as the percentage alternating conforming items was around 70-75% in each generation. However, because the remaining items were not largely non-alt (conforming) but alt (non-conforming), it rather appears that these participants learned a voicing exchange pattern parallel to the one found in the Exchange condition in Glewwe's (2019) Experiment 5. The reason a majority of the items were alternating conforming rather than non-conforming is that a majority of items in the input were conforming.

The behaviour in Chain 2e illustrates that the Exchange pattern tested by Glewwe (2019) can be both innovated and learned. Glewwe (2019) argued that such a pattern is complex, as participants must learn both final voicing and final devoicing. In her Experiment 5, however, participants learned this pattern better than they learned final devoicing, and Glewwe argues that this higher performance could be due to the high exposure to alternating items in the Exchange condition. In Chain 2e, Generation 2, who innovated the exchange rule, was not exposed to more alternating items than other participants. Furthermore, as seen in Figure 6.1, this pattern was very stable between Generations 2 and 5. In fact, it was more consistent than both final voicing and final devoicing in the other diffusion chains, with between 89% and 100% of items conforming to this rule throughout four generations.

Participants thus do not appear to have any difficulty learning this rule. In fact, in featural terms, the rule is not more complex than the final devoicing and final voicing rules. The exchange rule states that [α voice] becomes [–α voice], relating two instances of a single feature. In addition, this rule is simpler in that it states that all items are alternating, rather than some items being alternating and others being non-alternating. This could explain the high learnability and stability of this rule in Chain 2e.

Looking at the distribution of different types of learners, then, it appears that in both conditions, there were participants who learned the relevant rule, participants who chose item types at random, and participants who learned a different rule entirely. The results thus indicate that there were no major differences in the learnability of final voicing and final devoicing. However, in assessing the implications of the results for the voicing asymmetry and the role of substantive bias, it is also necessary to discuss the overall results for conforming and non-conforming as well as alternating and non-alternating items in the two conditions. Specifically, the alternation asymmetry between VOICING and DEVOICING found in Chapter 5 must be discussed in light of the patterns described in this section. These issues will be discussed in the next section.

6.2 Implications

In Section 3.3, I put forth the hypotheses that final devoicing would be learned to a greater extent than final voicing, and that this effect would be amplified throughout the diffusion chains in the iterated learning experiment. In this section, these hypotheses will be evaluated in light of the results presented in Chapter 5 and the discussion of the different patterns in Section 6.1.

The results show that participants in VOICING and DEVOICING overall learned the relevant phonotactic restriction in their language equally well. That is, there were as many items with a voiced final stop in VOICING as there were items with a voiceless final stop in DEVOICING. These results are parallel to those in Lysvik’s (2020) Experiment 1, where participants adhered to the phonotactic restriction in their language condition to the same extent in VOICING and DEVOICING. They also support the findings in Greenwood (2016) and Glewwe (2019) that participants do not show a preference for DEVOICING over VOICING. In both the current experiment and previous experiments, then, participants do not show a dispreference against voiced final stops.

What the results do show is that participants in the VOICING condition alternated conforming items to a lesser degree than those in DEVOICING. Looking at all generations combined, 52% of conforming items were alternating in VOICING, while the corresponding percentage in DEVOICING was 77%. Thus, the results appear to support Lysvik’s (2020) alternation asymmetry hypothesis, stating that because participants are substantively biased against the final voicing process, they avoid it by alternating items to a lesser degree.

However, the results do not directly fall in line with Lysvik’s (2020) findings. In his Experiment 1, Lysvik found that participants in DEVOICING alternated plurals significantly more than chance (82% in the production task), while participants in VOICING alternated plurals significantly less than chance (22% in the production task). Lysvik explains this finding in VOICING with a paradigm uniformity bias leading participants to preferer non-alternating items. As discussed in Section 6.1, participants in the current experiment do not appear to have been influenced by a paradigm uniformity bias to a great extent in either condition. The alternation asymmetry found here also involves a higher percentage alternating conforming items in DEVOICING than in VOICING, but contrary to in Lysvik’s experiment, participants in VOICING did not alternate below chance for critical items. In this condition, 52% of conforming items were alternating, and it thus appears that rather than avoiding the alternation, participants chose more or less randomly between alternating and non-alternating items. That is not to say that the alternation asymmetry cannot be the result of a substantive bias, but if this is the case, a different strategy than paradigm uniformity must have been used.

However, in attempting to discover what this strategy might be and whether the observed asymmetry was in fact due to a substantive bias effect, the results from Chains 1a and 2e must be taken into account. In Chain 1a, participants learned that items were non-alternating, while in Chain 2e, participants learned that items were alternating, and so it is necessary to investigate whether the alternation asymmetry is still found when these two chains are excluded. That is, if it is argued that the reason participants in DEVOICING chose alternating conforming items more often than those in VOICING is because they learned the relevant rule to a larger extent, we cannot include Chain 2e in the calculation, a chain with a very high percentage alternating conforming items, but where we know that participants in fact did not learn final devoicing. Similarly, Chain 1a cannot be taken to show that participants in VOICING adhered to the phonotactic restriction to the same degree as in DEVOICING, but alternated less, because we know that participants in this chain learned a different rule entirely.

It is possible that these two chains contribute substantially to the alternation asymmetry. Figure 6.2 shows that this was the case to some degree. This figure can be compared to Figure 5.3, which includes all chains, repeated for ease of comparison in Figure 6.3.

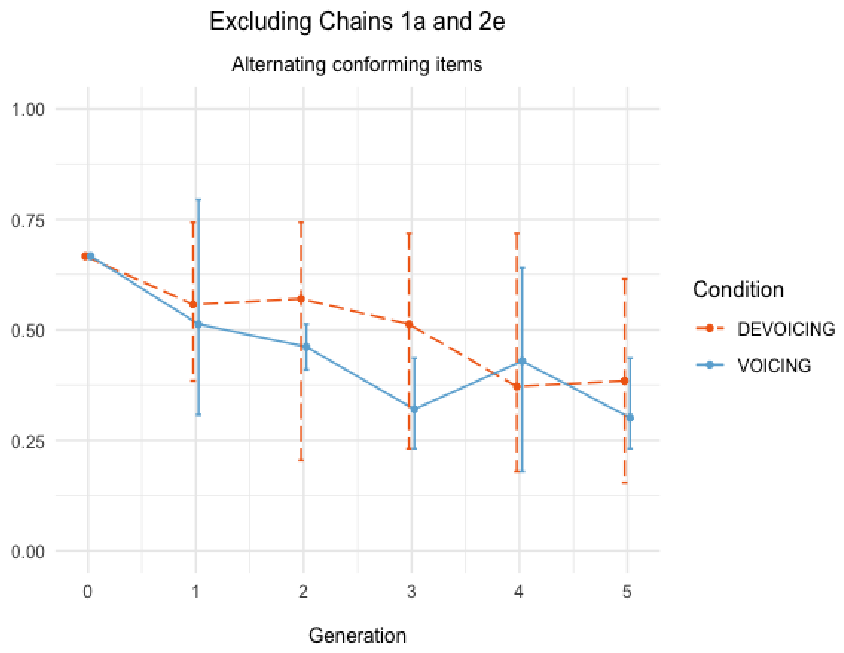


Figure 6.2 Mean percentage alternating conforming items in Voicing and Devoicing excluding Chains 1a and 2e.

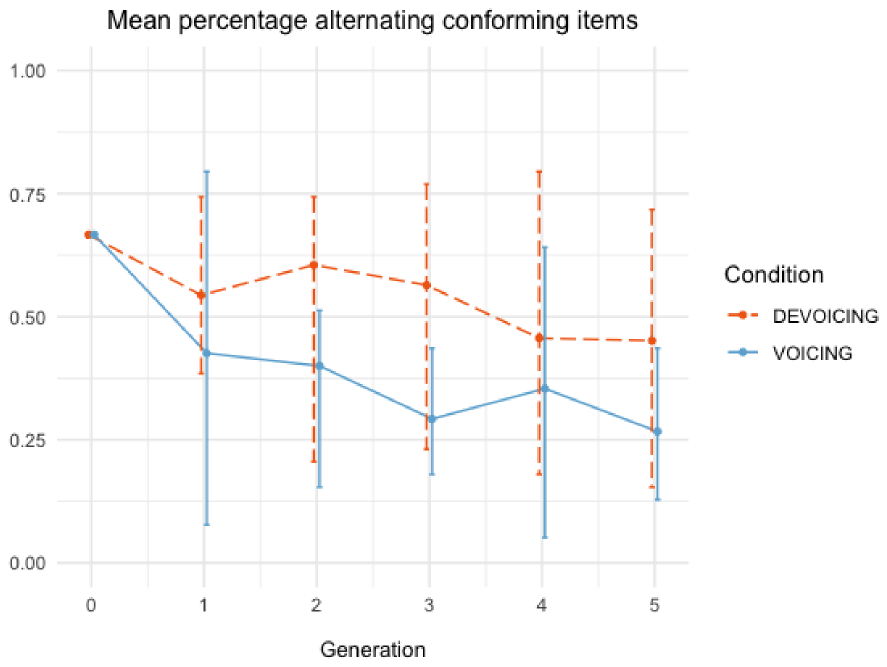


Figure 6.3 Mean percentage alternating conforming items in Voicing and Devoicing, including Chains 1a and 2e.

When Chains 1a and 2e are excluded, the difference between Voicing and Devoicing was largest in Generations 2 and 3, with 46% in Voicing compared to 57% in Devoicing in Generation 2 and 32% compared to 51% in Generation 3. In the remaining generations, the two conditions were fairly equal. A comparison of Figure 6.2 and Figure 6.3 illustrates that the alternation asymmetry diminished when Chains 1a and 2e were excluded. There was still an overall difference in the percentage alternating items out of all conforming items in the two conditions, with 60% in VOICING compared to 73% in DEVOICING, but it appears that this difference can to a large extent be attributed to the results in Generations 2 and 3. Recall that it is in exactly these generations the exchange pattern found in Chain 2e was also found in Chain 2b. In fact, in Generation 3, where the difference was largest, all conforming items in Chain 2b were alternating. Keeping in mind these findings, then, the hypothesis that the alternation asymmetry is indicative of an actual discrepancy in how well the final voicing and final devoicing processes were learned is weakened.

Moreover, when Chains 1a and 2e are excluded, the argument against a paradigm uniformity bias explanation is strengthened. When the proportion alternating conforming items is calculated for the eight remaining chains, a majority of conforming items in both VOICING and DEVOICING were alternating. Additionally, if the alternation asymmetry is in fact indicative of a substantive bias effect, it lacks a straightforward explanation. That is, if participants in VOICING and DEVOICING learned the relevant phonotactic restriction equally well and also alternated items in more than 50% of cases, there does not seem to be a principled reason to assume that the difference between the conditions is due to the final devoicing process being learned better than the final voicing process.

I stated above that it is possible that the alternation asymmetry was due to a substantive bias effect, but that participants employed a different strategy than making paradigms uniform to avoid the final voicing process. From the discussion of the results when Chains 1a and 2e are excluded, however, it is not immediately apparent what this strategy might be. This argument further weakens the hypothesis that the alternation asymmetry is due to a substantive bias effect.

Evaluating the first part of the hypothesis formulated in Section 3.3, then, there was no difference between VOICING and DEVOICING in the learning of the phonotactic restriction in the word-final position, which is the strongest indicator of having learned final voicing or final devoicing. There was a difference between conditions in that a higher percentage of conforming items were alternating in DEVOICING, although it is not certain that this difference resulted from a difference in learning, especially when the deviating Chains 1a and 2e are excluded. What does appear to be clear is that if this difference did in fact arise from a dispreference against final voicing, participants did not resort to paradigm uniformity to avoid the process. In Section 3.2.2, I mentioned that Lysvik (2020) argued in favour of a paradigm uniformity bias explanation for the results in his Experiment 1, but that he conceded that channel biases or the nature of the synthetic stimuli could also be responsible for the observed alternation asymmetry. The results of the current experiment suggest that these alternative explanations might be more accurate.

We now turn to the second part of the hypothesis, namely that any substantive bias effects should be amplified throughout the diffusion chains. If this hypothesis holds true, participants in VOICING, but not in DEVOICING, should gradually decrease the number of conforming items from Generation 1 to Generation 5. Furthermore, if the alternation asymmetry was in fact an effect of substantive bias, participants in VOICING are also expected to increasingly choose non-alternating conforming items leading up to Generation 5, while the percentage alternating items out of the conforming items should remain fairly stable in DEVOICING. In investigating these developments, it is possible to determine whether the overall results hide a declining trend in the VOICING condition, which could in fact be compatible with both a substantive bias effect and a paradigm uniformity bias effect. Moreover, in comparing the development in VOICING with that in DEVOICING, we can gain an understanding of whether a potential decline was unique to the VOICING condition.

Figure 6.4 illustrates the percentage conforming items in each generation when Chains 1a and 2e are excluded, parallel to Figure 5.4 in Chapter 5.

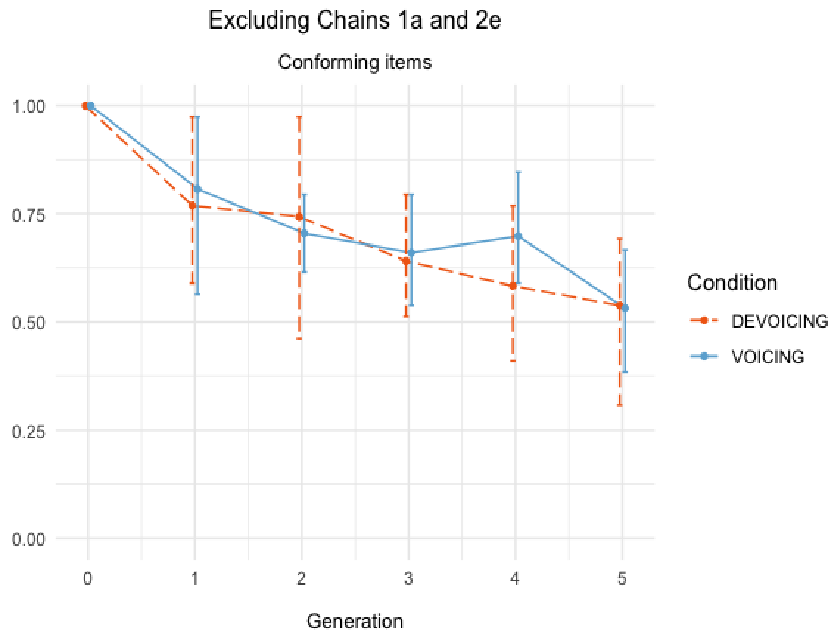


Figure 6.4 Mean percentage conforming items in VOICING and DEVOICING excluding Chains 1a and 2e.

In Generation 1, VOICING and DEVOICING both had a percentage around 80%, and by the end of the diffusion chains in Generation 5, the percentages had dropped to just above 50%. There was thus a clear decline in the percentage conforming items in both conditions, revealing that the final voicing and final devoicing processes both became less stable over the course of the five generations. In so far as participants in VOICING got rid of the unnatural final voicing process, then, participants in DEVOICING also got rid of the natural final devoicing process. As will be discussed further in Section 6.3 below, the languages in the two conditions became less consistent and thus more complex as they evolved through the diffusion chains. The decline shown in Figure 6.4 therefore appears to be an effect of a general difficulty with learning the relevant patterns when there were inconsistencies in the input rather than an effect of substantive bias in the VOICING condition.

As for the proportion of conforming items which were alternating in the two conditions, this is illustrated in Figure 6.5, which once again excludes Chains 1a and 2e.

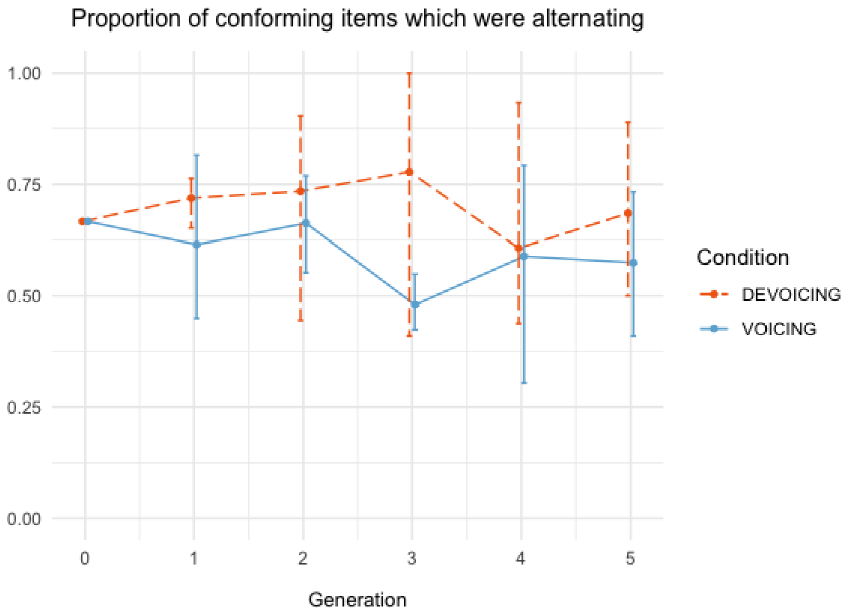


Figure 6.5 Proportion of conforming items which were alternating in VOICING and DEVOICING excluding Chains 1a and 2e.

In Generation 1, 61% and 72% of conforming items were alternating in VOICING and DEVOICING, respectively. In Generation 5, the corresponding percentages were 57% and 69%. Thus, there was no substantial decline throughout the diffusion chains in either condition. We can therefore entirely discard a paradigm uniformity bias as an explanation for the difference between VOICING and DEVOICING in the current experiment. Moreover, there is no evidence that the alternation asymmetry was amplified in the iterated learning experiment. There are two possible explanations for this. One the one hand, and perhaps most likely based on the discussion above, it is possible that the difference between conditions is coincidental and not due to substantive bias. On the other hand, it is possible that the effect can in fact be attributed to substantive bias, but that it was not amplified in the iterated learning experiment. We will return to this last point in Section 6.3 below.

Based on the discussion above, then, it can be concluded that this experiment did not find clear effects of substantive bias, which was the bias under investigation. The finding in Greenwood (2016), Glewwe (2019), and Lysvik (2020) that participants do not display a dispreference against voiced final obstruents is confirmed, and the hypothesis that such a dispreference could be too weak to manifest itself in just one generation, but that it could be amplified and become apparent throughout the diffusion chains, does not appear to hold true based on the current findings. Consequently, this experiment contributes to a growing body of evidence indicating that a learning bias disfavouring voiced final stops cannot underlie the voicing asymmetry. In the attempt to explain why final devoicing is typologically common, while final voicing is typologically very uncommon, then, it appears that a potential substantive bias must affect learning in a different way, or the typological asymmetry must have a different explanation entirely, as argued by proponents of the channel bias account.

A way in which substantive bias can affect the learning of final voicing without targeting voiced final stops is by impeding the learning of the unnatural alternation between voiceless intervocalic stops and voiced final stops. The alternation between voiced intervocalic stops and voiceless final stops, found in the final devoicing process, should under this account not be subject to the same avoidance strategy. The current experiment found an asymmetry in this direction, in that participants in VOICING alternated slightly less than those in DEVOICING. However, when taking into account the participants who learned a different pattern entirely and the lack of evidence for a paradigm uniformity strategy, the results provide little support for an interpretation of the alternation asymmetry as a substantive bias effect.

When these findings are seen in light of previous ALL research investigating the voicing asymmetry, such as Myers & Padgett (2014), Greenwood (2016), Glewwe (2019), and Lysvik (2020), it appears that the evidence for substantive bias is sparse. If ALL experiments are in fact able to tap into underlying biases, the fact that most experiments fail to find substantive bias effects weakens the hypothesis that such a bias is responsible for the typological asymmetry between final devoicing and final voicing. Some indications for the role of substantive bias have been found, most notably in Myers & Padgett (2014) and Lysvik (2020). However, the results of the current experiment call into question whether the alternation asymmetry found in Lysvik (2020) is best explained as a substantive bias effect. In any case, the question remains whether these effects, which are arguably weak, are sufficient to cause the typological asymmetry in natural language. Based on the findings in the experiment conducted in this thesis, future experiments investigating effects of substantive bias should focus specifically on the alternation asymmetry to determine whether this effect is replicated and, if so, whether it is in fact strong enough to cause the voicing asymmetry. However, considering the evidence presented here, future research should also specifically investigate effects of channel bias.

6.3 Iterated learning

As discussed in Section 3.2.3, the iterated learning paradigm is the most uncommon of the four paradigms used in ALL research, and no previous ALL experiment has used the iterated learning paradigm in the study of phonological processes. Therefore, the current experiment provides novel insights into this experimental design. The most evident experience from the current experiment is that the iterated learning design is very sensitive to individual differences. This became apparent already in Generation 1, where certain participants learned the voicing neutralisation rules close to perfectly, others failed to learn the relevant rule and appeared to choose more randomly, and others again learned a completely different rule. All types of participants were found in both conditions, and the stability of a chain appeared to largely depend on which participant coincidentally followed another.

Such individual variation will be found in any experiment, but in the iterated learning paradigm where each generation in a chain is made up of only one participant, it necessarily affects the results to a large degree. If one participant in the chain does not learn the relevant rule, there is no evidence of that rule in the input to the next generation, and so this participant will not learn the rule either. This was the case for example in Chains 1e and 2d, where the voicing alternation rules disappeared already in Generation 1. The following generations thus cannot be said to have failed to learn the relevant rule, and it is possible that they would in fact have learned it if they had received evidence for it.

Furthermore, throughout the generations, complexity became a factor. I argued in Section 2.2.3 that the final voicing process and the final devoicing process were of equal complexity and that the voicing asymmetry thus serves as a suitable asymmetry to investigate in ALL experiments. However, this is only true in Generation 1. In the following four generations, complexity is introduced into the chains in that a simple rule such as *[+voice]# or *[–voice]# cannot capture the pattern participants are exposed to in the training phase. In fact, apart from a few participants in DEVOICING who consistently alternated items, no rule can capture the relevant patterns because no pattern was consistent. When faced with this complexity, participants showed different behaviours. Certain participants learned the relevant pattern despite the inconsistencies in the input, for example Generation 2 in Chain 2c, while such inconsistencies led others to choose the types of items more randomly, as in Generation 5 in Chain 1c (see Figure 5.5). Ultimately, this complexity caused the final voicing and final devoicing rules to disappear in most diffusion chains.

Turning to the directions for future research, future iterated learning experiments should attempt to mitigate the large effect of individual differences discussed above. This can be done in different ways, for example by extending the training phase for each participant. As mentioned in Section 3.3, it is not entirely straightforward to include a threshold to which participants have to learn the relevant pattern in an iterated learning experiment, but it is possible to expose participants to more input. The current experiment was conducted online for practical reasons, and the training phase was therefore short, but it is possible that some of the participants who did not learn the relevant pattern would have learned it to a larger degree with more input.

What is more, it would be beneficial to include more diffusion chains with more generations. With five diffusion chains in each condition, it is not possible to generalise the results to a larger population, but if more chains were included, it would be possible to say more about performance on a group level. Furthermore, if each chain consisted of more generations, we could get a clearer picture of whether chains ultimately stabilise and, if so, which patterns emerge. First and foremost, it is interesting to find out whether the structures under investigation, in this case the final voicing and final devoicing processes, are retained over the course of for example ten generations, and if there is a difference between the two processes. In addition, it can be useful to discover which alternative patterns emerge. In most chains in the current experiment, the pattern found in Generation 5 sees an even distribution of the four types of items, as discussed above. It is possible that if there were ten generations in each chain, participants would get rid of the variability and construct languages conforming to a specific rule, such as Chains 1a and 2e, in which items were generally non-alternating and alternating, respectively (see Figure 6.1). Such a result would support the findings from iterated learning experiments in other subfields, such as Smith & Wonnacott (2010), showing that participants regularise variable input.

Yet, an important experience from this experiment is also that there were no indications that the potential substantive bias effect was amplified throughout the generations. Both the final voicing process and the final devoicing process were lost in most chains, and so as conducted here, this design reveals more about participants' ability to learn consistent and inconsistent patterns than the transmission of natural and unnatural patterns over time. That is not to say that the iterated learning design cannot inform on the stability of phonological processes, but it appears that it is necessary to conduct an experiment along the lines discussed above to be able to fully evaluate the usefulness of this design in the investigation of typological asymmetries in phonology. In providing the experiences discussed in this section, then, the current experiment constitutes a stepping stone in the further development of this experimental method.

6.4 Filler items

As seen in Section 5.8, participants in VOICING had a higher accuracy in filler items than participants in DEVOICING. Fillers never alternated in the input, but in the forced-choice task, participants were asked to choose whether the non-alternating form or an alternating confound form was the correct singular target when exposed to a plural trigger. Participants in DEVOICING chose the alternating confound form to a larger degree than those in VOICING.

A possible explanation for this difference is that participants' learning of the critical items, where participants in DEVOICING alternated items more than in VOICING, affected the learning of the filler items. If a large proportion of participants in DEVOICING hypothesised that a majority of critical items alternated, it is possible that they extended this hypothesis to filler items as well. However, if such a scenario was indeed the case, we would expect to find that those participants who alternated the most in both conforming and non-conforming items would show the lowest percentage accuracy in fillers. Generations 2-3 in Chain 2b and Generations 2-5 in Chain 2e show such a pattern in the critical items, but as seen in Figure

5.18 above, these participants generally had a high percentage accuracy in fillers. This hypothesis is thus weakened.

It can be hypothesised, then, that the explanation lies in the VOICING condition. That is, because participants in VOICING to a larger extent chose non-alternating critical items, they also chose non-alternating filler items, giving higher accuracy in this condition than in DEVOICING. If this was the case, participants in Chain 1a, a largely non-alternating chain, would be expected to have high accuracy on fillers. From Figure 5.18, it appears that this was in fact the case, and this hypothesis thus remains a plausible explanation for the higher performance in VOICING.

Yet another factor which must be discussed with regard to the filler items is the nature of the target consonants. Glewwe (2019) found that the inherent voicing of the fillers influenced how well a phonotactic restriction against word-final voiced stops and a phonotactic restriction against word-final voiceless stops was learned, and she argues that the fillers should therefore include both inherently voiced and inherently voiceless segments in the final position. This is the basis for including fillers with both nasals and voiceless fricatives as the target consonant in the current experiment.

However, the fact that voiceless fricatives are obstruents and can therefore be interpreted as patterning with stops in voicing neutralisation processes must be taken into account. Specifically, in the DEVOICING condition, a constraint stating that only voiceless obstruents are allowed word-finally can capture the pattern found in both the critical items and the fillers, whereas in the VOICING condition, obstruents do not pattern together, and participants must learn that stops are voiced word-finally and fricatives are voiceless word-finally. This could contribute to better learning of the final devoicing rule than the final voicing rule.

Nevertheless, there is reason to believe that this did not greatly influence the behaviour in the two conditions. If the hypothesis above were true, we would either expect the fillers to influence the learning of the critical items, or the critical items to influence the learning of the fillers. Recall that in the filler trials, participants had the choice between the voiceless fricative or nasal found in the input and a confound form with a stop at the same place of articulation as the target consonant and the voicing corresponding to the neutralisation rule in the relevant condition. In the VOICING condition, then, participants had the choice between the correct form *tipis* ‘hammer’ and the confound *tipid* in the trials containing a voiceless fricative as the target consonant. If participants in VOICING grouped obstruents together, it is reasonable to assume that they would to a larger degree than those in DEVOICING choose the confound form in these trials because the confound, but not the voiceless fricative, conformed to the final voicing rule. However, as discussed, participants in VOICING chose the confound form in fillers to a lesser degree than those in DEVOICING. As seen in Figure 5.19, this was also the case when looking at fillers with a voiceless fricative separately, and so this hypothesis does not appear to be accurate.

Alternatively, if participants in VOICING grouped obstruents together, it is possible that they would have a lower percentage conforming items among the critical items than participants in DEVOICING. Specifically, if these participants learned that obstruents were voiceless in the word-final position, as evidenced by the fillers, they might choose a lower number of items with a voiced final stop. As discussed in Section 6.2, however, there was no difference in the percentage conforming items in VOICING and DEVOICING, which indicates that the voiceless fricative fillers did not influence the learning of the critical items in VOICING.

The discussion above indicates that the inclusion of voiceless fricative fillers did not lead to the final voicing rule being learned to a lesser degree than the final devoicing rule. Future research should nonetheless take this issue into account when designing the critical items and the fillers. One way of making sure the nature of the fillers does not influence the learning of the critical items is to include vowel-final fillers only. In doing so, the fillers can to a larger extent be excluded as a possible confounding factor.

6.5 Summary

In this chapter, the implications of the findings presented in Chapter 5 were discussed. Firstly, the different patterns which emerged in the testing phase of the experiment were discussed. Subsequently, the hypotheses stated in Section 3.3 were evaluated faced with the current results, and it was concluded that this experiment did not find clear effects of substantive bias. The first hypothesis, namely that final devoicing would be learned better than final voicing, did not hold true when looking at word-final stops. There was, however, a difference between VOICING and DEVOICING in that a slightly higher proportion of conforming items were alternating in the latter group, referred to as the alternation asymmetry. It is possible that this effect resulted from a substantive bias disfavouring the learning of final voicing, but due to the small size of this difference and the fact that it cannot be explained as a paradigm uniformity bias strategy, it appears more likely that the asymmetry arose by chance.

The second hypothesis, that any substantive bias effects would be amplified throughout the diffusion chains, was not confirmed in this experiment. The percentage conforming items declined to the same degree in both the VOICING and the DEVOICING condition, and the proportion of conforming items which were alternating also remained stable in both conditions. This latter finding has two possible explanations. Either the alternation asymmetry was in fact not an effect of substantive bias, or the iterated learning design used here failed to facilitate transfer between generations.

As conducted here, the iterated learning method is very sensitive to individual differences between participants and to complexity effects. I argue that only by conducting a more extensive, large-scale iterated learning experiment can the usefulness of this paradigm in ALL research in phonology be properly assessed. Future experiments should thus take the experiences made in this experiment into account.

Finally, the nature of the filler items was discussed. It was argued that although these did not appear to have affected the learning of the critical items in the current experiment, future research should attempt to further control for their influence, for example by including vowel-final fillers only.

7 Conclusion

In this thesis, I have investigated the typological asymmetry between final devoicing and final voicing through the use of an artificial language learning experiment. Specifically, the experiment aimed at uncovering why the final devoicing process is typologically common, while the final voicing process is typologically uncommon. In previous work on this and other typological asymmetries, two main explanations have been posited for such skews in distribution, the channel bias account and the substantive bias account. The former attributes the voicing asymmetry to final voicing being phonetically unnatural, while the latter holds that this phonetic unnaturalness is encoded in a language-specific learning bias disfavouring final voicing. Most previous research has attempted to find effects of substantive bias, but the results have been varying, and the role of substantive bias in shaping the typological asymmetry between final devoicing and final voicing remains unsettled.

This thesis presented an ALL experiment isolating possible effects of substantive bias. The experiment was conducted using the iterated learning paradigm, in which participants were organised into diffusion chains consisting of multiple generations. The purpose of this design was to investigate whether potential effects of substantive bias which were not found in previous, individual-based experiments were amplified and thus became more apparent in the iterated learning experiment. Specifically, the experiment tested the hypotheses that participants learning a final voicing language would learn the relevant pattern to a lesser extent than participants learning a final devoicing language, and that this effect would be amplified throughout the generations in the diffusion chains.

The results did not show clear effects of substantive bias, as participants in the VOICING condition and participants in the DEVOICING condition learned the relevant phonotactic restriction in their language to the same degree. The results did show an alternation asymmetry, in that a higher proportion of the items conforming to the relevant phonotactic restriction were alternating in the DEVOICING condition than in the VOICING condition. However, this effect was small, and it diminished when different learning strategies among participants were taken into account. As for the second hypothesis, there was no effect of transmission between generations with regard to potential substantive bias effects. The alternation asymmetry remained stable throughout the generations, and overall, both the final voicing process and the final devoicing process largely disappeared by the end of the diffusion chains.

This experiment contributes to the body of research on the asymmetry between final devoicing and final voicing using the ALL method in several ways. Firstly, it confirms the finding in previous experiments (Greenwood, 2016; Glewwe, 2019; Lysvik, 2020) that participants do not show a dispreference against voiced final stops. Based on these experiments, then, it can be argued that such a dispreference is unlikely to cause the voicing asymmetry in natural language. Secondly, an alternation asymmetry in the same direction as in Lysvik's (2020) Experiment 1 was found in the current experiment, although this asymmetry was small and did not arise from a strategy to make paradigms more uniform. Thirdly, this experiment is the first ALL experiment making use of the iterated learning paradigm in phonology. It therefore provides valuable experiences to the further study of the voicing asymmetry and the use of this paradigm in future research.

The findings in the current experiment, then, indicate that if there is a substantive bias impeding the learning of final voicing, it is most likely to be found in the alternation between voiceless intervocalic stops and voiced final stops. Because it is not possible to make strong claims based on the descriptive analysis presented here, the alternation asymmetry is worth further investigation. Nevertheless, the asymmetry turned out to be small when different learning strategies were accounted for, and the results thus provide little support for the substantive bias account. The most promising way forward in the investigation of the voicing asymmetry is therefore to explicitly investigate effects of channel bias.

Finally, although the iterated learning design did not show an effect of transmission in this experiment, it can prove useful in future research on the voicing asymmetry. In experiments probing for effects of substantive bias, the scope of the experiment can be expanded to include more input to each generation, more diffusion chains, and more generations. Furthermore, the iterated learning design can also be used to investigate effects of channel bias, for example by including a production task. If the final devoicing and final voicing processes develop differently in such an experiment, which was not the case in the current experiment, the channel bias account will be strengthened. Such an experiment would be practically more challenging, but due to the scarcity of experiments finding substantive bias effects, it could provide important insights into which bias underlies the asymmetry between final devoicing and final voicing.

References

Avcu, E., Rhodes, R., & Hestvik, A. (2019, May 7-11). *Neural tracking of implicit vs. explicit phonotactic learning* [Conference presentation]. GLOW42: Workshop on Rules and Learning Strategies in the Acquisition of Signed and Spoken Phonologies, Oslo, Norway.

Baer-Henney, D. (2019, May 7-11). *Phonetics and L1 as learners' little helpers* [Conference presentation]. GLOW2019, Workshop on Rules and Learning Strategies in the Acquisition of Signed and Spoken Phonologies, Oslo, Norway.

Baker, C. L. (1979). Syntactic Theory and the Projection Problem. *Linguistic Inquiry, 10*(4), pp. 533-581.

Bates, D., & Sarkar, D. (2016). *Lme4: Linear mixed-effects modeling using S4 classes.* R package, version 2.15.3.

Beguš, G. (2018). *Unnatural Phonology: A Synchrony-Diachrony Interface Approach* [Doctoral dissertation]. Harvard University, MA.

Bermúdez-Otero, R., & Börjars, K. (2006). Markedness in phonology and in syntax: the problem of grounding. *Lingua, 116*, pp. 710–756.

Blevins, J. (2004). *Evolutionary Phonology: The emergence of sound patterns.* Cambridge: Cambridge University Press.

Blevins, J. (2006). A Theoretical Synopsis of Evolutionary Phonology. *Theoretical Linguistics, 32*(2), pp. 117-166.

Carpenter, A. C. (2010). A naturalness bias in learning stress. *Phonology, 27*(3), pp. 345-392.

Culbertson, J. (2015, June 22 - 26). *Syntactic universals in the lab: New methods and approaches* [Conference presentation]. Advanced Core Training in Linguistics (ACTL), University College London, UK.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition, 122*(3), pp. 306–329.

De Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics, 28*(4), pp. 441-465.

Do, Y. (2013). *Biased learning of phonological alternations* [Doctoral dissertation]. Massachusetts Institute of Technology, MA.

Do, Y. (2018). Paradigm uniformity bias in the learning of Korean verbal inflections. *Phonology, 35*(4), pp. 547-575.

Do, Y., Zsiga, E., & Havenhill, J. (2016, Jan. 7-10). *Naturalness and frequency in implicit phonological learning* [Conference presentation]. 90th Annual Meeting of the Linguistic Society of America, Washington, D.C.

Ettlinger, M., Morgan-Short, K., Faretta-Stutenberg, M., & Wong, P. C. (2015). The relationship between artificial and second language learning. *Cognitive Science, 40*(4), pp. 822–847.

Fedzechkina, M., Newport, E., & Jaeger, T. F. (2016). Miniature artificial language learning as a complement to typological data. In L. Ortega, A. E. Tyler, H. I. Park, & M. Uno (Eds.), *The usage-based study of language learning and multilingualism* (pp. 211–232). Washington, DC: Georgetown University Press.

Finley, S. (2012). Typological asymmetries in round vowel harmony: Support from Artificial Grammar Learning. *Language and Cognitive Processes, 27*(10), pp. 1550–1562.

Glewwe, E. (2019). *Bias in Phonotactic Learning: Experimental Studies of Phonotactic Implicationals* [Doctoral dissertation]. University of California, Los Angeles.

Glewwe, E., Zymet, J., Adams, J., Jacobson, R., Yates, A., Zeng, A., & Daland, R. (2018, Jan. 4-7). *Substantive bias and the acquisition of final (de)voicing patterns* [Conference presentation]. 92nd Annual Meeting of the Linguistic Society of America (LSA), Salt Lake City, UT.

Greenwood, A. (2016). *An Experimental Investigation of Phonetic Naturalness* [Doctoral dissertation]. University of California, Santa Cruz.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using Category Structures to Test Iterated Learning as a Method for Identifying Inductive Biases. *Cognitive Science, 32*(1), pp. 68-107.

Grijzenhout, J., & Krämer, M. (2000). Final devoicing and voicing assimilation in Dutch derivation and cliticization. In B. Stiebels, & D. Wunderlich (Eds.), *Lexicon in focus* (pp. 55–82). Berlin: Akademie Verlag.

Harris, J. (2009). Why final obstruent devoicing is weakening. In K. Nasukawa, & P. Backley (Eds.), *Strength relations in phonology* (pp. 9–46). Berlin: Mouton de Gruyter.

Hayes, B., & White, J. (2013). Phonological Naturalness and Phonotactic Learning. *Linguistic Inquiry, 44*(1), pp. 45-75.

Hayes, B., Kirchner, R., & Steriade, D. (Eds.). (2004). *Phonetically-Based Phonology.* Cambridge: Cambridge University Press.

Kiparsky, P. (1995). The phonological basis of sound change. In J. A. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 640–670). Cambridge, MA; Oxford, UK: Blackwell.

Kiparsky, P. (2006). The Amphichronic Program vs. Evolutionary Phonology. *Theoretical Linguistics, 32*(2), pp. 217-236.

Kiparsky, P. (2008). Universals constrain change; change results in typological generalizations. In J. Good (Ed.), *Linguistic universals and language change* (pp. 23-53). Oxford: Oxford University Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences - PNAS, 105*(31), pp. 10681-10686.

Kopkalli, H. (1993). *A phonetic and phonological analysis of final devoicing in Turkish* [Doctoral dissertation]. University of Michigan.

Kristoffersen, G. (2000). *The phonology of norwegian (The Phonology of the world's languages).* Oxford: Oxford University Press.

Lieberman, E., Nowak, M. A., Jackson, J., Michel, J.-B., & Tang, T. (2007). Quantifying the evolutionary dynamics of language. *Nature, 449*(7163), pp. 713-716.

Lysvik, J. K. (2020). *Where does naturalness in phonology come from? – Insights from Artificial Language Learning* [Doctoral dissertation]. University of Oslo.

Moreton, E. (1999). *Non-Computable Functions in Optimality Theory.* ROA-364. Rutgers Optimality Archive. http://roa.rutgers.edu/.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology, 25*(1), pp. 83-127.

Moreton, E., & Pater, J. (2012a). Structure and Substance in Artificial-phonology Learning, Part I: Structure. *Language and Linguistics Compass, 6*(11), pp. 686-701.

Moreton, E., & Pater, J. (2012b). Structure and Substance in Artificial-Phonology Learning, Part II: Substance. *Language and Linguistics Compass, 6*(11), pp. 702-718.

Moreton, E., Pater, J., & Pertsova, K. (2017). Phonological Concept Learning. *Cognitive Science, 41*(1), pp. 4-69.

Myers, S., & Padgett, J. (2014). Domain generalization in artificial language learning. *Phonology, 31*(3), pp. 399–433.

Ohala, J. J. (1981). The listener as a source of sound change. In C. S. Masik, R. A. Hendrick, & M. F. Miller (Eds.), *Papers from the Parasession on Language and Behavior* (pp. 178–203). Chicago: Chicago Linguistic Society.

Ohala, J. J. (1993). The phonetics of sound change. In C. Jones (Ed.), *Historical linguistics: Problems and perspectives* (pp. 237–278). London: Longman Academic.

Peperkamp, S., & Dupoux, E. (2007). Learning the mapping from surface to underlying representations in an artificial language. In J. Cole, & J. Hualde (Eds.), *Laboratory Phonology 9* (pp. 315–338). Berlin: Mouton de Gruyter.

Prince, A., & Smolensky, P. (1993/2004). *Optimality Theory: Constraint interaction in generative grammar.* Malden, MA; Oxford, UK: Blackwell. [Published in 2004. First circulated in 1993].

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition, 111*(3), pp. 317-328.

Sanders, L., Moore-Cantwell, C., Pater, J., Staubs, R., & Zobel, B. (2015). *Event-related potential evidence of abstract phonological learning in the laboratory* [Manuscript]. University Massachsettes, Amherst.

Shapp, A. (2012). *Substantive bias in the learning of harmony patterns* [Qualifying Paper]. New York University.

Skoruppa, K., & Peperkamp, S. (2011). Adaptation to novel accents: Feature-based learning of context-sensitive phonological regularities. *Cognitive Science, 35*(2), pp. 348–366.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition, 116*(3), pp. 444-449.

Stanton, J. (2017). *Constraints on the Distribution of Nasal-Stop Sequences: An Argument for Contrast* [Doctoral dissertation]. Massachusetts Institute of Technology, MA.

Wedel, A. (2012). Lexical contrast maintenance and the organization of sublexical contrast systems. *Language and Cognition, 4*(4), pp. 319-355.

Westbury, J. R., & Keating, P. A. (1986). On the naturalness of stop consonant voicing. *Journal of Linguistics, 22*(1), pp. 145-166.

Wilson, C. (2003). Experimental Investigation of Phonological Naturalness. In G. Garding, & M. Tsujimura (Eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics (WCCFL 22)* (pp. 533–546). Somerville, MA: Cascadilla Press.

Winter, B. (2020). *Statistics for Linguists: An Introduction Using R.* New York, New york: Routledge.

Yu, A. C. (2004). Explaining final obstruent voicing in Lezgian: Phonetics and history. *Language*, *80*(1), pp. 73–97.

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. Retrieved from https://doi.org/10.17605/OSF.IO/MD832.

Appendix

Table A.1 Percentages for all four possible item types in Chain 1a (VOICING).

Participant	Alt (conforming)	Non-alt (conforming)	Alt (non-conforming)	Non-alt (non-conforming)
1.1a	8%	54%	0%	38%
1.2a	15%	44%	8%	33%
1.3a	18%	44%	3%	36%
1.4a	5%	60%	0%	36%
1.5a	13%	38%	8%	41%

Table A.2 Percentages for all four possible item types in Chain 1b (VOICING).

Participant	Alt (conforming)	Non-alt (conforming)	Alt (non-conforming)	Non-alt (non-conforming)
1.1b	62%	33%	3%	3%
1.2b	51%	15%	13%	21%
1.3b	28%	38%	5%	28%
1.4b	59%	15%	10%	15%
1.5b	26%	26%	18%	31%

Table A.3 Percentages for all four possible item types in Chain 1c (VOICING).

Participant	Alt (conforming)	Non-alt (conforming)	Alt (non-conforming)	Non-alt (non-conforming)
1.1c	79%	18%	0%	3%
1.2c	41%	33%	5%	21%
1.3c	33%	31%	10%	26%
1.4c	31%	31%	26%	13%
1.5c	28%	10%	33%	28%

Table A.4 Percentages for all four possible item types in Chain 1d (VOICING).

Participant	**Alt (conforming)**	**Non-alt (conforming)**	**Alt (non-conforming)**	**Non-alt (non-conforming)**
1.1d	33%	41%	8%	18%
1.2d	46%	33%	3%	18%
1.3d	44%	36%	3%	18%
1.4d	64%	21%	0%	15%
1.5d	44%	23%	21%	13%

Table A.5 Percentages for all four possible item types in Chain 1e (VOICING).

Participant	**Alt (conforming)**	**Non-alt (conforming)**	**Alt (non-conforming)**	**Non-alt (non-conforming)**
1.1e	31%	26%	10%	33%
1.2e	46%	15%	21%	18%
1.3e	23%	31%	13%	33%
1.4e	18%	41%	21%	21%
1.5e	23%	33%	15%	28%

Table A.6 Percentages for all four possible item types in Chain 2a (DEVOICING).

Participant	**Alt (conforming)**	**Non-alt (conforming)**	**Alt (non-conforming)**	**Non-alt (non-conforming)**
2.1a	74%	23%	0%	3%
2.2a	74%	23%	3%	0%
2.3a	23%	33%	3%	41%
2.4a	31%	38%	3%	28%
2.5a	33%	28%	15%	23%

Table A.7 Percentages for all four possible item types in Chain 2b (DEVOICING).

Participant	**Alt (conforming)**	**Non-alt (conforming)**	**Alt (non-conforming)**	**Non-alt (non-conforming)**
2.1b	67%	26%	5%	3%
2.2b	72%	8%	21%	0%
2.3b	69%	0%	31%	0%
2.4b	28%	18%	5%	49%
2.5b	15%	15%	15%	54%

Table A.8 Percentages for all four possible item types in Chain 2c (DEVOICING).

Participant	**Alt (conforming)**	**Non-alt (conforming)**	**Alt (non-conforming)**	**Non-alt (non-conforming)**
2.1c	44%	15%	13%	28%
2.2c	62%	13%	18%	8%
2.3c	72%	8%	5%	15%
2.4c	72%	5%	15%	8%
2.5c	62%	8%	8%	23%

Table A. 9 Percentages for all four possible item types in Chain 2d (DEVOICING).

Participant	**Alt (conforming)**	**Non-alt (conforming)**	**Alt (non-conforming)**	**Non-alt (non-conforming)**
2.1d	38%	21%	10%	31%
2.2d	21%	26%	21%	33%
2.3d	41%	10%	26%	23%
2.4d	18%	23%	28%	31%
2.5d	44%	10%	31%	15%

Table A.10 Percentages for all four possible item types in Chain 2e (DEVOICING).

Participant	**Alt (conforming)**	**Non-alt (conforming)**	**Alt (non-conforming)**	**Non-alt (non-conforming)**
2.1e	49%	33%	8%	10%
2.2e	74%	3%	21%	3%
2.3e	77%	3%	21%	0%
2.4e	79%	0%	21%	0%
2.5e	72%	8%	18%	3%