TEL AVIV אוניברסיטת
UNIVERSITY תל אביב

The Lester and Sally Entin Faculty of Humanities

The School of Philosophy, Linguistics and Science Studies

Department of Linguistics

Evolution of Phonological Typology: an Iterated Learning Model of the Emergence of Phonological Patterns

MA thesis submitted by

Itamar Shefi

Thesis advisors:

Dr. Roni Katzir

Dr. Ezer Rasin

December 2020

Abstract

Phonological typology is highly skewed. For example, while final devoicing of obstruents occurs in many languages, Lezgian is the single documented case that has been argued to possess the opposite pattern of final voicing (Yu, 2004). One approach to explaining this kind of asymmetry is attributing it to an *analytic bias*, cognitive biases which ease the learning of some phonological patterns over others, as done in classical Optimality Theory (OT; Prince and Smolensky, 1993). Another approach is that the source for the asymmetry is *channel bias* – recurring systematic errors which push languages towards one pattern and away from its opposite pattern, as Evolutionary Phonology (EP; Blevins, 2004) suggests.

The division of labor between analytic and channel bias is an empirical question. In my work I present a model for channel bias which can help us to reason about this division of labor by examining whether typological asymmetries in phonology can emerge as a result of transmission of phonological knowledge between generations, using the asymmetry between final voicing and devoicing as a case study. My model, which builds on the Iterated Learning Model of language transmission (ILM; Kirby, 2001, 2002), includes corruption of the data by applying noise which models the channel bias described by Blevins (2004). The learning agent in my model is Rasin et al.'s (2018) Minimum Description Length (MDL; Rissanen, 1978) learner which I modified so it can handle some amounts of such noise. I show how this model succeeds in simulating the emergence of final (de-)voicing asymmetry from a (de-)voicing neutral starting point. I also show the model can simulate a decay of final voicing which weakens Kiparsky's (2006) response to EP that final voicing is expected to be a more common sound pattern if there were no innate constraints against it. The success of the model to simulate the emergence of this phonological asymmetry opens the door to theories that attribute less of the typology to analytic bias and more of it to channel bias.

Acknowledgments

First and foremost, I would like to thank my advisors, Roni Katzir and Ezer Rasin. Their breadth and depth of knowledge and invaluable guidance and insights allowed this work to happen. Their contribution was critical for completing this project. During my last year as an undergraduate student, Roni's fascinating Learning seminar motivated my passion for language evolution. Roni and Ezer's guidance allowed me to pursue my interest in asking meaningful questions about the human mind and the forces that shape languages.

I would also like to express my gratitude to Evan Cohen. This work, and particularly the phonological aspects of it, benefited a great deal from his insightful comments.

This project could not have happened without Nur Lan and Iddo Berger. This is true in a very literal sense, as Iddo implemented the learner's algorithms and Nur improved the search process significantly, allowing the heavy simulations I needed for this work to run. But beyond that, Nur's guidance through the code and help in debugging errors, sometimes even late on Friday nights, enhanced my progress tremendously.

I extend my gratitude to Taly Rabinerson and Iddo Yadlin, who challenged me with great questions about my research and happily provided feedback about the writing style. Finishing this work without their academic and emotional support would have been considerably harder.

I wish to thank my teachers in the linguistics department at Tel Aviv University:

Aya Meltzer-Asscher, Tal Siloni, Fred Landman, Outi Bat-El, Mira Ariel, Nirit Kadmon and Julia Horvath; members of the Tel Aviv University Computational Linguistics Lab; and wonderful Ruti Zussman.

Finally I would like to thank my parents, Dina and Ilan – I could not have asked for better parents than you; and my friends and loved ones who were there when I needed them. I have no idea how I could done this without your support.

Contents

Chapter 1

Phonological typology

The study of typology is one of the most important tools that the linguist might have: a linguistic theory that cannot account for typological generalizations is incomplete. Some theories attribute key aspects of the typology to *analytic bias*, claiming that certain features are impossible to represent or harder to learn than others, while according to other theories the source of certain typological asymmetries is *channel bias*, systematic errors in transmission which directs the cultural evolution of languages. As Moreton (2008) states, analytic bias and channel bias should not be treated as mutually exclusive. The question is what is the division of labor between the two. This chapter introduces the descriptive and theoretical threads of the study of phonological typology and the ways in which typological evidence can be used in order to evaluate phonological theories. In section 1.2 I provide two examples of theories which deal with typological generalizations differently, by relying on either analytic bias or channel bias, and present how these theories deal with these generalizations. In section 1.3 I provide examples for such generalizations and explain how these theories account for them.

Ultimately the division of labor between the innate and external factors is an empirical question. In chapter 2 I present a model for transmission of phonological knowledge between generations which allows us to look into this question. I also discuss criteria for learning, and present a proposal for learning phonology in a noisy environment. In chapter 3 I present simulations I ran using this model which can help us reason about the division of labor between channel bias and analytic bias in shaping the final (de-)voicing typology.

1.1 Introduction

Languages differ from one another in their structural properties. While word order in Italian is subject-verb-object (SVO), in Japanese it is subject-object-verb (SOV), and in fact each one of the six possible orders is witnessed in some language.[1] The study of typology groups languages according to their structural properties and surveys the variation and distribution of these properties. If we take the example above, we can classify Italian, Modern Hebrew and English as SVO languages[2], while Japanese, Kannada and Amharic are classified as SOV languages.

There seem to be properties that are universal to all languages: for example, there is no known language that does not have the property "has verbs", while some properties are not universal but appear in many unrelated languages, like the property of *final devoicing*: "an obstruent in a final position is devoiced". On the other hand, there are properties that are not attested in any language: according to Horn (1972) no known language possesses logical entities corresponding to NAND (=not and) nor NALL (=not all). In addition, the property *final voicing* "an obstruent in a final position is voiced" which, as its name suggests, is symmetric to *final devoicing* is arguably unattested. Lezgian is the single documented case that has been argued to possess the

[1] Out of the possible six orders SOV, SVO, and VSO are much more common than the others (Greenberg, 1963), and it is not clear if each of the orders can appear on surface without syntactic derivation. For example, Coon (2010) argues that VOS order in Chol Mayan is derived by a fronting of the predicate, and according to Polinskaja (1989) in some OSV cases the word order is a result of object fronting.

[2] McCawley (1970) argues that English is actually underlyingly VSO. Either way, the point here is that we can categorize languages by their word order.

pattern (Yu, 2004), and even in Lezgian the phenomenon is extremely restricted both phonologically and morphologically and cannot be considered a mirror image of final devoicing.

This kind of consistency can be used to evaluate linguistic theories: a theory that cannot account for typological asymmetries is incomplete. Some theories attribute this kind of asymmetries to analytic bias which allows certain properties while constraining others. For example, Kiparsky (2006, p. 2) explains the final (de-)voicing typological asymmetry under the framework of Optimality Theory (OT; Prince and Smolensky, 1993) by "*the existence of constraints that prohibit marked features in weak positions, and the absence of constraints that prohibit unmarked features in them*". Other theories attribute some of the asymmetries to channel bias, suggesting these asymmetries are a result of systematic errors in language transmission. For example, Evolutionary Phonology (EP; Blevins, 2004, 2006) theory suggests that some recurrent phonological patterns might be a result of a common *sound change* – diachronic events which change the sound system of a given language over time. Regarding the final (de-)voicing asymmetry, Blevins proposes that phonetic factors such as phrase-final lengthening of stops and absence of audible release phrase-finally, word-finally, and before certain consonants are some of the sources for channel bias which derives the emergence of final devoicing. The lack of phonetic factors which support final voicing is the reason that the pattern is not attested according to Blevins.

It is plausible to attribute certain universal phenomena to analytic bias, like in the NAND, NALL example above. Certain phonological universals, on the other hand, might be a result of channel bias. As Moreton (2008) showed, it is possible to test empirically if a specific linguistic feature should be attributed to analytic bias. In a series of experiments he showed that humans are better at learning the dependencies between the heights in two adjacent syllables than learning the dependencies between the height of a vowel and the voice features of the following consonant, or the dependencies between the voice features of consonants in two neighboring syllables, suggesting that the analytic bias is the source of the common phonological pattern of vowel harmony. If we can show how certain universals emerge by using an evolutionary model such as Blevins's, it will allow us to consider the possibility to attribute these universals to channel bias.

Phonological typology, which this work focuses on, can be divided into two interdependent studies. The first one is descriptive: studying the distribution and frequency of sound patterns in human languages as done, for example, by Trubetzkoy (1939) who noticed that all languages have voiceless obstruents but not all languages have voiced obstruents, or Jakobson (1962) who noticed the universal asymmetry between syllable structures – while all languages allow syllables that start with CV and syllables that end with V, some languages do not allow syllables that start with V or syllables that end with C. The other area of phonological typology relates to examining how different theories explain the typological observations as done in generative phonology since its inception: Chomsky and Halle (1968, p. 19, footnote 5) provide typological arguments for serial rule application, showing it can predict that certain languages cannot exist; Anttila and Magri (2018) argue against Maximum Entropy (ME; Goldwater and Johnson, 2003) theory as it predicts the existence of unattested phonological patterns; and Gordon (2016) provides in his book different types of explanations for phonological typology, ranging from phonetic factors to analytic bias and explains how these factors can be incorporated within phonological theories. Gordon shows how the different theories deal with typological observations related to the structure of phoneme inventories, syllable structure, phonological processes, stress, and additional patterns while surveying patterns' distribution, variation, and frequency.

1.2 Typology and theoretical phonology

In this section I present the different approaches to typological explanation by surveying two phonological theories: Standard OT which attributes much of the typology to analytic bias in the shape of innate phonological constraints, and EP which argues that channel bias accounts for certain asymmetries. Concrete examples for how these theories account for typological generalizations will be provided in section 1.3.

1.2.1 Optimality Theory

Unlike Chomsky and Halle's Sound Pattern of English (SPE; 1968) which proposes a rule-based system, Standard OT assumes there is a set of universal constraints. The ranking between these constraints determines the phonological grammar of the language, and the sound patterns of the language are formed by the interaction between the constraints. OT consists of three components: (i) a Generator which takes the lexical input and produces candidates for the final output; (ii) a Constraint component which includes a set of the universal constrains mentioned above; and (iii) an Evaluator which chooses the optimal candidate according the ranking of the constraints. The ranking of the constraints defines the phonological grammar: if a candidate c_1 violates a constraint r_1 that is not violated by another candidate c_2 then c_2 is preferred over c_1, given c_2 does not violate any constraint r_2 which is ranked higher than r_1 and is not violated by c_1. A simple example is given in the tableau in (1) which shows how the Evaluator chooses the output [sap] from the three candidates generated by the Geneator for the lexicon input /stap/. The Constraints in this example are Max ("no deletion"), Dep ("no epenthesis") and *CC ("no consonant clusters") and they are ranked *CC >> Dep >> Max.

(1)

/stap/	*CC	Dep	Max
a. stap	*!		
☞ b. sap			*
c. satap		*!	

Unless certain constraints are universally ranked, all rankings should be possible in some language. The fact that the set of constraints is universal limits the typologies to the number of possible constraint rankings. Moreover, as we will see in section 1.3.1.1, some subsets of constraints do not interact with other subsets of constraints and it does not matter if a constraint of the first subset is ranked above or below a constraint of the other subset, so some re-rankings of constraints lead to identical patterns. Namely, *n*! different rankings of *n* constraints do not necessarily yield *n*! different typologies. The specification of the constraints limits the typology as well. Consider for example the fact that all languages possess the CV syllable structure (Jakobson, 1962) while only few languages have tautosyllabic VVVV (Gilbertese) or CCCCCC (Georgian) sequences (Blevins, 2004, p. 213). Take a look at OT's markedness constraint and its hypothetical symmetric constraint in (2):

(2) **Markedness constraints**

a. *Complex (Prince and Smolensky, 1993, p. 96)

No more than one C or V may associate to any syllable position node.

b. *Simplex

Any syllable position node must associate with more than one C or V.

*Complex (2a) supports the pattern "syllables with the structure CV, V, CVC, VC etc." while it blocks the pattern "syllables with the structures CCVVCC, VCVCVC, CCCCCC etc.", while *Simplex (2b) supports the second pattern while blocking the first. However, within OT a constraint like *Simplex is simply not viable, as markedness constraints are based on universal markedness observations. To generalize, let us assume two symmetric patterns, p and p^{-1}. The constraint C_p supports the pattern p (something like "the pattern p^{-1} is not allowed") and the constraint $C_{p^{-1}}$ supports the pattern p^{-1} ("the pattern p is not allowed"). If according to typological observations p occurs in human languages while there is no evidence for the existence of p^{-1}, it is easy to claim that the constraint $C_{p^{-1}}$ simply does not exist because p^{-1} is a marked pattern while p is an unmarked pattern. This is how OT explains final (de-)voicing as we shall see in section 1.3.2.1.

1.2.2 Evolutionary Phonology

Standard OT is an example for a theory which attributes typological asymmetries to analytic bias. EP, on the other hand, is an example for a theory that attributes certain asymmetries to channel bias. EP examines recurring phonological patterns from a diachronic perspective and its explanation for typological generalizations is that different instances of a frequent pattern stem from a common sound change between generations which according to Blevins (2006) falls into one of the following types:

(3) **Typology of sound change**

a. CHANGE: The utterance perceived by the listener is different from the utterance produced by the speaker due to acoustic similarities or some innate perceptual bias.

For example: speaker produces [binba] consistently, listener perceives [bimba]

b. CHANCE: The utterance perceived by the listener can be parsed in multiple ways, and the listener parses it differently form the underlying form in the grammar of the speaker.

For example: speaker produces [ʔaʔ] consistently meaning /aʔ/, listener perceives [ʔaʔ] and interprets as /ʔa/

c. CHOICE: A single speaker might produce different variants of the same phonological form (for example, due to change in speech rate or other types of noise). The listener perceives those forms correctly but chooses a different variant from the variant in the speaker's grammar.

For example: speaker produces [kakˈata], [kăkˈata], [kkˈata] inconsistently for /kakata/, listener perceives all of them correctly but interprets as /kkata/

Blevins compares phonological evolution to biological evolution and states that the sound patterns of two languages may be similar due to: (i) their sharing of the same common ancestor-language; (ii) a similar common sound change that affected both languages; (iii) physical constraints on form and function and phonological universals; (iv) external factors such as language contact; (v) mere chance. Table 1, taken from Blevins (2006), provides examples for these factors and their analog in biological evolution.

Source of Similarity	BIOLOGICAL	LINGUISTIC
a. Direct genetic inheritance	Shared genetic traits of identical twins, e.g. eye color	Shared inherited features of British and Australian English, e.g. r-loss
b. Adaptation by natural selection	Independent development of toepads in *Iguanidae*, *Scincidae*, and *Gekkonidae*	Independent development of final obstruent devoicing in Indo-European, Turkic, Cushitic, etc.
c. Physical constraints on form and function	Patterns of spots and stripes on cats and seashells, as determined by chemistry/-physics	Universal gross category boundaries for consonant types, as determined by categorical perception
d. "Non-natural" or external factors	Grafting, hybridization, genetic modification	Language contact/diffusion, prescriptive norms, literacy and second language learning
e. Chance	Arctic hares and albino rabbits have white coats, but...	Japanese and Gilbertese only allow nasal Cs word-finally, but...

Table 1: **General sources of similarity** of recurrent sound patterns and biological characteristics, with a biological and a linguistic example for each one of the sources. From Blevins (2006, p. 121).

1.3 Examples of typological asymmetries in phonology

In order to illustrate the differences between analytic bias and channel bias typological explanation, this section shows how the phonological theories surveyed in 1.2 deal with two phonological patterns: Jakobsonian syllable structure and the asymmetry of final (de-)voicing.

1.3.1 Jakobsonian syllable structure

The different ways that words are divided into syllables seem to be limited. As mentioned above, only few languages have tautosyllabic VVVV (Gilbertese) or CCCCCC (Georgian) sequences (Blevins, 2004, p. 213). Here I focus on a much simpler typological observation, described by Jakobson (1962):

(4) a. Syllables that end with C are not allowed in some languages.

b. Syllables that start with V are not allowed in some languages.

c. Syllables that start with CV and syllables that end with V are allowed in every language.

Since (4c) is an absolute, there is a trivial implicational universal here: every language that has either of the properties “Allow the syllable structure described in (4a)” or “Allow the syllable structure described in (4b)” also has the property “Allow the syllable structures described in (4c)”. A phonological theory that fails to describe grammars that obey this implicational universal or can generate grammars that do not obey it is not complete. Let us see how the theories mentioned above deal with the Jakobsonian syllable structure.

1.3.1.1 OT and Jakobsonian syllable structure

Prince and Smolensky (1993) show that OT can account for the Jakobsonian syllable structure using the five constraints in (5).

(5) a. **Onset**

A syllable must have an onset.

b. ***Coda**

A syllable must **not** have a coda.

c. **Parse**

Underlying segments must be parsed into syllable structure.

d. **Fill$^{\text{Nuc}}$**

Nucleus positions must be filled with underlying segments.

e. **Fill$^{\text{Ons}}$**

Onset positions must be filled with underlying segments.

Five constraints can be ranked in 5! = 120 different ways. However, this does not lead to 120 different syllable structures: first, because onset distribution is not affected by the *Coda, Fill$^{\text{Nuc}}$ constraints and coda distribution is not affected by the Onset, Fill$^{\text{Ons}}$ constraints; and second, some rankings lead to the same patterns as can be seen in Table 2, which is a simplified version of the table in Prince and Smolensky (1993, p. 104). We can see that in this case any ranking of the five constraints is mapped to one of nine possible constraint interactions, which in turn yield only four different possible syllable structures.

To summarize, no matter how we order these constraints they will not support the existence of a language without syllables that start with CV and without syllables that end with V, even though these constraints do not say anything directly about these structures. In addition, the number of syllable structures that this set of constraints can create is small relative to the size of the set. These limitations lead to all possible patterns observed by Jakobson on the one hand, and no impossible patterns on the other hand.

			Onsets		
			required		not required
			Onset, Fill^{Ons} >> Parse	Onset, Parse >> Fill^{Ons}	Parse, Fill^{Ons} >> Onset
Codas	forbidden	*Coda, Fill^{Nuc} >> Parse	CV	CV	(C)V
		*Coda, Parse >> Fill^{Nuc}	CV	CV	(C)V
	allowed	Parse, Fill^{Nuc} >> *Coda	CV(C)	CV(C)	(C)V(C)

Table 2: **Syllable structure typology**. Each of the 120 possible rankings of the constraints in (5) will fall into one of nine cells in the table, which contains four possible syllable structures: CV, (C)V, CV(C) and (C)V(C).

1.3.1.2 EP and Jakobsonian syllable structure

We saw how analytic bias can account for the Jakonsonian syllable structure by examining Standard OT's explanation, which captures the asymmetry of how human languages treat onsets and codas: favoring the first while trying to avoid the second. Let us see a channel bias explanation for the asymmetry which is composed from multiple different arguments. First, EP attempts to explain the pattern by stating that a preference for a certain structure within a language might be a result of the tendency to Structural Analogy, as described in (6) (Blevins, 2004, p. 154).

(6) Structural Analogy
In the course of language acquisition, the existence of a phonological contrast between A and B will result in more instances of sound change involving shifts of ambiguous elements to A or B if no contrast between A and B existed.

Blevins provides the following example for why the learner might prefer the CV syllable structure:

"Now imagine that the language at large has many unambiguous CV syl-

> *lables, but very few closed CVC syllables, which have been segmented by the learner. Structural Analogy allows the lexical dominance of open vs. closed syllables to play a role in the categorization of the ambiguous string. In this case, the result would be a higher probability of a /ʔa/ parse than a /aʔ/ parse."* (Blevins, 2006, p. 128)

Structural Analogy can explain how a relatively large number of syllable structures might collapse into few syllable structures over the course of generations of language transmission. However, we can replace CV with VC in Blevins's example which will result in a language which does not match the Jakobsonian syllable structure, hence some other explanation is needed for the reason that CV is the preferred syllable structure. This means that EP's explanation in fact results in typological symmetry between CV and VC distribution, and something to break that symmetry is required. Easterday (2019) surveys several explanations for preferring the CV syllable structure: first, according to Kawasaki-Fukumori (1992), it is easier to distinguish CVs from one another than VCs as they are more spectrally dissimilar. Second, Content et al. (2001) show that listeners identify onsets better than they identify codas, and third, Segui et al. (1991) show that the processing time of CVCs is longer than the processing time of CVs. In addition to the perceptual explanations, Easterday provides physiological explanations: for example, the fact that the timing between the production of the onset and the nucleus is relatively stable as they are produced almost simultaneously, while the timing between the production of the nucleus and the coda is more variable, making codas less stable (Byrd, 1996; Browman and Goldstein, 1995; Gick et al., 2006; Marin and Pouplier, 2010). All of these mean that CVs are more likely to survive language transmission between generations than VCs, both because CVs are more likely to be perceived correctly by the learner, and also because they are more stable than VCs. This is enough to break the symmetry between the two, allowing the tendency to Structural Analogy (6) to serve as a possible source of diachronic pressure which supports the emergence of the CV syllable structure.

1.3.2 Final (de-)voicing

We saw examples for how analytic bias and channel bias can account for the Jakobsonian syllable structure. This section discusses another typological generalization, the asymmetry between final voicing and final devoicing, and shows how the two approaches deal with this asymmetry.

Final devoicing of obstruents is a pattern that appears in many unrleated human languages: Russian, Ingush, Chadic Arabic (Halle, 1959; Zeltner and Tourneaux, 1986; Guerin, 2001, respectively; in Blevins, 2006) and German (Kenstowicz and Kisseberth, 1979) are only some of them. The Russian data in Table 3 taken from Kenstowicz and Kisseberth (1979, p. 49) illustrates the pattern: we see that obstruents in final position are always voiceless: /p t s ʃ k/. We can tell that it is not the case that an obstruent is voiced when contacted with a vowel in morpheme boundaries because we have examples such as [sobak-a] and [sobak-e] vs. [sobak], all have the devoiced obstruent k as stem's final segment. Within rule-based phonology, the pattern can be stated as follows:

(7) **Final devoicing**

$[-\textit{son}] \rightarrow [-\textit{voice}]\ /\_\_\#$

The symmetric pattern, final voicing (8) is arguably unattested. Lezgian (Yu, 2004) is the only documented case that has been argued to possess final voicing, and it is limited: it occurs in some cases to obstruents in coda position and word-finally only in certain monosyllabic nouns. Kiparsky (2006, 2008) offers an alternative analysis for the Lezgian pattern which does not include a process of final voicing.

(8) **Final voicing**

$[-\textit{son}] \rightarrow [+\textit{voice}]\ /\_\_\#$

Nom. sg.	Dat. sg.	Gen. pl.	Gloss
ryba	rybe	ryp	‘fish‘
tropa	trope	trop	‘path‘
pobeda	pobede	pobet	‘victory‘
sirota	sirote	sirot	‘orphan‘
groza	groze	gros	‘storm‘
kryʃa	kryʃe	kryʃ	‘rat‘
lyʒa	lyʒe	lyʃ	‘ski‘
dusa	duse	dus	‘soul‘
noga	noge	nok	‘leg‘
sobaka	sobake	sobak	‘dog‘

Table 3: Final devoicing in Russian

1.3.2.1 OT and final (de-)voicing

Let us examine an explanation to the (de-)voicing asymmetry which relies on analytic bias. Constraints in OT either demand for the output to preserve the input form, “faithfulness constraints”, or for the well-formedness of the output, “markedness constraints”. Kiparsky (2006) uses a markedness constraint of the form “a marked feature value is not allowed” in order to explain final devoicing. Assuming that the [+*voice*] feature value is marked in coda position, we can define a constraint such as *VOICED-CODA which leads to coda devoicing. Since final obstruents are a subset of the word’s codas, a modification of this constraint or an additional constraint that prohibits the change of feature values word-internally leads to a final devoicing pattern. This explanation relies on the assumption that the pattern of final (de-)voicing is *universal*: the existence of final voicing is ruled out by the fact that there are no constraints that do not allow the unmarked feature value [−*voice*] in the coda position. There are constraints that do not allow [−*voice*] in more general positions, for example the constraint *P[−*voice*] which prohibits voiceless bilabial stop in Algerian Spoken Arabic (Kessar and Mahadin, 2020). But even with constraints like *P[−*voice*], the existence of *VOICEDCODA and the lack of its opposite constraint means that no constraint ranking can result in final voicing. This is illustrated in the tableau in (9) which includes the two markedness constraints *P[−*voice*], *VoicedCoda and the faithfulness constraint Ident(voice) which requires that the value of the [*voice*] feature of every segment to be the same in the input and the output:

(9)

/pop/	*P[−*voice*]	Ident(voice)	*VoicedCoda
a. pop	**		
b. bob		**	*
c. pob	*	*	*

If we rank *P[−*voice*] above the two other constraints, candidate (9b) wins and we get voicing *everywhere*. But no matter how we rank the three constraints, candidate (9c) never wins meaning that final voicing cannot be represented.

Note that devoicing is not the only way to satisfy the constraint on voiced obstruents word-finally. Consider the underlying form /kb/ that violates this constraint. Possible grammatical responses to this violation can be devoicing (/kb/ → [kp]), C-deletion (/kb/ → [k]), V-insertion (/kb/ → [kbə]), segment reversal (/kb/ → [bk]) and more. But as Steriade (2009) mentions, devoicing is the only attested response to the violation of the constraint. To solve this, she suggests a mechanism she calls *P-map*, a mental representation of how different contrasts are distinct in various positions which favors devoicing over other responses to the constraint violation.

1.3.2.2 EP and final (de-)voicing

OT's explanation to the final (de-)voicing typology attributes the asymmetry to analytic bias: final voicing cannot be represented while final devoicing can. EP's explanation attributes the asymmetry to channel bias. Blevins (2006) suggests that the source of the (de-)voicing asymmetry stems in the existence of multiple change-supporting factors that direct the change towards final devoicing while there are no documented factors which do the same towards final voicing. According to her, there is no need to assume this pattern is a result of an innate phonological knowledge; rather, she suggests that it is an emergent property. She provides multiple different factors that can support the emergence of final devoicing. First, the following phonetic factors:

(10) **Phonetic factors for final devoicing**

a. Many languages use **laryngeal gestures** in order to mark phrase boundaries (Blevins, 2008). A phrase boundary is also a word boundary thus some words might have both finally-voiced and finally-voiceless versions, allowing CHOICE (3c) to take place.

b. **Phrase-final lengthening of stops** leads to segmental lengthening (Blevins, 2004, pp. 104-105), which might either:

   i. lead to an unintended devoicing due to the decay of voicing when intraoral air pressure rises in the stop.

   ii. be misperceived by the listener as voiceless, as the duration of voiceless stops is longer than the duration of voiced stops – CHANGE (3b) in EP terms.

c. In some languages the main cue for voicing contrast is present in audible release (VOT for example). A possible **absence of audible release** in phrase-final position, word-final position and before obstruents (Steriade, 1999) might lead such cues to be missing or hard to perceive, allowing CHANGE (3b) to take place.

According to another factor that Blevins mentions the directionality of the change is expceted to be phrase to word: during early language acquisition, up to 60% of the utterances that the child hears are single-word utterances. As a result, the child might interpret phrase-final patterns as word-final patterns. Finally, according to Ohala's (1997) Aerodynamic Voicing Constraint voicing is inhibited on obstruents, which means that along with the factors in (10), obstruents are more likely to be devoiced than other consonants. All of these factors are proposed by EP as sources for channel bias whose expected results are described in (11).

(11) **EP expectations regarding the final obstruent devoicing**

a. Final devoicing is expected to be a common sound pattern

b. Final devoicing emergence is expected to be gradient. In early stages, it is expected to:

i. appear only phrase-finally.

ii. occur as an optional pattern.

These expectations are borne out: as mentioned in the beginning of this section, final devoicing occurs in many unrelated languages and it is indeed a common sound pattern as stated in (11a). The second expectation is attested in Gulf Arabic (Holes, 1990, p. 261) where voiced plosives tend to be devoiced at the end of utterances (11b-i) and in Iraqw (Mous, 1993, p. 38) where stops are optionally devoiced at word end (11b-ii).

As for final voicing, Blevins claims that the pattern is not impossible to represent. But unlike the factors mentioned before which support the emergence of final devoicing, there is no documented factor that would support final voicing.

1.4 Conclusion

We saw two approaches to typological explanation: analytic bias and channel bias. We saw how Standard OT deals with Jakobsonian syllable structure and the final (de-)voicing asymmetry by attributing the two generalizations to analytic bias, and we saw how EP tries to explain the two using channel bias.

The debate regarding the division of labor between the two approaches is ongoing. In response to EP's explanation for the final (de-)voicing typology, Kiparsky (2006, p. 5) provides examples for known natural changes that, when chained together, might produce final voicing. According to him, these scenarios should have lead to an emergence of final voicing if there was no innate bias against it. Here are two of Kiparsky's examples:

(12) *Scenario 1: chain shift resulting in markedness reversal*

Stage 1:	tatta	tata	tat	(*tatt)	(gemination contrast)
Stage 2:	tata	tada	tad	(*tat)	(lenition)

- Result at stage 2: new voicing contrast, word-final phonological voicing.

Scenario 2: lenition plus apocope

Stage 1:	takta	tada	(*tata, *data, *tat, *dat)	(allophonic V_V voicing, no final -C)
Stage 2:	takta	tad	(*tat, *dat, *dad, *dat)	(apocope, unless final *-CC would result)

- Result at stage 2: allophonic voicing of word-final stops.

In the first scenario there is a gemination contrast in the first stage, and geminates are not allowed word-finally. All consonants in this stage are voiceless. In stage two the language undergoes a sound change of lenition: geminates are degeminated (but remain voiceless) and voiceless singletons become voiced. This means that the gemination contrast we had in the first stage becomes a voicing contrast and since geminates were not allowed word-finally before, now final voiceless consonants are not allowed. In other words, the result at stage two is word-final voicing.

In the first stage of the second scenario consonants are not allowed word-finally, and consonants between two vowels are voiced. In other words, no voiceless consonants appear in V_V position. In the second stage the language undergoes apocope: the vowels that appeared word-finally are lost, except for the ones that appeared after a cluster of two consonants. This means that the voicing process in V_V position now occurs in V_# position as well. V_# is the only context where consonants can appear word-finally, since apocope did not apply to words ending with CCV. This means that stage two results in an allophonic voicing word-finally.

The processes that Kiparsky describes are documented: Finnish is an example for lenition and degemination (Kiparsky, 2003), Korean is an example for intervocalic voicing (Cho, 1990) and apocope occurs in Lardil (Round, 2011). The fact that final voicing is not attested given these scenarios is problematic to EP according to Kiparsky (2006): the phonetic factors (10) only explain why final devoicing is common but not the lack of final voicing languages. However, nothing guarantees the pattern will last. Even if final voicing emerged the phonetic factors (10) would still be there, and we should take into account their interaction with the pattern.

It is hard to examine EP empirically as spoken languages do not leave traces. In the following chapter I present a computational model for EP which includes transmission of phonological knowledge between generations, and noise that models Blevins's (2006) phonetic factors. Using this model I put Blevins's claims to test, as well as Kiparsky's response to these claims. The simulations I ran using this model show that channel bias can explain the emergence of final devoicing, as well as the decay of final voicing had it emerged given scenarios like the ones Kiparsky mentions.

Chapter 2

Modeling channel bias

As mentioned in the previous chapter, phonetic factors which corrupt the data transmitted between generations might lead to typological asymmetries according to EP. Over generations, these similar cumulative changes to the data lead to the creation of phonological structure that is consistent with the corruption pattern. In this chapter I will try to model these assumptions. In section 2.1 I provide a model for transmitting phonological knowledge between generations based on Kirby's (2001) Iterated Learning Model. In section 2.2, in order to evaluate the influence of learnability on the emergence of phonological patterns, I discuss the task of learning and provide a model for this task. I survey MDL as a criterion for learning and explain why an MDL learner is a good model for an individual phase of sound change. In section 2.3 I discuss challenges to learning phonology under noise and present a model for dealing with such noise under a rule-based framework. Once we have such model we can put EP's claims to test along with the question regarding the division of labour between analytic bias and channel bias in shaping phonological systems.

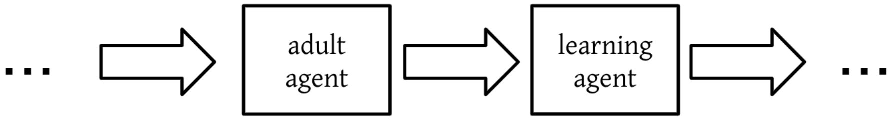


Figure 1: **A simple iterated learning mode**l. In every generation there is a single learning agent which observes utterances produced by an adult agent which was the learning agent in the previous generation. The learning agent induces a grammar based on the observed utterances and then becomes the adult agent of the next generation. Since the learning agent is not exposed to all of the utterances the adult agent can produce, the grammars of the two learners might be different.

2.1 Iterated Learning Model

The term 'channel bias' refers to systematic errors in transmission between speaker and hearer. This means that in order to model channel bias we need a model for the transmission of linguistic knowledge. An example for such model is the Iterated Learning Model (ILM; Kirby, 2001), where one or more learning agents try to induce a grammar based on utterances generated by one or more adult agents. The adult agents in every generation were the learning agents of the previous generation, meaning that this is a model for cultural evolution. As human learners, the learning agents are not exposed to the entire knowledge of the adult agents. There is a learning bottleneck which enforces the learners to generalize in order to generate utterances for meanings they were not exposed to. This bottleneck is the main force which supports cultural evolution in the model. According to Kirby et al. (2004), the tightness of the bottleneck is expected to affect the cultural evolution. If the bottleneck is too loose there is no pressure towards generalizations and if it is too tight the learner might not be exposed to enough data for the generalizations to be evident. A simple illustration of this model is presented in Figure 1.

ILM has been mostly used to show the emergence of compositionality, for example in Kirby (2000, 2001, 2002); Kirby et al. (2004); Smith et al. (2003) and more. In order to model phonological channel bias some modifications to the model are needed. First, the signals should be phonological surface forms. This means that the learning agent

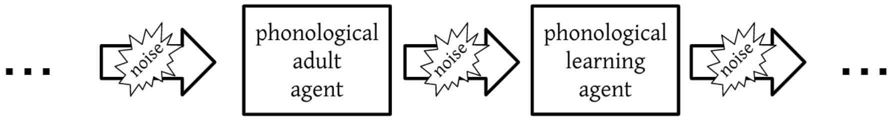


Figure 2: **A phonological channel bias iterated learning model**. The adult agent produces surface forms to which phonological noise is applied. The learning task of the learning agent is to induce a lexicon and a phonological grammar from the data after the noise has been applied to it. As before, there is a learning bottleneck which is expected to push the learner to generalize.

needs to be capable of learning phonology, i.e. to induce a lexicon and a phonological grammar from the surface forms. Since we want to use the model in order to examine whether channel bias can result in certain typological asymmetries, the learner should be capable of representing both sides of the asymmetry equally. Lastly, we need to model the channel bias itself. Transmitting phonological forms between generations and letting the bottleneck do the entire job of cultural evolution is typologically aimless. Patterns might be interpreted as regularities by chance, but since the learner is not biased towards any side of the asymmetry there is no reason these patterns will be of one side of the topological asymmetry but not the other. If we apply noise modeling channel bias to the data between generations, we should expect an emergence of a pattern that matches the noise but not of the opposite pattern. The rate of the noise is expected to affect the emergence of the pattern: if the noise is applied to all of the forms passed to the next generation, an immediate complete emergence of the pattern is expected as there will be no evidence for the learning agent that the noise is not a part of the adult's intended behavior. On the other hand, if the rates of the noise are very low, an emergence of the pattern is expected to take a large number of generations. The learner might even filter out such low rates of noise as explained in section 2.3 below, preventing from the pattern to emerge altogether. An illustration of a phonological channel bias ILM is shown in Figure 2.

Note that according to Niyogi and Berwick (2009) linear models such as ILM cannot explain certain kinds of language change. They show that these models cannot predict the changes that led to the evolution of Middle English syntax while nonlinear models they call Social Learning Models can. In addition, according to Brochhagen et al. (2018) ILM does not model all of the forces that shape languages. It models *learnability*: the pressure towards generalization due to the learning bottleneck leads to simple and more regular languages. But they mention another force that ILM does not take into consideration, *communicative efficiency*, which might push a language towards the opposite direction from learnability. They illustrate the contrast between the two forces by providing the following extremes: a language that maps every different meanings to distinct forms is good for communication since no form is ambiguous – but each form needs to be learned, while a language that maps all meanings to a single form is very easy to learn but worthless communication-wise because of the complete ambiguity of the single form. In order to balance between learnability and communicative efficiency, Brochhagen et al. use the replicator-mutator dynamic and show that it can describe coevolution of lexical meaning and semantic use. However, it is not clear that models more complex than ILM are needed to describe evolution of phonological patterns and its simplicity and transparency make it a good starting point for investigating phonological patterns evolution. If ILM turns out to be insufficient for this task, other methods should be considered.

Now that we have a model for phonological cultural evolution, we need to scope in and discuss the phonological learner itself, which I do in the following section.

2.2 Modeling learning

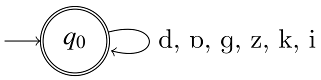
In order for us to model phonological channel bias, we need to model phonology learning properly. First, our learner needs to be able to induce a lexicon and a phonological grammar from surface forms. Second, the learner should not have any preference towards any side of the typological asymmetry we are investigating, as we want to examine whether channel bias can account for the asymmetry. Third, the learner should be able to deal with optionality, as according to EP the emergence of certain phonological patterns is expected to be gradual, in which cases these patterns might surface as optional patterns. In this section I show that Rasin et al.'s (2018) SPE MDL learner matches these three criteria.

2.2.1 Learning phonology

Children acquire key aspects of language, and specifically various phonological patterns, simply by listening. The child acquires a grammar, a characterization of their knowledge, which allows them to generate the data they were exposed to along with additional words which obey the internal logic of the language. We can think of the task of learning as finding a *target grammar* which is the best way to generate the set of observed data *D*. But what is this "best way"? Assuming there are multiple grammars that can generate *D*, how does a learner choose one grammar over the other?

One possible theory of the learners' strategy is the tendency to *simplicity* in the same terms defined by William of Occam in the 'Occam's Razor' principle: "*entities should not be multiplied without necessity*". An evaluation metric for comparing phonological grammars which is based on the idea of simplicity was suggested by Chomsky and Halle (1968, p. 334): the *economy* evaluation metric. If two grammars G', G'' can both generate *D* and G' contains fewer symbols than G'', prefer G' over G''. However, such bias towards a simpler grammar might result in over-generalization. Consider the following toy version of English which contains the English phonological forms [dɒg], [kid], [gig] and the plural forms of the first two: $D = \{$dɒg, dɒgz, kid, kidz, gig$\}$. While there are no phonological patterns in *D*, it can be used in order to demonstrate the economy metric. Let us assume that a learner is exposed to *D* and tries, using the metric, to induce a grammar that allows it to generate *D*. Choosing nondeterministic finite automata (NFA) in order to represent string-generating grammars, the grammar $G_{minimal}$ that generates strings based on concatenation of the inventory of segments found in *D* can generate *D*, like in the following example:

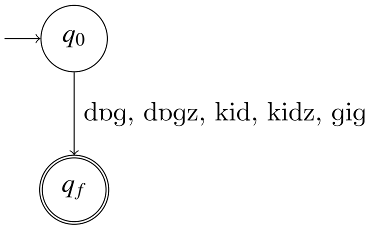
(13) $\mathbf{G}_{minimal}$



By Chomsky and Halle's economy metric, $G_{minimal}$ should be preferred on any other grammar that can describe *D* as it contains only the minimal subset of building blocks required to generate the utterances in *D*. This grammar clearly over-generates, as it is hard to think of a language in which every arbitrary concatenation of segments is grammatical. (13) can generate all the forms in *D* and even generate [gigz] which do not appear in *D* but seems like a proper generalization, as it is the plural form of [gig]. However, (13) can also generate forms like kiiiiikz, dgzk, dddddd which do not seem to obey English (nor our toy English) phonotactics and are not expected to be a part of learner's language given the data in *D*.

This means the evaluation metric used by human learners restricts the target grammar to the observed data somehow. An example for such restriction is the *subset principle* (Dell, 1981; Berwick, 1985; Wexler and Manzini, 1987): if two grammars G', G'' generate $L' \subseteq D, L'' \subseteq D$ respectively and $L' \subsetneq L''$ – prefer G' over G''. Using this metric to evaluate its hypotheses for a target grammar will solve the over-generalization issue. Specifically, it allows the learner to choose a grammar G_{repeat} which just repeats the observed data, as in the NFA in (14) which can generate *D*.

(14) $\mathbf{G}_{repeat}$



The grammar in (14) can generate only the forms in the data set and unlike (13) cannot generate the form gigz. This is obviously not what human learners do as humans are capable of generating unheard utterances. In other words, a bias towards restrictiveness alone leads to under-generalization.

If a bias towards economy results in over-generalization and a bias towards restrictiveness results in under-generalization, a metric that balances between the two seems reasonable. Such metric is Kolmogorov Complexity (KC), built upon the pioneering work by Solomonoff (1964) and developed independently by Kolmogorov (1965) and Chaitin (1966). KC formalizes the notion of Occam's Razor by measuring object's complexity by the length of the shortest description of the object in a Turing complete description language. For example, if we use Turing machines as our description language and use it to describe a string s, then KC of s is the encoding length of the shortest Turing machine that prints s and halts, and if we use Python programming language as our description language then KC of s is the length of the shortest Python script that prints s and halts.

A learner looking to minimize the description length of the data will achieve a grammar that balances between economy and restrictiveness. Economy is guaranteed because we are looking for the *shortest* description of the data, and restrictiveness is guaranteed because we are looking for a description which is specific for the data.

2.2.2 Minimum description length

While KC serves as a good metric for finding the target grammar, it is not computable. However, there are approximations for it such as Minimum Description Length (MDL; Rissanen, 1978), a criterion for comparison of hypotheses, which concentrates on minimizing the combined length of the hypothesis and the data encoded given the hypothesis. Namely, given a data set D and a hypotheses space H, the best hypothesis to describe D is the one that satisfies:

(15) $argmin_{G \in H}\{|G| + |D{:}G|\}$

In the case of phonology, G is a phonological grammar and a lexicon. The size $|G|$ is the number of bits it takes to encode G. We can think of $|D{:}G|$ as the length of the instructions needed to be fed to G in order to generate D, or more formally, $|D{:}G|$ is the number of bits that are needed in order to encode D using G. Every binary choice adds a bit to $|D{:}G|$, so a choice between n options adds $log_2(n)$ bits to $|D{:}G|$.

MDL, along with the closely related Bayesian approach to learning, has been shown to be successful in tasks of learning grammar from distributional data, for example in Horning (1969); Berwick (1982); Rissanen and Ristad (1994); Lan (2018); Rasin et al. (2018, 2020) and more, and there are conceptual arguments in favor of MDL as a learning criterion. The MDL criterion comes almost for free, as discussed in Katzir (2014). The grammar is stored in the speaker's memory, taking $|G|$ space. It is used to parse inputs, so if the speaker stores the instructions on how to parse the input using the grammar, or $D{:}G$, it should take $|D{:}G|$ space. The sum of these two, $|G| + |D{:}G|$, is the MDL quantity. If this quantity is available to the learner, the only thing it needs in order to learn is the ability to compare this quantity for its current hypothesis with neighboring hypotheses and move gradually towards the best grammar in MDL terms. There is also empirical support for MDL as a learning criterion: word learning (Xu and Tenenbaum, 2007), casual reasoning (Sobel et al., 2004) and visual scene chunking (Orbán et al., 2008) are some of the lab experiments mentioned by Rasin and Katzir (2020) where human subjects seem to balance between the general plausibility of the hypothesis, analogous to $|G|$, and its specificity, analogous to $|D{:}G|$.

As Rasin and Katzir (2016) point out, if we think of the length of the hypothesis, notated by $|G|$, as the number of bits it takes to encode G then $|G|$ is analogous to economy: a simpler hypothesis usually takes fewer bits to encode. The length of $D{:}G$, the data described using G, is analogous to restrictiveness: an hypothesis that requires fewer bits in order to describe the data accounts for data's idiosyncrasies hence minimizing |*D*:*G*| ensures us a plain explanation of the observed data.

If a grammar is too simple, the limitations on the possible forms generated by this grammar are fewer. Intuitively, accounting for the data observed given such a grammar will require a complicated explanation. On the other hand, a complicated grammar might overfit the data as in the case of G_{repeat} discussed in the previous section, and prevent us from generalizing. This means that a good hypothesis for a target grammar by the MDL criterion is one which is simple, yet still allows us to provide a simple explanation for data regularities.

2.2.3 No (de-)voicing bias

We saw that MDL is a good metric for phonology learning, but we still needs to make sure our learner is not biased towards final devoicing nor final voicing, as this is one of EP's main assumptions: nothing is innate nor grounded regarding the two patterns, and noise does the entire job. In order to model this assumption, we can use a substance-free version of SPE. As shown in (16), the two patterns can be represented in such framework.

(16) a. $[-son] \rightarrow [-voice]/\_\_\#$

b. $[-son] \rightarrow [+voice]/\_\_\#$

MDL is a complexity metric. Rasin et al.'s (2018)'s SPE MDL learner treats the two rules in (16) as having the same complexity, since the ± value of the *voice* feature in the rules does not change the rule's complexity. This means that this learner fits our phonological channel bias model, as long as it can deal with optional patterns – which it indeed can, as we shall see next.

2.2.4 Optional rules in MDL

According to EP, patterns are not expected to emerge suddenly: it is a gradual process in which cumulative changes which seem arbitrary at first turn out to be systematic. It means that there are stages between phase *A* where the pattern does not exist and phase *B* where the pattern is completely systematic. During these middle stages speakers might produce both utterances where the pattern occurs and utterances where it does not occur (as EP expects to happen in final devoicing (11b-ii)). Note that optionality does not entail instability and it might be the case that an optional pattern is stable.

Optional patterns can be modeled in SPE as rules that apply optionally, and an MDL learner might represent these optional rules as follows: the choice whether to activate the rule or not is translated to a binary choice, adding a bit to $|D{:}G|$ for every such choice when parsing the data. Using this method, Rasin et al.'s (2018) SPE MDL learner has successfully learned the optional pattern of final liquid-deletion that appears in French (Dell, 1981). This means their learner can represent the middle stages during the emergence of a phonological pattern that EP expects, and as such it fits as a model for human learner while investigating EP's assumptions.

However, there is still one issue that needs to be addressed. Channel bias is modeled as noise in our model. Highly infrequent forms are not expected to change grammars, which means human learners are capable of filtering them out as noise somehow. In the next section I consider the task of learning under noise and suggest a noise-correction mechanism within SPE.

2.3 Learning with noise

As mentioned in section 2.2.1, children acquiring language have to induce a target grammar that can describe the data they are being exposed to. This task seems hard enough even if the set of data contains only forms that are valid in the language, but we know that the actual distributional data that the child is exposed to contain exceptions due to performance errors or variation among speakers. According to EP, in some cases these errors are learned as systematic and lead to the emergence of phonological patterns, but in some cases mistakes are not a result of a recurring and predicted phonetic factor: they might be a one-time slip of the tongue and make the data less systematic and harder to find generalizations in, i.e. harder to learn. Children should be able to filter out these non-systematic errors for their learning mechanism to be efficient.

The Gradual Learning Algorithm (Boersma, 1997) was suggested as a solution to this issue within constraint-based theories. Boersma and Hayes (2001) show that the algorithm is robust in the face of noisy data, and that perturbations are minimal if the learning algorithm is exposed to a small number of erroneous forms. In SPE there are multiple ways of representing grammatical exceptions, for example, using diacritics. We can mark certain exceptions in the lexicon with a feature, say [−rule *n*] which excludes the exception from the domain of rule *n*, telling the rule to not apply. Another way of dealing with exceptions in SPE is modifying their representation in some ad-hoc way. For example, in order to explain why the stress pattern of words bearing the *-ion* suffix in English does not obey English Main Stress Rule, Chomsky and Halle (1968, p. 87) suggest that the UR of the suffix is $/\mathrm{i\breve{V}n}/$, where $/\breve{\mathrm{V}}/$ represents the archi-segment “lax vowel”. This way words like *prohibition* are represented as /$\mathrm{prohibit} + \mathrm{i\breve{V}n}$/. According to the Main Stress Rule the primary stress is assigned to the final vowel of the word unless this vowel is lax, in which case the stress is placed on the penultimate syllable. This means this ad-hoc lax vowel suffix solution leads to a correct assignment of the primary stress in *prohibition*.

An SPE learner that has no noise correction mechanism must provide an ad-hoc explanation for the exceptions and cannot filter them out as noise. Every form that such a learner is exposed to must be explained, and if the data contains errors the learner might need to provide a complicated explanation to the data in order to overcome them. For example, consider the following surface forms of the English words dog, cat, bee, tap and their plural forms:

(17) dɒg, dɒgz, kæt, kæts, biː, biːz, tæp, tæps

A learner that is exposed to these data should successfully learn that the data can be described using the morphemes /dɒg/, /kæt/, /biː/, /tæp/, an optional plural suffix /z/ (or /s/, or another phoneme from which the two are derived) and some kind of voicing assimilation rule, turning [z] to [s] in the forms [kæts], [tæps]. If we add to the data an exceptional form such as [kætz], an SPE learner that has no error handling mechanism will have to represent the erroneous form explicitly: it could, for example, choose to represent /z/ and /s/ as two separate suffixes or to have an optional voicing assimilation rule instead of the mandatory rule used to describe the exception-less data.

Note that frequency plays a role here. In this example one out of five plural forms was exceptional, i.e. only 80% of the forms follow the voice assimilation pattern. Such a high exception rate might be expected to change the learner's grammar, making it plausible for the learner to choose to represent both suffixes or change the assimilation rule to be optional as suggested above. But human learners are exposed to much larger data sets which do not contain this rate of consistent errors. Even if an English learner is exposed to forms like [kætz], the evidence in favor of voicing assimilation would be much more significant than 80% of the forms. Even so, an SPE learner that has no way to correct or filter out noise must choose a grammar in which [kætz] is represented as any other form. This does not match actual child learners that manage to acquire a grammar that does not allow generation of such noisy exceptions.

In order to model language change properly our model should be able to filter out noise somehow. Noise correction can be achieved within learning strategy details (for example, "try to explain 95% of the observed data") or within the phonological framework. Below I show how exceptions are represented in SPE and present a modification to SPE which allows correcting noise using optional SPE rules.

2.3.1 SPE and exceptions

As mentioned, our learner needs to deal with noise in the data in some way. Before I present my suggestion for noise-correction mechanism, let us see how SPE deals with exceptions. In order to compare these methods with this noise-correction mechanism, let us first formulate the issue at hand. In order to make things simple, I limit the data set D so the target grammar G that describes it is a *simple grammar* (18): every form in D can be generated by G by concatenating m morphemes and then applying to the result the (possibly empty) set of rules. Each morpheme k has o_k options to choose from when generating a form, with the empty morpheme ϵ as a valid option. So *simple* refers to the fact that the morphology is concatenative and there are no limitations on morpheme concatenation (such as lexical categories) except for the order of the morphemes.

(18) **Simple grammar**

Lexicon: $\{morph^1_1, \ldots, morph^1_{o_1}\} + \cdots + \{morph^m_1, \ldots, morph^m_{o_m}\}$

Rule set: $\{\ldots\}$

To illustrate, I will use Russian final l-drop pattern formulated as the following rule in Kenstowicz and Kisseberth (1979, p. 57):

(19) $\mathrm{l} \rightarrow \emptyset / C\_\_\#$

In order for a learner to learn a grammar with a rule set that consists only of rule (19) under a plain SPE framework and no noise-correction mechanism, the entire data has to be consistent with it. The Russian data in (20) from Kenstowicz and Kisseberth (1979, p. 55) is an example of such data. Since I use MDL to model learning, I included the MDL score (21c) of Rasin et al.'s (2018) learner for this grammar. $|D{:}G|$ represent the number of binary choices multiplied by 10 in order to make sure $|D{:}G|$ is not negligible compared to $|G|$.

(20) nes, nesla, pek, pekla, sek, sekla, pas, pasla, pisal, pisala

(21) a. Lexicon: {nesl, pekl, sekl, pasl, pisal} + {a, ϵ}

b. Rule set:

i. $\mathrm{l} \rightarrow \emptyset / C\_\_\#$

c. $|D{:}G| + |G| = 332.19 + 169.9 = 502.09$

The simple grammar in (21) which can describe the data in (20) has only stems {nesl, pekl, sekl, pasl, pisal} and suffixes {a, ϵ}, namely $m = 2$ with $o_1 = 5$ and $o_2 = 2$. Now, let us see what happens if our data do not fall into regularities perfectly due to performance errors either made by the producer (articulatory errors) or by the learner (perception errors). As we shall see, noisy data sometimes cannot be described elegantly by a classic SPE rule system. Let us assume that due to a slip of the tongue a form $f \in D$ was produced slightly differently: one of the m morphemes that this noisy form consists of, say the k-th morpheme, is different in a bundle of phonetic features from the k-th morpheme in f. A segment or multiple segments might be missing or added to this morpheme, or some feature values of the morpheme's segments might have shifted. This new exceptional form which we shall mark as f' is added to the data fed to the learner, creating a *noisy data set* $D' = D \cup \{f'\}$, as summarized in (22):

(22) **A noisy data set** is a data set D that can be generated using a simple grammar G, plus at least one form f' which G cannot generate and differs from $f \in D$ in the k-th morpheme.

In our example we can assume that due to a slip of the tongue the form [nes] was produced as [nesl]. Adding it to the to the data in (20) results in the following D':

(23) nes, nesla, pek, pekla, sek, sekla, pas, pasla, pisal, pisala, nesl

The new data set might break the generalizations of G just as happens in (23), as the final l of /nesl/ is not deleted even though it comes after a consonant. We can use this data set in order to examine different ways which SPE can deal with exceptions without explicitly marking them as exceptional. Let us look at ways of changing G minimally so it can generate D'. Such a minimal change is described in (24).

(24) **Treating noise with optional rule(s)**: Given a noisy data set (22) D'; we shall define a grammar G' which differs from G only by its set of rules: it either

contains additional optional rule(s), or some of G's obligatory rules are optional in G'. Applying or not applying these rules can change the k-th morpheme in f to match the surface form of the k-th morpheme in f'.

G' can generate D by mimicking G's choices, not applying any of the new optional rules and applying every optional rule that was obligatory in G. It can also generate f' by selecting f's morphemes and applying only the rules that lead to the surface form f'. In our Russian example we can apply (24) by changing the l-deletion rule to an optional rule which yields the grammar in (25). Choosing not to apply the rule will generate the noisy form [nesl].

(25) a. Lexicon: {nesl, pekl, sekl, pasl, pisal} + {a, ϵ}

b. Rule set:

i. l→ $\emptyset / C$__# (optional)

c. $|D'$:$G'|$ + $|G'|$ = 415.41 + 169.9 = 585.31

For every optional rule added, including obligatory rules changed to optional, one bit is to be added to $|D$:$G|$ for every form that the rule applies or could apply to. In our example this applies to four of the forms in D: {nes, pek, sek, pas} plus the noisy form [nesl]: a total of 5 bits which add 50 to the MDL score. There is no change to $|G|$ since the lexicon was not changed nor new rules were added. The grammar in (25) over-generates as it can generate a version with final /l/ for each one of the /l/-deleted forms in the data which cannot be generated by the grammar in (21). As expected, this is indeed reflected in higher $|D$:$G|$: the average cost[1] of generating a string in the data in (21) is 33.22 and in (25) it is 37.76.

Another way noise can be dealt with is changing G's lexicon instead of its rule set so it can generate D':

[1] The reason I mentioned the average cost here is that I am comparing grammars that deal with different sizes of data, which inevitably affects $|D$:$G|$, while $|G|$ might stay the same as can be seen in (25).

(26) **Add an option for the noisy morpheme**: Given a noisy data set (22) D'; we shall define a grammar G' which differs from G by its lexicon: adding another option to the k-th morpheme which either (i) differs from the k-th morpheme in f in the same bundle of features that f differs from f'; or (ii) if the difference is due to a phonological rule that should not be applied, the new option can have a "protective" environment which prevents the rule from applying (in which case we might need to add a rule to "clear" the protective environment).

It is clear why G' in this case can generate D': the old options for the k-th morpheme are valid as before, and f' can be generated using the new morpheme. For our Russian example (26-ii) can be applied:

(27) a. Lexicon: {neslk, nesl, pekl, sekl, pasl, pisal} + {a, ϵ}

b. Rule set:

i. $\mathrm{l} \rightarrow \emptyset / C\_\_\#$

ii. $\mathrm{k} \rightarrow \emptyset / \mathrm{l}\_\_$

c. $|D'{:}G'| + |G'| = 394.35 + 232.94 = 627.29$

(27) can generate all of the forms that (21) can generate plus the form [nesl]: in the option /neslk/ which was added to the stems in (27) the segment /l/ does not occur in a word-final position when (27b-i) applies, blocking l-deletion. Then, the final /k/ is deleted due to (27b-ii), generating [nesl]. Unlike (25), it cannot generate any other l-final forms. The grammar (27) is based on the idea of abstract URs: the morpheme /neslk/ is used here as an ad-hoc solution for the exception [nesl] and its last segment /k/ is not supported by an alternation. Such an explanation was given by Chomsky and Halle (1968) for the English word *nightingale*: their claim was that it contained /x/ which is never realized in English. There is also evidence from Nupe that supports abstract URs (Hyman, 1970). On the other hand, there are arguments in favor of limiting the abstractness of phonological representations. For example, Kiparsky (1968)

discusses two kinds of neutralization – *absolute neutralization* where the alleged underlying contrast never surfaces and *contextual neutralization* where the merge of phonetic contrast occurs only in certain contexts. While contextual neutralization is reversible as in Yiddish which lost the final devoicing pattern (Sapir, 1915; Weinreich, 1963) which it inherited from German, there is no evidence that absolute neutralization is reversible. Kiparsky claims this means that if absolute neutralization emerges in a language future generations of learners will analyze the pattern as lexical rather than absolute phonological neutralization, leading him to claim that absolute neutralization does not exist in synchronic grammars and to suggest a constraint he calls the alternation condition, formulated as follows by Kenstowicz and Kisseberth (1979, p. 215):

(28) **Alternation Condition**: Each language has an inventory of segments appearing in underlying representations. Call these segments phonemes. The UR of a morpheme may not contain a phoneme /x/ that is always realized phonetically as identical to the realization of some other phoneme /y/.

Kenstowicz and Kisseberth mention that it is not clear whether the internal evidence supporting abstract URs is valid or not. The MDL criterion, however, can help us explain how abstract URs can be learned as shown in Rasin et al. (2020). And even so, this solution means the exception or the error is treated as every other form in the data and not as noise. Returning to the grammar in (27), since we added an option for the k-th morpheme, generating each of the forms in D' requires choosing from $o_k + 1$ options for the k-th morphemes instead of o_k options. Let us see how this affects $|D{:}G|$. The cost of every choice between n options when generating a form in D' is $log_2 n$. Since we added a new option for the k-th morpheme, the cost of choosing an option for this morpheme when generating a form is now $log_2(o_k + 1)$ instead of $log_2(o_k)$. So, the increase in $|D{:}G|$ for a single form is $log_2(o_k+1)-log_2(o_k) = log_2(1+\frac{1}{o_k})$ bits. Since this increase is relevant for all forms in D', the total increase in $|D{:}G|$ is $|D'| \cdot log_2(1 + \frac{1}{o_k})$, and generally, if e_k options were to be added to the k-th morpheme, the increase in

$|D{:}G|$ for choosing an option for this morpheme would be $|D'| \cdot log_2(1 + \frac{e_k}{o_k})$.

In our example in (27), the average $|D{:}G|$ is 35.85 per form which is higher than the average of the exceptionless grammar (21), as expected. In addition, $|G|$ was affected badly by the addition of a rule and a morpheme option.

Next I present a noise-correction component which allows the learner to filter out noise and discuss in which scenarios it should be worthwhile for the learner to use this component over modifying its grammar.

2.3.2 Noise correction component

The changes to G presented above treat exceptions as every other form in the data, either by softening the generalization or providing an ad-hoc solution to the exceptions, meaning these exceptions can be generated by the speaker and passed to the next generation.

There should be a room for lexicalized exceptions. But the grammar should distinguish inconsistent errors from consistent exceptions based on frequency. Exceptions which survive generations are the highly frequent ones and might play a role in language change. But not all exceptions are treated alike as it would be easier for the learner to generalize if it somehow filtered out one-time slips of the tongue. The learner could, for example, ignore some of the forms it cannot parse with its current evaluated hypothesis – but this solution might be too strong and good irregular forms might be filtered out. Another option is to provide the learner with a “noise correction” component: we let our learner assume that the data was exposed to articulatory noise or perception noise – i.e., that the strings that it perceived might slightly differ from the ones it was intended to perceive.

Such a noise correction component can be easily achieved within SPE system by adding optional phonological rules to the system which apply after all other rules applied, meaning they do not interfere with “regular” rules but only can fix (or corrupt) the final output. For example, in order to model noise that changes the [*voice*] feature of any segment in a form perceived by the learner, we can add the optional rule in (29). I am using the notion *(noise)* to represent the fact that this rule is optional, that it can take effect only after every non-noise rule has taken effect[2] and that it is dealt differently in terms of MDL as will be discussed in section 2.3.3.

(29) $[\alpha voice] \rightarrow [\beta voice]/\_\_$ (noise)

Similarly, we can add the optional rule in (30) to model noise of vowel insertion or the rule in (31) to model noise of consonant deletion.

(30) $\emptyset \rightarrow [-cons]/\_\_$ (noise)

(31) $[+cons] \rightarrow \emptyset/\_\_$ (noise)

Using this method we can choose which types of noise we accept and which ones we do not. For example, we can limit rules to apply only in certain contexts (only word-finally, only after vowels, etc.) or assign different weights to different types of noise.

Another important limitation when determining whether a form is an error or not is the frequency of either the token or the phonological pattern, as we want frequent patterns to be not considered as noise. As we shall see in the next section, by assigning the right MDL cost of noise rules we can determine how frequent a pattern has to be in order to force the MDL learner to explicitly represent it in its grammar.

2.3.3 MDL cost of noise correction

Noisy forms are less likely than non-noisy ones. Since $|D{:}G|$ of a single form represents the binary choices taken in order to parse or generate this form, it is tied directly to the form's likelihood. For this reason we want $|D{:}G|$ of an error to reflect the fact that noise

[2] This is achieved by rule order in SPE. In other words, we can say that a rule set that has a rule with a *(noise)* notion before a rule without the notion is an invalid rule set.

correction was needed in order to parse it. As explained in section 2.2.4, in regular optional rules the choice whether to activate the rule or not when parsing a form is translated to a binary choice meaning that both options are translated to a single bit in $|D{:}G|$. In the case of noise-correction optional rules we do not want the choices to be equal: activation of the rule should be expensive to the learner and not activating the rule should cost a little if anything at all as no additional assumptions about the data are required.[3] Since we want the cost of the rule to represent likelihood, the choice of the cost should be related to the expected noise ratio in the data. If we want noise correction to apply instead of changing G's rule set (24) or lexicon (26), the entire noise we want to allow should cost less than the cost each one of these options: it must be less than the sum $|D| \cdot \sum_{i=1}^{m} log_2(1 + \frac{e_i}{o_i})$ where e_i is the number of the options added to morpheme i in order to deal with errors and o_i is the current number of options for this morpheme, and less of the number of the contexts that the optional rule(s) in (24) can apply (or indeed apply) to. This is summarized in (32).

(32) In order for a simple grammar G to fix errors in a data set of size $|D|$ using noise-correction rules, the total cost of applying the rules should be:

a. less than $|D| \cdot \sum_{i=1}^{m} log_2(1 + \frac{e_i}{o_i})$
where o_i is the number of options for the i-th morpheme in G and e_i is the minimal number of morpheme options to add to this morpheme in order to deal with the exceptions (without noise-correction rules);

b. less than the total number of phonological contexts in D that matched obligatory rules in G which have to be turned optional in order to deal with the exceptions (without noise-correction rules); or, if no such obligatory rules, less than the cost of adding such optional rules plus the number of contexts as mentioned.

[3] This method is chosen because Rasin et al.'s (2018) learner allows only uniform probability distributions over optional rule application. In a system that allows for nonuniform probability distributions, this amounts to very unlikely optional rules that are ordered after all other rules.

Note that if the noise is systematic a single optional rule or single new option for morpheme might be able to deal with multiple exceptions, which might make the change to *G* more worthwhile for the learner than applying a noise rule. This is a desired result, as we expect consistent and frequent noise to play role in changing grammars.

Regarding the effect on $|G|$, even though noise-correction is represented as rules I chose not to include them in the calculation of $|G|$ as these rules are an integral part of the new rule system and will be added to every hypothesis, so they will not affect the decision between two hypotheses. (33) is an example for a such a noise-correction rule set: a rule for toggling any feature value (33a), a rule for segment insertion (33b) and a rule for segment deletion (33c). The notation $[\alpha feature] \rightarrow [\beta feature]$ marks a toggle of any single phonological feature value, and the notion $[feature{-}bundle]$ marks any bundle of phonological features to insert (33b) or delete (33c).

(33) **General noise-correction rules**.

a. $[\alpha feature] \rightarrow [\beta feature]/\_\_$ (noise)

b. $\emptyset \rightarrow [feature{-}bundle]/\_\_$ (noise)

c. $[feature{-}bundle] \rightarrow \emptyset/\_\_$ (noise)

As mentioned above, highly frequent forms are not expected to change grammars. Adding noise-correction optional rules to an SPE grammar can help the learner to filter out such noisy forms, making the learner suitable for our channel bias model. To illustrate, let us return to the data in (23) repeated in (34) and try to find a grammar that can explain it using noise-correction optional rules.

(34) nes, nesla, pek, pekla, sek, sekla, pas, pasla, pisal, pisala, nesl

If we assume every hypothesis has general noise-correction rules (33), [nesl] can be explained using the following modified version of the grammar in (21):

(35) a. Lexicon: {nesl,pekl, sekl, pasl, pisal} + {a, ϵ}

b. Rule set:

i. l→ $\emptyset/C\_\_\#$

c. General noise-correction rule set (33)[4]

Let us see what should be the cost of applying the noise-correction rule for insertion (33b) for the grammar in (35). If we want our learner to use it when parsing [nesl], it must be less than the cost of applying (24) on the original grammar (changing the l-drop rule to be optional) which is 5 bits or 50 in MDL score. The cost of applying the noise-correction rule for insertion (33b) in this case should also be lower than the cost of applying (26) on the original grammar, which is $|D'|{\cdot}log_2(1{+}\frac{1}{o_k}) = 11{\cdot}log_2(1{+}\frac{1}{5}) \approx 2.89$ bits, 28.9 in MDL score, plus the cost of adding another rule to the grammar which is 79.8 – a total of 108.7. This means that the MDL cost of applying the noise-correction rule for insertion (33b) when parsing [nesl] must be lower than 50. If we set the cost of applying the rule to be 4.9, the increase of the MDL score will be only 49. Therefore, the MDL score in this case is calculated as follows: $|D{:}G|$ of every single form in (34) except [nesl] remains the same, as the grammar can explain it without using noise correction rule (33b). As for [nesl], the grammar can generate it by following the generation instructions of [nes] plus applying the noise-correction rule for insertion (33b). This means that $|D{:}G|$ of [nesl] is $|D{:}G|$ of [nes] + noise cost, which in our case is 4.9 multiplied by 10, giving us 33.22 + 49 = 82.22.[5] As mentioned, $|D{:}G|$ of all other surface forms remains the same, which means $|D{:}G|$ here is the same one from (21c) + 82.22. As for $|G|$: the only difference in G is the added noise rule which should

[4] The segment-insertion noise-correction rule (33b) is correcting an addition of any segment to nes. Intuitively, this not the source of the noise in [nesl] – one would assume that this form was created by not applying (35b-i) by mistake or by chance. However, the described noise system is more general: while allowing the learner to deal with language-specific noise as in this case, it can account for noise which is not directly related to the phonological system of the language.

[5] $|D{:}G|$ of [nes] is calculated as follows: the first morpheme is chosen out of {nesl, pekl, sekl, pasl, pisal} and the second morpheme is chosen out of {a,ϵ}. The first choice costs $log_2 5$ bits and the second costs $log_2 2$ bits. The sum is multiplied by 10, giving us 33.22. Since the only non-noise rule in G is obligatory it does not affect $|D : G|$. See Lan (2018) p. 18 for more details (though Lan uses $\lceil log_2(n) \rceil$ as the cost of a choice between n options and in this work I use $log_2(n)$).

not change $|G|$, which means it is the same as in (21c). This means the MDL score of (35) for the data in (34) is $|D{:}G| + |G| = (332.19 + 82.22) + 169.9 = 584.31$ and the average encoding a of surface form is 37.67. Even though this average is higher than the one of the grammar in (27), $|G|$ is much lower and so is the total MDL score. As expected, the total MDL score here is better by one point than the one in (25). The cost of 4.9 allows only a single exception of this kind: for example, if we add to D' the form [pekl] we will need to apply rule (33b) once for each of the noisy forms, which will cost $4.9 + 4.9 = 9.8$ bits (plus 98 to the MDL score), while changing the l-drop rule to be optional will cost 6 bits (plus 60 to the MDL score). This means that while noise rules allow certain rate of noise, if the noise is systematic it will be learned as grammatical.

We now have a model for a learner that can induce phonological rules from distributional data and filter out any amount of noise we desire. For a complete EP model we still need to model the channel bias itself – which is modeled in the following chapter, where I show how the model presented in this chapter can explain the final (de-)voicing typology based on EP's assumptions. I suggest a model of noise representing the phonetic factors that are argued by Blevins (2006) to be the source of the asymmetry. I also present simulations that address some of Kiparsky's (2006) arguments against EP, showing how an existing pattern of final voicing is expected to interact with these phonetic factors.

Chapter 3

Simulations: emergence of final (de-)voicing

In the previous chapters I surveyed how different theories deal with typological asymmetries in phonology. While theories such as Standard OT attribute key aspects of the typology to analytic bias, according to theories like EP some asymmetries are the result of channel bias. I also presented a model which allows us to test EP's claims by simulating the transmission of somewhat noisy phonological knowledge between generations.

In this chapter I show how an ILM with MDL agents can be used to model the emergence of the final (de-)voicing typology. In section 3.1 we shall see how applying noise modeling the phonetic factors discussed in 1.3.2.2 indeed leads to the emergence of the final devoicing pattern from initial data with no (de-)voicing patterns. In addition, I suggest a possible explanation for why final devoicing does not emerge in every language, even if the phonetic factors which are considered by EP as the source of the pattern are similar in all languages. My suggestion relies on the different ways languages distinguish between voiced and voiceless consonants which might change the ratio of noise in the data, which might make certain languages less prone to be affected by these phonetic factors.

Later, in section 3.2 I present simulations that weaken Kiparsky's (2006) claims that if EP's assumptions were right, we would expect to see languages with final voicing. These simulations are similar to the ones mentioned before, except their initial state is final voicing and they show how the pattern decays over generations.

Each simulation ran for 25 ILM generations.[1] As for the genetic algorithm search parameters,[2] each ILM generation learning simulation ran using a genetic algorithm with 100 islands with a population of 200 hypotheses for grammar each (total population of 20,000). Each learning simulation had 675 genetic algorithm generations. Two AWS h1.2xlarge machines with 8 vCPUs each (2.3 GHz Intel Xeon) were used, with each machine running 50 islands.

3.1 Emergence of final devoicing

According to EP the source of the final (de-)voicing typology is the existence of phonetic factors which support final devoicing and the lack of phonetic factors which support final voicing. Modeling the devoicing phonetic factors as noise and applying this noise to the data transmitted between generations of speakers allows us to test EP's assumptions. I start with initial data which is not biased towards voicing nor devoicing as presented in section 3.1.1. The way the data is transmitted between two generations is described in section 3.1.2, and in section 3.1.3 I explain about the characteristics and the different rates of noise I introduced between any two consecutive generations in order to model the phonetic factors Blevins (2006) mentions.

[1] The code for the model is available at `https://github.com/taucompling/morphophonology_spe`

[2] I provide the parameters here but do not elaborate on them. See Lan (2018) for more details.

3.1.1 Initial data

The following lexicon of 16 stems and four suffixes, including the null suffix, was used to generate the initial data in all of the simulations I ran:

stems	pad, zooz, sab, nag, aab, koz, bop, taot, gaak, koas, oak, at, dam, moon, doa, maa
suffixes	on, ka, bam, ϵ

Every segment in the data is represented by specifying ± value for each one of the phonological features *voice*, *cont*, *low*, *coronal*, *labial*, *son*. The final segments of six of the stems are voiced obstruents $\left[\begin{smallmatrix}-son\\+voice\end{smallmatrix}\right]$ (d, z, b, g). The rest of the stems are not affected by the obstruent devoicing noise. Six of them end with voiceless obstruents $\left[\begin{smallmatrix}-son\\-voice\end{smallmatrix}\right]$ (t, s, p, k), and the remaining four stems end with sonorants [+*son*] (m, n, a). The final segments of the suffixes are all sonorants, since I wanted to reduce any effect they might have on the simulations: suffix final segments always come word-finally and are not contrasted with any other environment, so there is no reason for the learner to assume their underlying voicedness is different than appears on surface.[3] This means that if all occurrences of a specific suffix are devoiced, it will not support the final devoicing pattern. The final segment of the suffix will be simply reanalyzed as a voiceless segment by the learning agent.[4] In addition, while a stem may appear word-finally a small number of times, when concatenated with the null suffix, every suffix appears word-finally a relatively large number of times. This means that the impact of a single devoiced stem on the overall final devoicing pattern is significant, while a single devoiced suffix might delay the emergence of an obligatory pattern or

[3] See discussion on abstract URs in section 2.3.1, especially Kiparsky's (1968) Alternation Condition (28).

[4] This is due to the learner's implementation. A proper learner should be able to generalize the pattern, even in a data set where there are no alternations and all of the word-final segments in the data are voiceless (for example, by underspecifying the [*voice*] feature of the final segments). It is not clear how such change would affect the simulation results. The question is left for future work.

even prevent it, even if it supports the emergence of optional final devoicing.[5]

The learner in the first generation was exposed to the 64 forms that can be generated by concatenating these stems and suffixes.

3.1.2 Data between generations

An important part of the ILM is the bottleneck: only part of the data the adult agent can generate is transferred to the next generation, forcing the learning agent to generalize in order to generate forms it was not exposed to. This way systematic irregularities in the data might be learned as regularities and are more likely to be transferred to the following generations.

The data fed to the learning agent in the first generation consist of 64 unique forms. When this agent becomes an adult agent in the following generation it will be able to generate all of the 64 forms and them only. I chose the bottleneck arbitrarily to be 80% of the forms the adult agent can generate, so the first adult agent can pass to the next generation 51 unique forms. As explained in 2.3.3, the size of the data is used to determine the cost of noise-correction so it would be helpful if the data passed between every two generations is of the same size. In this case where the data passed to the next generation consist of 51 forms, we can randomly choose 13 forms and repeat them. Note that this does not mean that the number of unique forms is expected to decrease along the generations: it is sufficient for the learner to see partial data in order to generate all of the forms the adult agent was capable of generating, as long as it is exposed to every morpheme option in the adult agent's grammar. Still, in some of the simulations a stem was lost along the generations. This did not affect noise-correction as explained in section 3.1.3.2 below.

At some generation the learning agent might choose to describe the data using an optional rule as EP expects (11b-ii). In this case the learner will be able to generate

[5] This can be solved by assigning weights which match distributional observations to optional rules. This way, final devoicing would increase gradually. This kind of change to the learner is left for future work.

more than 64 forms, since the optional rule can either apply or not apply to some of the 64 original forms, creating multiple forms for the same UR. In such cases, if 80% of the forms the adult agent can generate are more than 64 forms, only 64 forms will be selected in random and passed on to the next generation.

Prior to transmitting the 64 forms to the next generation, noise will applied to them as explained next.

3.1.3 Noise

EP proposes that certain phonetic factors support the emergence of final devoicing. However, if these phonetic factors are universal and affect language transmission between generations of speakers in a similar way, we should expect final devoicing to occur in every language that has lexical items with final voiced obstruents. In this section I show how these phonetic factors can be modeled as noise, and provide a possible explanation for why these factors might be interpreted as different rates of noise in different languages.

3.1.3.1 Noise characteristics

The explanation of EP to the final (de-)voicing typology discussed in 1.3.2.2 can be summarized to the following points:

(36) **Summary of EP explanation for the final (de-)voicing typology**

a. Languages use certain gestures to mark phrase boundaries (10a, 10b), or the cue for voicing is missing in phrase-final positions (10c). This blurs the distinction between voiced and voiceless consonants in phrase-final positions, devoicing voiced consonants.

b. There are no phonetic factors which blur the distinction in the opposite direction, i.e. voicing voiceless consonants.

c. Up to 60% of the utterances that the child hears are single-word utterances.

(36a, 36c) can be modeled as the existence of noise of the form in (37a), while (36b) can be modeled as the lack of noise of the form in (37b).

(37) a. $[-son] \rightarrow [-voice]/\_\_\#$

b. $[-son] \rightarrow [+voice]/\_\_\#$

But what is the rate of the noise that should be applied between generations? First, let us figure out how many of the words are in an utterance-final position during language acquisition. In the data used by Spinelli et al. (2018) in a language acquisition experiment, the mean length of utterances (MLU) of maternal input was 2.7 words when the child was six months old, 2.48 when the child was 9 months old, and 2.78 when the child was 12 months old, averaging at 2.65. This means that on average, one of every 2.65 words is utterance-final, or 37.74% of the words the child is expected to hear. This means that the maximum possible rate of words that might be affected by the noise, if we assume that the factors in (36a) necessarily lead to final devoicing phrase-finally, is 37.74%. In our case, 37.74% of 64 words is about 24.15 words. So a noise rate of 100% might affect 24 forms at most.

However, the noise rate might be lower than 100%, affecting only some of the phrase-final forms. For example, Wang and Bilge (1973) show that under different rates of noise English speakers misperceive b, d, g, z in coda position for p, t, k, s respectively only about 4.12% of the cases, and there are differences in the confusion rates of different obstruents. In fact, it is likely that the noise rate varies between different languages. Even if all languages use the exact same phonetic gestures to mark phrase end there are reasons to believe these gestures are perceived differently in different languages, leading to different rates of noise. According to Laufer (1998), the relevant cues which mark the voiced-voiceless contrast in human languages are release duration, amount of aspiration and voice timing (VOT and Voice into Closure – ViC). For example, Laufer shows that in Hebrew the difference in VOT between voiceless and voiced stops is about 125 milliseconds, and the difference in ViC between

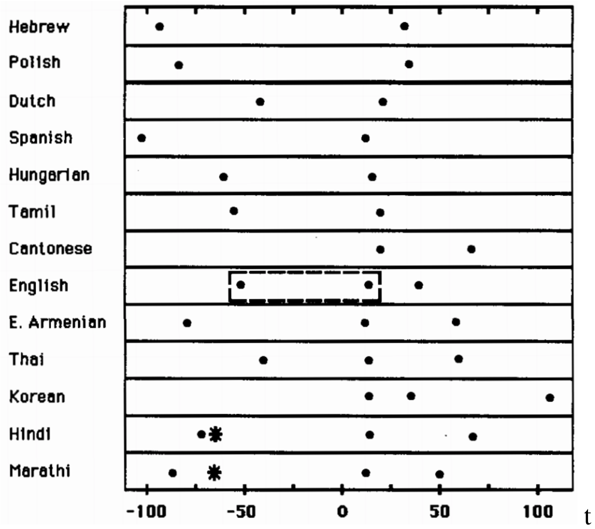


Figure 3: **VOT averages of initial stops in 13 languages**. The horizontal axis represents the VOT in milliseconds, where 0 is the time of release. Except for English, the leftmost dot in every row represents the average for voiced stops, the one after it represents voiceless stops and the one after it (if exists) represents aspirated voiceless stops. The asterisk in the last two rows represent aspirated voiced stops. For English there were two discontinuous ranges of measurements for the voiced stops – both dots in the box represent voiced stops. From Laufer (1998, p. 161).

voiced and voicless stops is about 116 milliseconds. Based on his findings along with others (Lisker and Abramson, 1964, 1971; Keating, 1984; Ladefoged, 1971), Laufer constructs Figure 3 which compares VOT averages of initial stops in 13 languages.

We see that voice timing is different in each language: for example, the VOT of voiced stops in Spanish averages around -100 milliseconds while the average VOT of voiced stops in Thai is around -50 milliseconds. This means that phonetically speaking Spanish voiced stops are "more voiced" than the voiced stops in Thai. In addition, the distinction between voiced and voiceless stops is different within each language: in Hebrew the difference in VOT between voiceless and voiced stops is about 125 milliseconds while in Cantonese it is only about 50 milliseconds. If the meaning of the phonetic factors (36a) is "add about 25 milliseconds to VOT", it will blur the distinction between voiced and voiceless stops in Cantonese but not in Hebrew. In other words, even if the articulatory noise is the same, it does not always perceived as noise. I model these differences in different rates of devoicing noise (37a), running a simulation for each of the following noise rates:

noise rate	max. number of affected forms	% of all words
100%	24	37.5%
75%	18	28.125%
50%	12	18.75%
25%	6	9.375%
4.17%	1	1.5625%

Note that I do not claim that languages which have final devoicing pattern are expected to have a small difference between voiced and voiceless VOTs. Sometimes the opposite is true: according to Figure 3, the difference between voiced and voiceless VOTs in Polish is relatively large while in Hungarian this difference is much smaller – but Polish has final devoicing while Hungarian does not. A possible explanation for this is that when final devoicing started emerging in Polish the difference between VOTs was smaller. Once the pattern emerged and it is part of language's grammar, it is expected to remain stable even if VOT differences change, as there is no counter noise which will destroy final devoicing.

3.1.3.2 Noise-correction

As mentioned in section 2.3.3, noise-correction rules should be general in order to avoid any bias. The noise-correction component of the learner in all of the simulations is limited and includes only two noise-correction rules corresponding to the noise in (37), as our base assumption is that the learner is not biased towards voicing nor devoicing. Other than final obstruent devoicing no other type of noise is applied to the data, so noise-correction rules fixing changes in other feature values is not added to the learner for computational reasons – the more rules there are in an hypothesis the more complex the calculation of its $|D{:}G|$. The grammar used to generate the initial data is a simple grammar (18): concatenation of two morphemes, and applying (zero) rules on them. The data in the following generations was expected to be described by simple grammars as well, with a similar lexicon and a (possibly empty) set of rules. This means we can use the analysis in 2.3.3 in order to determine the desired MDL cost of noise-correction rules. Let us recall the instructions of how to determine the cost of noise-correction (32), repeated in (38):

(38) In order for a simple grammar G to fix errors in a data set of size $|D|$ using noise-correction rules, the total cost of applying the rules should be:

a. less than $|D| \cdot \sum_{i=1}^{m} log_2(1 + \frac{e_i}{o_i})$
where o_i is the number of options for the i-th morpheme in G and e_i is the minimal number of morpheme options to add to this morpheme in order to deal with the exceptions (without noise-correction rules);

b. less than the total number of phonological contexts in D that matched obligatory rules in G which have to be turned optional in order to deal with the exceptions (without noise-correction rules); or, if no such obligatory rules, less than the cost of adding such optional rules plus the number of contexts as mentioned.

Let us calculate the cost for (38a) if we want to allow a single noisy form: $|D| \cdot \sum_{i=1}^{m} log_2(1 + \frac{e_i}{o_i})$. In our case, $m = 2$ since there are only stems and suffixes, and $e_1 = 1, e_2 = 0$ as our noise can affect only stems (no suffix ends with an obstruent). $|D| = 64, o_1 = 16, o_2 = 4$ so in total we get $64 \cdot (log_2(1 + \frac{1}{16}) + log_2(1 + \frac{0}{4})) = 5.59$ bits, namely a total increase of 55.9 in $|D{:}G|$. As for (38b), an optional devoicing rule can apply to the 6 forms which are generated by a concatenation of each of the voiced-obstruent-final stems with the empty morpheme. As mentioned above, the number of the binary choices made when parsing a word is multiplied by 10 when calculating $|D{:}G|$, leading to a total of 60 plus the cost of adding a rule to the grammar, which is

32.13 – i.e., a total increase of 92.13 to the MDL score.[6]

This means that if we want the learner to ignore a single error (which is 1.5625% of the forms it is exposed to) or more correctly, fix it using its devoicing-noise-correction rule, the total cost of the correction should be less than 55.9. We should be careful not to choose a cost which is too low, otherwise the learner could correct more than a single form. Generally, allowing the learner to fix up to j rules is calculated as follows: using (38) we calculate the cost of correcting $j + 1$ errors and ensure the cost of noise-correction is higher than this number. In our case, the increase in $|D{:}G|$ of correcting two errors should be less than 46.065 for each error. I arbitrarily chose a cost of 5.3 for the noise rule which increases $|D{:}G|$ in 53 each time the rule applies, allowing the learner to deal with a single noisy form in the data and no more. Since we assume the learner is not biased towards the existence of voice feature nor the lack of it, this is also the cost of the voicing noise-correction rule.

Along the generations of each one of the 25% and 100% noise rate simulations, a single stem was lost: bop and zooz respectively. They both end with an obstruent, but it did not affect the learner's ability to handle a single noisy form and no more with our noise-correction cost of 5.3.[7] In addition, in the 100% noise rate simulation the stem /aap/ was induced instead of /aab/ in the last two generations, but this happened after final devoicing of obstruents had emerged.

[6] Theoretically, the learner could choose an optional rule whose left context is more specific thus affecting only some of the 12 forms. However, even if the learner chooses this option, such a more specific rule will probably be useful only for a small number of generations as the devoicing noise applies to a wider context, affecting forms which the rule does not apply to in the following generations.

[7] This was calculated as follows: first, the cost of adding a single morpheme option (38a) was 5.96, or 59.6 to $|D{:}G|$, which is higher of our cost of noise-correction. Second, if the lost stem ended with an obstruent, an optional rule that could affect the voicedness of this morpheme (38b) had less contexts it could apply to, lowering the cost of this option to 82.13, which is higher than the cost of noise-correction as well. The MDL cost of fixing two forms according to (38) in these cases is 41.065 to each one of the forms, which means the learner still could fix only a single form using noise-correction rules.

3.1.4 Simulation results

Some level of final-devoicing emerged in all simulations except for the 4.17% noise rate simulation. A pattern of obligatory final devoicing of all obstruents emerged in generation 13 of the 75% noise rate simulation and in generation 15 of the 100% noise rate simulation, and optional final devoicing of all obstruents emerged in both 25% and 50% noise rate simulations. Simulation results are summarized in Figure 4.

A higher noise rate was expected to lead to a faster emergence of final devoicing. In the presented simulations, however, obligatory final devoicing had emerged in the 75% noise rate simulation two generations before it emerged in the 100% noise rate simulation. But this is not very interesting: the forms which were transmitted to the following generation were selected at random, as well as the ones that were affected by the noise. Namely, it was likely that final devoicing would emerge faster in the 100% noise rate simulation than in the 75% one, but it was not guaranteed.

In two generations of the 75% noise rate simulation and in one generation of the 100% noise rate simulation the learner chose to represent a partial final voicing pattern. Let us analyze one of these cases. The data that was fed to the learner in the twelfth generation of the 100% noise simulation was:

> aab, aabbam, aabbam, aabon, aabon, aap, at, at, atbam, atbam, atka, aton, aton, bopbam, bopka, dambam, damka, damon, doa, doa, doabam, doaka, doaka, doaon, gaak, gaakka, gaakon, gaakon, koasbam, koaska, kos, kozbam, kozka, maa, maabam, maaka, maaka, moon, moonbam, moonka, moonon, nagbam, nagka, nagka, nagon, nak, oak, oakbam, oakka, oakon, padbam, padka, pat, pat, sab, sabbam, sabka, sabka, sabon, sap, taot, taotbam, taotka, taoton

The only two forms in the data with final voiced obstruents are [aab] and [sab], both end with the labial obstruent [b]. There are two other forms that end with a labial obstruent, [aap] and [sap], which allowed the learner to describe the data using the

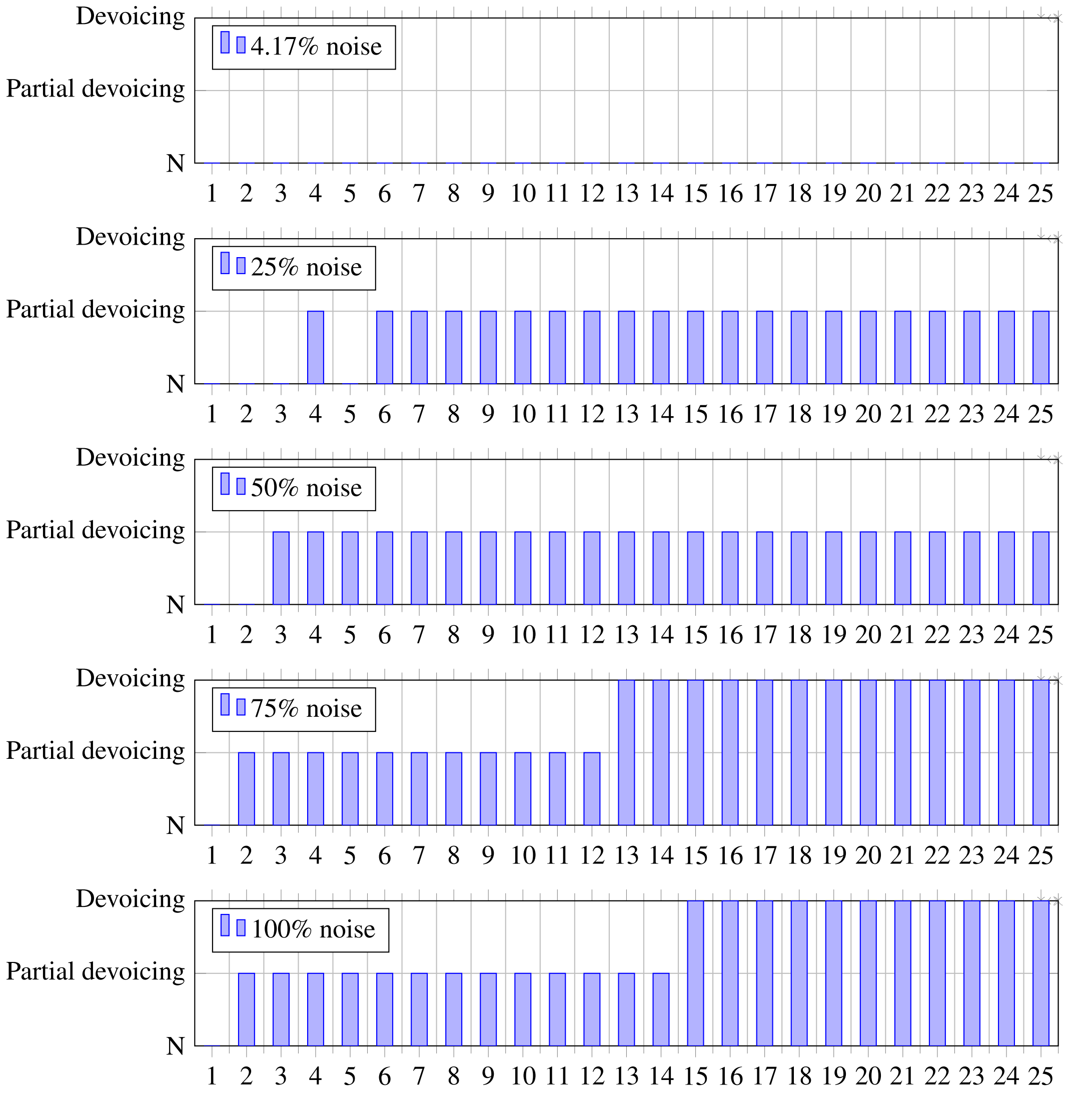


Figure 4: **Emergence of final devoicing from (de-)voicing neutral environment.**
The horizontal axis shows generation number.
N: No devoicing pattern.
Partial devoicing: pattern is (i) optional; or (ii) applies only in limited left context (e.g. only after [+*low*] segments); or (iii) applies only for some of the voiced obstruents (e.g. only [+*coronal*]); or (iv) any combinations of these.
The graph is missing partial voicing pattern learned in generations 7–8 of the 75% noise simulation and in generation 12 of the 100% simulation.

following rule system:

a. $[-son] \rightarrow [-voice]/\_\_\#$

b. $[+labial] \rightarrow [+voice]/\_\_\#$ (*optional*)

The relevant stems for these four forms are $/\mathrm{aab}/$, $/\mathrm{sab}/$ (as can be induced from forms like $[\mathrm{aabon}]$, $[\mathrm{sabka}]$). When these stems are concatenated with the null suffix, they are turned to aap, sap by the first phonological rule which is obligatory. These are the only labial-final forms the learner can generate which means the second rule may apply only to them, allowing the learner to generate the forms $[\mathrm{aap}]$, $[\mathrm{sap}]$ by not applying the optional rule, and the forms $[\mathrm{aab}]$, $[\mathrm{sab}]$ by a "Duke-of-York" derivation, reverting the devoicing of the first rule. These are the only four cases when the learner needs to choose whether to use the optional rule in order to describe the data, hence the rule's contribution to $|D{:}G|$ is four bits. Using noise correction was not an option for the learner here, because the cost of noise correction allowed the learner to tolerate only one error. If the learner chose to use its voicing-noise-correction rule, correcting $[\mathrm{aab}]$ and $[\mathrm{sab}]$ so they match the devoicing pattern, it would have cost it 5.3 bits for correcting each form. Another option for the learner was to describe the data using a single phonological rule of optional final devoicing of obstruents. But this would have cost it a great deal: there is no evidence for an optional pattern in the forms $[\mathrm{pat}]$ which appeared twice and $[\mathrm{kos}]$, $[\mathrm{nak}]$ which appeared once. Namely, a limited and optional voicing rule indeed describes well the generalization that the cases where final obstruents are not devoiced are limited and exceptional. This ad-hoc solution, however, did not last more than one generation. In the following generation the learner preferred to describe the data using an optional final devoicing rule that could apply to all obstruents, and two generations later the pattern became obligatory. A similar thing happened in the seventh generation of the 75% noise rate simulation, again with final labial obstruents. This time the pattern remained for another generation but disappeared afterwards.

These simulations show how channel bias alone can account for the existence of final devoicing, explaining why the pattern is common. In the next section we will show that EP's phonetic factors are not only capable of explaining the existence of final devoicing, but can also destroy the pattern of final voicing had it emerged.

3.2 Decay of final voicing

The previous section shows how final devoicing emerges from initial data without any (de-)voicing patterns, modeling Blevins's (2006) suggestion that phonetic factors serve as a channel bias towards final devoicing. However, as mentioned in section 1.4, Kiparsky (2006) provides examples for chains of documented natural phonetic processes which should have lead to final voicing if there was no innate constraint against it according to him. While the simulations in the previous section show how final devoicing emerges due to the channel bias, they do not address Kiparsky's claims. Theoretically final voicing could emerge given the scenarios Kiparsky describes if there was no channel bias, but there is no reason to assume channel bias is not there during or after the emergence of final voicing. In this section I present simulations which model how the phonetic factors described by Blevins interact with an existing pattern of final voicing had it emerged in a scenario similar to the ones Kiparsky describes.

The method of transmitting data between generations and the noise applied to these data are similar to the ones in the simulations in the previous section, as are the rates of noise in the different simulations. The only difference is the data introduced to the learning agent in the first generation in each one of the simulations. The same lexicon was used to generate these data, but the following final voicing pattern was applied to all (relevant) forms:

$$[-son] \rightarrow [+voice] / \_\_\#$$

This means that six out of the 64 forms showed the pattern of final voicing in the data fed to the learning agent in the first generation. The pattern was learned correctly by the first learning agent in all five simulations.

3.2.1 Simulation results

In the 4.17% and 75% noise rate simulations a single morpheme was lost but as before, it did not affect the learner's capability of handling a single noisy form. In addition, in some cases stems were replaced with their final voiced/voiceless alternative, and in the 12-21 generations of the 100% noise rate simulation both stems /kos/ and /koz/ were learned. The fact that this kind of errors occurred here but not in the set of simulations in the previous section might be a result of the clash between the two opposite final (de-)voicing patterns. Other than that, a wrong segmentation of stem/suffix lead the learner to induce stems such as /mo/ instead of /moon/ or /oakka/ instead of /oak/. It is not clear why it did not happen in the previous set of simulations.

In three out of the five simulations a pattern of obligatory final devoicing of all obstruents emerged: the 50%, 75%, and 100% noise rate simulations, at generations 19, 24 and 21 respectively. A decay of the final voicing pattern was evident in all simulations by the fifth generation, including the 4.17% noise rate simulation. In the 25% and 50% noise rate simulations the voicing pattern decayed completely before a pattern of final devoicing showed up, while all other simulations showed a mixed pattern of partial voicing and devoicing at some point. Partial could mean the pattern was optional, or it occurred only in a limited left context (e.g. only after [+*low*] segments) or it occurred only for some of the obstruents (e.g. only [+*coronal*]), or any combination of these. The results are shown in Figure 5. It is interesting to see that even a noise rate of 4.17% managed to destroy the full final voicing pattern, as along the generations some obstruents did not appear voiced word-finally, making it more worthwhile for the learner to assume the voicing pattern does not apply to them. In addition, partial final devoicing emerged alongside the final voicing pattern in this simulation while it did

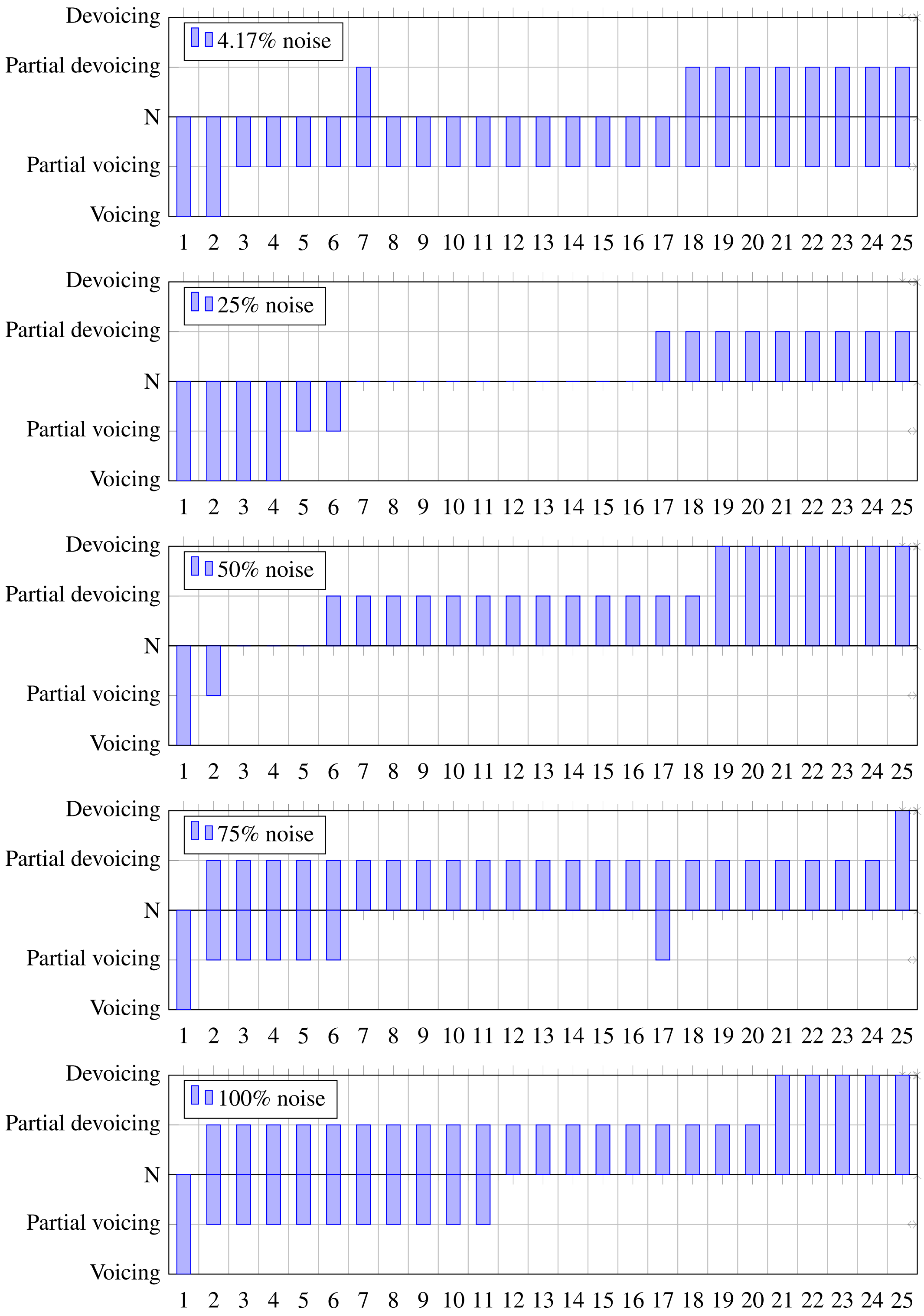


Figure 5: **Decay of final voicing.**
The horizontal axis shows generation number.
N: No (de-)voicing pattern.
Partial (de-)voicing: pattern is (i) optional; or (ii) applies only in limited left context (e.g. only after [+*low*] segments); or (iii) applies only for some of the obstruents (e.g. only [+*coronal*]); or (iv) any combinations of these.

not happen in a simulation with the same noise rate and a neutral (de-)voicing environment. It seems like the contrast between the noise and the voicing pattern made the noise more prominent. In other words, not only final voicing is not stable under EP's phonetic factors, but in some cases it might even enhance the effects of these factors. The results of these simulations suggest that even if final voicing had emerged in scenarios such as the ones Kiparsky describes, the pattern is not expected to last. This might be one of the reasons that the pattern is not attested.

Chapter 4

Discussion

In this work I surveyed how different theories deal with asymmetries in phonological typology. While some theories attribute key aspects of the typological asymmetry to analytic bias, such as OT with innate constraints, according to theories such as EP the asymmetry stems in channel bias. I presented a computational model of EP which can put the theory's claims to test by simulating the emergence of phonological patterns, thus helping us to figure out the division of labor between analytic bias and channel bias in shaping phonological typology. I suggested a noise-correction mechanism in order to make sure that only noise which is consistent and frequent enough would lead to sound change. Using this model I showed that patterns such as final devoicing can arise simply due to consistent noise, as at some point the learners might interpret the noise as part of the language's generalizations. In addition, I showed that even if natural processes would lead to the emergence of final voicing, channel bias would destroy this pattern, which might be one of the reasons that the pattern is not attested.

In some cases innate natural constraints lead to predictions that do not match typological generalizations. For example, Prickett (2017) shows that Stratal OT with innate natural constraints can generate word-final voicing. Such evidence, along with the presented model which can help account for typological generalizations via channel bias, opens the door to theories which do not build the typology into analytic bias. An example for a direction worth exploring is learnability of constraints, or at of least some of them. Rasin and Katzir (2016) show that an MDL learner can induce constraints from general constraint schemata, and Moreton et al. (2015) and Hayes and Wilson (2008) show that phonotactic knowledge can be induced with minimal restrictions on constraints.

There is more work to be done in the future. Kiparsky (2006) mentions scenarios which, according to him, were expected to give rise to final voicing if there was no innate constraint against it. The results of the final devoicing decay simulations I presented in section 3.2 show that final voicing would die out quickly if it emerged, but it is interesting to see what happens before the pattern's emergence. We can imagine a situation where Kiparsky's scenarios occur over and over, continuously interfering with EP's phonetic factors. Such cases should be examined in order to properly address Kiparsky's claims.

In addition, in this work I presented the emergence of a specific typological asymmetry, and it is interesting to see what the model can tell us about other types of typological generalizations such as prosodic typology – for example, Jakobsonian syllable structure mentioned in 1.3.1 or stress patterns.

It also has to be noted that the bottleneck I chose as well as the rates of noise were somewhat arbitrary. In addition, instead of tampering with the data transmitted between generations in order to ensure the rate of allowed noise is similar, one might consider to determine the cost of noise correction dynamically, according to the size of the data. It is interesting to see how changing these parameters might affect the emergence of phonological patterns.

While these issues need to be addressed, the model I presented can help us to test whether certain typological asymmetries can emerge without analytic bias, which opens the door to theories which attribute less of the typology to analytic bias and more of it to channel bias.

Appendix A

Grammars induced in the simulations

This appendix describes the grammar induced in each one of the generations in the different simulations. Only missing/added stems are mentioned because all suffixes where induced correctly in all generations of all simulations. Note that an obligatory rule followed by a partial rule of the opposite pattern was treated as a partial pattern in my analysis of simulation results.

A.1 Neutral initial environment

4.17% noise rate

generations	**phonological rule set**
1-25	∅

25% noise rate

generations	**missing stems**	**phonological rule set**

1		$\emptyset$
2		$\emptyset$
3	bop	$\emptyset$
4	bop	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/[-\mathit{low}]\_\_\#$ (optional)
5	bop	$\emptyset$
6-12	bop	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/[-\mathit{low}]\_\_\#$
13	bop	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
14	bop	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/[+\mathit{low}]\_\_\#$ (optional)
15	bop	$[+\mathit{labial}, -\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
16-25	bop	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)

50% noise rate

generations	**phonological rule set**
1-2	$\emptyset$
3	$[+\mathit{coronal}, -\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
4-17	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
18	$[-\mathit{labial}, -\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
19-25	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)

75% noise rate

generations	**missing stems**	**phonological rule set**
1		$\emptyset$
2-5		$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
6		$[-\mathit{labial}, -\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ $[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)

7		[−*son*] → [−*voice*]/__# [+*labial*] → [+*voice*]/__# (optional)
8		[−*son*] → [−*voice*]/__# [+*labial*] → [+*voice*]/[+*voice*]__# (optional)
9		[−*son*] → [−*voice*]/__# (optional)
10-11		[+*coronal*, −*son*] → [−*voice*]/__#
12-13		[−*labial*, −*son*] → [−*voice*]/__#
12-13		[−*labial*, −*son*] → [−*voice*]/__#
14-25		[−*son*] → [−*voice*]/__#

100% noise rate

generations	**missing stems**	**phonological rule set**
1		∅
2-8		[−*son*] → [−*voice*]/__# (optional)
9-10	zooz	[−*son*] → [−*voice*]/[+*low*]__# (optional)
11	zooz	[−*son*] → [−*voice*]/__# (optional)
12	zooz	[−*son*] → [−*voice*]/__# [+*labial*] → [+*voice*]/__# (optional)
13-14	zooz	[−*son*] → [−*voice*]/__# (optional)
15-23	zooz	[−*son*] → [−*voice*]/__#
24-25	zooz, aab (replaced with aap)	[−*son*] → [−*voice*]/__#

A.2 Final voicing initial environment

4.17% noise rate

generations	missing stems	phonological rule set
1-2		[−*son*] → [+*voice*]/__#
3	doa	[−*son*] → [+*voice*]/__#
4	doa	[−*cont*] → [+*voice*]/__#
5-17	doa	[−*coronal*] → [+*voice*]/__#
18-24	doa	[−*son*] → [−*voice*]/__# [−*coronal*] → [+*voice*]/__#
25	doa	[−*son*] → [−*voice*]/__# [−*coronal*] → [+*voice*]/[+*low*]__#

25% noise rate

generations	missing stems	phonological rule set
1-4		[−*voice*] → [+*voice*]/__#
5	gaak (replaced with gaag)	[−*cont*] → [+*voice*]/__#
6	gaak (replaced with gaag)	[−*son*] → [+*voice*]/[−*low*]__#
7-16	gaak (replaced with gaag)	∅
17-25	gaak (replaced with gaag)	[−*son*] → [−*voice*]/__# (optional)

50% noise rate

generations	missing stems	phonological rule set
1		[−*son*] → [+*voice*]/__#
2		[−*coronal*] → [+*voice*]/__# [+*cont*] → [+*voice*]/__#
3-5	nag (replaced with nak)	∅
6-18	nag (replaced with nak)	[−*son*] → [−*voice*]/__# (optional)

19-23	nag (replaced with nak)	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$
24	koz (replaced with kos) nag (replaced with nak)	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$
25	koz (replaced with kos) nag (replaced with nak) zooz (replaced with zoos)	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$

75% noise rate

generations	missing stems	phonological rule set
1		$[-\mathit{low}] \rightarrow [+\mathit{voice}]/\_\_\#$
2		$[+\mathit{coronal}, -\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ $[-\mathit{voice}] \rightarrow [+\mathit{voice}]/\_\_\#$ (optional)
3		$[-\mathit{labial}] \rightarrow [+\mathit{voice}]/\_\_\#$ $[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
4		$[+\mathit{coronal}] \rightarrow [+\mathit{voice}]/\_\_\#$ $[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
5		$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ $[+\mathit{coronal}] \rightarrow [+\mathit{voice}]/\_\_\#$ (optional)
6		$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional) $[+\mathit{cont}] \rightarrow [+\mathit{voice}]/[+\mathit{low}]\_\_\#$
7		$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
8-12	at	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
13	at	$[+\mathit{coronal}, -\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ $[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
14-16	at	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$ (optional)
17	at	$[-\mathit{son}] \rightarrow [-\mathit{voice}]/\_\_\#$

		[+*labial*] → [+*voice*]/__# (optional)
18-24	at moon (replaced with mo)	[−*son*] → [−*voice*]/__# (optional)
25	at, aab (replaced with aap), moon (replaced with mo)	[−*son*] → [−*voice*]/__#

100% noise rate

generations	missing stems	added stems	phonological rule set
1			[−*low*] → [+*voice*]/__#
2			[−*labial*, −*son*] → [−*voice*]/__# [+*coronal*] → [+*voice*]/__# (optional)
3-4			[+*coronal*] → [+*voice*]/[+*low*]__# [−*son*] → [−*voice*]/__# (optional)
5			[+*coronal*] → [+*voice*]/__# [−*son*] → [−*voice*]/__# (optional)
6			[+*coronal*] → [+*voice*]/[+*low*]__# [−*son*] → [−*voice*]/__# (optional)
7			[−*son*] → [−*voice*]/__# [+*coronal*] → [+*voice*]/__# (optional)
8-11	oak (replaced with oakka)		[−*son*] → [−*voice*]/__# [+*cont*] → [+*voice*]/__# (optional)
12-20	oak (replaced with oakka)	kos (duplicates koz)	[−*son*] → [−*voice*]/[+*low*]__#
21	oak (replaced with oakka)	kos (duplicates koz)	[−*son*] → [−*voice*]/__#
22	oak (replaced with oakka)		[−*son*] → [−*voice*]/__#
23-25	oak (replaced with oakka),		[−*son*] → [−*voice*]/__#

	koz (replaced with kos)		

Bibliography

Anttila, Arto, and Giorgio Magri. 2018. Does MaxEnt Overgenerate? Implicational Universals in Maximum Entropy Grammar. In *AMP 2017: Proceedings of the 2017 Annual Meeting on Phonology*, ed. Gillian Gallagher, Maria Gouskova, and Sora Yin, 1–12.

Berwick, Robert C. 1982. Locality principles and the acquisition of syntactic knowledge. Doctoral Dissertation, MIT, Cambridge, MA.

Berwick, Robert C. 1985. *The acquisition of syntactic knowledge*. Cambridge, Massachusetts: MIT Press.

Blevins, Juliette. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

Blevins, Juliette. 2006. A theoretical synopsis of evolutionary phonology. *Theoretical Linguistics* 32(2):117–166.

Blevins, Juliette. 2008. Consonant epenthesis: Natural and unnatural histories. In *Linguistic universals and language change*, ed. Jeff Good, 79–107. Oxford: Oxford University Press.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21:43–58.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45–86.

Brochhagen, Thomas, Michael Franke, and Robert van Rooij. 2018. Coevolution of lexical meaning and pragmatic use. *Cognitive Science* 42:2757–2789.

Browman, Catherine P., and Louis M. Goldstein. 1995. Gestural syllable position effects in American English. In *Producing speech: Contemporary issues*, ed. Fredericka Bell-Berti and Lawrence J. Raphael, 19–33. Woodbury, NY: AIP Press.

Byrd, Dani. 1996. A phase window framework for articulatory timing. *Phonology* 13:139–169.

Chaitin, Gregory J. 1966. On the length of programs for computing finite binary sequences. *Journal of the ACM* 13:547–569.

Cho, Young-mee Yu. 1990. Syntax and phrasing in Korean. In *The Phonology-Syntax Connection*, ed. Sharon Inkelas and Draga Zec, 47–62. Chicago: University of Chicago Press.

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

Content, Alain, Ruth K. Kearns, and Uli H. Frauenfelder. 2001. Boundaries versus onsets in syllabic segmentation. *Journal of Memory and Language* 45:177–199.

Coon, Jessica. 2010. VOS as Predicate-fronting in Chol. *Lingua* 120:345–378.

Dell, François. 1981. On the learnability of optional phonological rules. *Linguistic Inquiry* 12:31–37.

Easterday, Shelece. 2019. *Highly complex syllable structure: A typological and diachronic study*. Berlin: Language Science Press.

Gick, Bryan, Fiona Campbell, Sunyoung Oh, and Linda Tamburri-Watt. 2006. Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids. *Journal of Phonetics* 35:49–72.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Stockholm Workshop on Variation Within Optimality Theory*, ed. Jennifer Spenader, Anders Eriksson, and Östen Dahl, 111–120. Stockholm University.

Gordon, Matthew Kelly. 2016. *Phonological Typology*. Oxford University Press.

Greenberg, Joseph H. 1963. *Universals of Language*. MIT Press, Cambridge, MA.

Guerin, Francoise. 2001. *Description de l'ingouche: Parler du centre nord du caucase*. Munich: Lincom Europa.

Halle, Moris. 1959. *The sound pattern of Russian*. The Hague: Mouton.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.

Holes, Clive. 1990. *Gulf Arabic*. London: Croom Helm.

Horn, Laurence. 1972. On the Semantic Properties of the Logical Operators in English. Doctoral Dissertation, UCLA.

Horning, James. 1969. A study of grammatical inference. Doctoral Dissertation, Stanford.

Hyman, Larry M. 1970. How concrete is phonology? *Language* 46(1):58–76.

Jakobson, Roman. 1962. *Selected writings 1: phonological studies*. The Hague: Mouton.

Katzir, Roni. 2014. A cognitively plausible model for grammar induction. *Journal of Language Modelling* 2:213–248.

Kawasaki-Fukumori, Haruko. 1992. An acoustical basis for universal phonotactic constraints. *Language and Speech* 35:73–86.

Keating, Patricia A. 1984. Phonetic and phonological representation of stop consonant voicing. *Language* 60:286–319.

Kenstowicz, Michael, and Charles Kisseberth. 1979. *Generative phonology: Description and theory*. New York: Academic Press.

Kessar, Sara, and Radwan S. Mahadin. 2020. An Optimality analysis of the phonology of French loanwords as manifested in the eastern part of Algeria. *International Journal of Linguistics* 12:171–185.

Kiparsky, Paul. 1968. *How abstract is phonology?*. Indiana University Linguistics

Club.

Kiparsky, Paul. 2003. Finnish noun inflection. In *Generative Approaches to Finnic Linguistics*, ed. Diane Nelson and Satu Manninen, 109–161. CSLI, Stanford.

Kiparsky, Paul. 2006. The amphichronic program vs. evolutionary phonology. *Theoretical Linguistics* 32(2):217–236.

Kiparsky, Paul. 2008. *Universals constrain change; change results in typological generalizations*, 23–53. Oxford: Oxford University Press.

Kirby, Simon. 2000. Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In *The evolutionary emergence of language: Social function and the origins of linguistic form*, ed. Michael Studdert-Kennedy Chris Knight and James R. Jurford, 303–323. Cambridge: Cambridge University Press.

Kirby, Simon. 2001. Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation* 5(2):102–110.

Kirby, Simon. 2002. Learning, bottlenecks and the evolution of recursive syntax. In *Linguistic evolution through language acquisition: Formal and computational models*, ed. Ted Briscoe, 173–203. Cambridge: Cambridge University Press.

Kirby, Simon, Kenny Smith, and Henry Brighton. 2004. From UG to universals: linguistic adaptation through iterated learning. *Studies in Language* 28:587–607.

Kolmogorov, Andrei Nikolaevic. 1965. Three approaches to the quantitative definition of information. *Problems of Information Transmission (Problemy Peredachi Informatsii)* 1:1–7.

Ladefoged, Peter. 1971. *Preliminaries to linguistic phonetics*. Chicago: University of Chicago Press.

Lan, Nur. 2018. Learning morpho-phonology using the Minimum Description Length principle and a genetic algorithm. Master's thesis, Tel Aviv University.

Laufer, Asher. 1998. Voicing in contemporary Hebrew in comparison with other languages. *Hebrew Studies* 39:143–179.

Lisker, Leigh, and Arthur S. Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20:384–422.

Lisker, Leigh, and Arthur S. Abramson. 1971. Distinctive features and laryngeal control. *Language* 47:767–785.

Marin, Stefania, and Marianne Pouplier. 2010. Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. *Motor Control* 14:380–407.

McCawley, James D. 1970. English as a VSO language. *Language* 46(2):286–299.

Moreton, Eliot. 2008. Analytic bias and phonological typology. *Phonology* 25:83–128.

Moreton, Elliott, Joe Pater, and Katya Pertsova. 2015. Phonological concept learning. *Cognitive Science* 41:4–69.

Mous, Marten. 1993. *A grammar of Iraqw*. Hamburg.

Niyogi, Partha, and Robert C. Berwick. 2009. The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences* 106:10124–10129.

Ohala, John J. 1997. Aerodynamics of phonology. In *Proceedings of the Seoul International Conference on Linguistics*, 92–97. Seoul: Linguistic Society of Korea.

Orbán, Gergő, József Fiser, Richard N Aslin, and Máté Lengyel. 2008. Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences* 105:2745–2750.

Polinskaja, Maria S. 1989. Object initiality: OSV. *Linguistics* 27:257–303.

Prickett, Brandon. 2017. Post-nasal devoicing as opacity: A problem for natural constraints. In *Proceedings of the 35th West Coast Conference on Formal Linguistics*, ed. Wm. G. Bennett, Lindsay Hracs, and Dennis Ryan Storoshenko, 331–340. Somerville, MA: Cascadilla Proceedings Project.

Prince, Alan, and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers University, Center for Cognitive Science.

Rasin, Ezer, Iddo Berger, Nur Lan, and Roni Katzir. 2018. Learning phonological optionality and opacity from distributional evidence. In *Proceedings of NELS 48*, ed. Sherry Hucklebridge and Max Nelson, 269–282. Amherst, MA: GLSA.

Rasin, Ezer, and Roni Katzir. 2016. On evaluation metrics in Optimality Theory. *Linguistic Inquiry* 47:235–282.

Rasin, Ezer, and Roni Katzir. 2020. A conditional learnability argument for constraints on underlying representations. *Journal of Linguistics* 1–29.

Rasin, Ezer, Itamar Shefi, and Roni Katzir. 2020. A unified approach to several learning challenges in phonology. In *Proceedings of NELS 50*, ed. Mariam Asatryan, Yixiao Song, and Ayana Whitmal, volume 3, 73–87. Amherst, MA: GLSA.

Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Rissanen, Jorma, and Eric Sven Ristad. 1994. Language acquisition in the MDL framework. In *Language computations: DIMACS Workshop on Human Language, March 20-22, 1992*, 149. Amer Mathematical Society.

Round, Erich. 2011. Word final phonology in Lardil: Implications of an expanded data set. *Australian Journal of Linguistics* 31:327–350. URL `https://doi.org/10.1080/07268602.2011.598630`.

Sapir, Edward. 1915. Notes on Judeo-German phonology. In *Selected writings of Edward Sapir*, ed. David Mandelbaum, 252–272. Berkeley: University of California Press.

Segui, Juan, Emmanuel Dupoux, and Jacques Mehler. 1991. The role of the syllable in speech segmentation, phoneme identification, and lexical access. In *Cognitive models of speech processing*, ed. Gerry T. M. Altmann, 263–280. Cambridge, MA:

MIT Press.

Smith, Kenny, Simon Kirby, and Henry Brighton. 2003. Iterated learning: A framework for the emergence of language. *Artificial Life* 9:371–386.

Sobel, David M., Joshua B. Tenenbaum, and Alison Gopnik. 2004. Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive Science* 28:303–333.

Solomonoff, Ray J. 1964. A formal theory of inductive inference, parts I and II. *Information and Control* 7:1–22, 224–254.

Spinelli, Maria, Mirco Fasolo, Prachi E. Shah, Giuliana Genovese, and Tiziana Aureli. 2018. The influence of early temperament on language development: The moderating role of maternal input. *Frontiers in Psychology* 9:1527.

Steriade, Donca. 1999. Phonetics in phonology: The case of laryngeal neutralization. In *Papers in Phonology 3 (UCLA Working Papers in Linguistics 2)*, ed. Matthew Gordon, 25–145. Los Angeles: Department of Linguistics, University of California.

Steriade, Donca. 2009. The Phonology of Perceptibility Effects: the P-map and its consequences for constraint organization. In *The Nature of the Word: Studies in Honor of Paul Kiparsky*, ed. Kristin Hanson and Sharon Inkelas, 151–179. MIT Press.

Trubetzkoy, Nikolai S. 1939. *Grundzüge der phonologie*. Göttingen: Vandenhoeck and Ruprecht.

Wang, Marilyn D., and Robert C. Bilge. 1973. Consonant confusions in noise. *Journal of the Acoustical Society of America* 54:1248–1266.

Weinreich, Uriel. 1963. Four riddles in bilingual dialectology. In *American Contributions to the Fifth International Congress of Slavists*, 335–358. The Hague: Mouton.

Wexler, Kenneth, and Rita M. Manzini. 1987. Parameters and learnability in binding theory. In *Parameter setting*, ed. Thomas Roeper and Edwin Williams, 41–76. Dordrecht, The Netherlands: D. Reidel Publishing Company.

Xu, Fei, and Joshua B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological review* 114:245–272.

Yu, Alan C. L. 2004. Explaining final obstruent voicing in Lezgian: Phonetics and history. *Language* 80:73–97.

Zeltner, Jean-Claude, and Henri Tourneaux. 1986. *L'arabe dans le bassin du tchad: Le parler des ulad eli*. Paris: Karthala.

תוכן עניינים

תקציר

הטיפולוגיה הפונולוגית מוטית מאוד. למשל, בעוד הדפוס של ביטול קוליות סופית של חוסמים מופיע בשפות רבות, לזגית היא השפה המתועדת היחידה בה נטען כי מופיע הדפוס ההפוך, הוספת קוליות סופית (Yu, 2004). אחת הגישות להסבר של אסימטריות כאלו היא לייחס אותן ל־Analytic bias, הטיות קוגניטיביות שמקלות על למידה של דפוסים פונולוגיים מסוימים על פני אחרים, כפי שעושה למשל תיאוריית האופטימליות (Optimality Theory - OT; Prince and Smolensky, 1993). גישה אחרת היא לייחס את מקור האסימטריות ל־Channel bias, שגיאות שיטתיות שחוזרות על עצמן ודוחפות שפות לעבר דפוס מסוים והרחק מהדפוס ההפוך לו, כפי שעושה תיאוריית הפונולוגיה האבולוציונית (Evolutionary Phonology - EP; Blevins, 2004).

חלוקת העבודה בין Analytic bias לבין Channel bias היא שאלה אמפירית. בעבודה זו אני מציג מודל ל־Channel bias שיכול לסייע לנו לספק הסברים לגבי חלוקת עבודה זו, על־ידי בחינה האם אסימטריות טיפולוגיות בפונולוגיה יכולות להתהוות כתוצאה מהעברה של ידע פונולוגי בין דורות, תוך התמקדות באסימטריה בין הוספת קוליות סופית לבין ביטול קולית סופית כמקרה בוחן. המודל שלי מבוסס על מודל למידה חזרתית של העברת שפה (Iterated Learning Model - ILM; Kirby 2001, 2002) וכולל השחתה של המידע בדמות רעש המדמה את ה־Channel bias שתואר על ידי Blevins (2006). הסוכן הלומד במודל שלי הוא לומד אורך־תיאור מזערי (Minimum Description Length - MDL; Rissanen, 1978) של Rasin et al. (2018) ששיניתי על־מנת שיוכל להתמודד עם כמויות מסוימות של רעש שכזה. אני מראה כיצד המודל מצליח לדמות את ההתהוות של האסימטריה של הוספת/ביטול קוליות סופית מנקודת התחלה ניטרלית. כמו־כן, אני מראה שהמודל יכול לדמות דעיכה של הוספת קוליות סופית, מה שמחליש את הטענות של Kiparsky (2006) נגד EP, לפיהן התבנית של הוספת קוליות סופית צפויה להיות נפוצה יותר אלמלא היו אילוצים מולדים נגד התבנית. ההצלחה של המודל שלי לדמות את האסימטריה הפונולוגית הזו פותחת את הדלת לתיאוריות שמייחסות חלקים פחות משמעותיים של טיפולוגיה פונולוגית ל־Analytic bias ומעדיפות לייחס אותם ל־Channel bias.

אוניברסיטת תל אביב
TEL AVIV UNIVERSITY

הפקולטה למדעי הרוח ע״ש לסטר וסאלי אנטין
בית הספר לפילוסופיה, בלשנות ולימודי מדע
החוג לבלשנות

אבולוציה של טיפולוגיה פונולוגית: מודל למידה חזרתית של התהוות תבניות פונולוגיות

חיבור זה הוגש כעבודת גמר לקראת התואר
״מוסמך אוניברסיטה״ - M.A. באוניברסיטת ת״א

על־ידי
איתמר שפי

העבודה הוכנה בהדרכת:
ד״ר רוני קציר
ד״ר עזר ראסין

דצמבר 2020