UNIVERSITY OF CALIFORNIA

Los Angeles

Studies in morphophonological copying: Analysis, experimentation and modeling

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Linguistics

by

Yang Wang

2024

ABSTRACT OF THE DISSERTATION

Studies in morphophonological copying: Analysis, experimentation and modeling

by

Yang Wang
Doctor of Philosophy in Linguistics
University of California, Los Angeles, 2024
Professor Timothy Hunter, Co-Chair
Professor Bruce P. Hayes, Co-Chair

Reduplication, the copying operation employed in natural language morphophonology (e.g., Ilokano pluralization, [kaldíŋ] 'goat'; [kal-kaldíŋ] 'goats'; Hayes and Abad, 1989, p. 357), creates repetition structures within surface word forms. Though reduplication and surface repetitions have been extensively studied, two questions remain unresolved. First, what are the possible natural language word forms with reduplication? Secondly, how can reduplication be characterized and learned in a unified way with other (morpho)phonological regularities? This dissertation approaches these questions through three studies that combine experimental and computational methods with previous typological studies and phonological theory.

Chapter 2 offers experimental evidence on how human learners tend to generalize reduplicative patterns. Two series of artificial grammar learning experiments using the *poverty of the stimulus paradigm* (Wilson, 2006) yield the following results. First, human learners rapidly extrapolate and generalize reduplicative hypotheses to novel forms after being exposed to only a small number of familiarized forms. Their generalizations align with coarse-grained phonological abstractions characterizable by the vocabulary of prosody (e.g., syllables, feet, prosodic words), supporting the core claims of Prosodic Morphology (McCarthy

and Prince, 1986, *et seq.*). Moreover, there are strong correlations between participants' spontaneous responses and naturally occurring reduplicative patterns. The universally preferred patterns followed typological trends, while variations in individually learned grammar reflected the variations attested in natural languages. Lastly, patterns whose empirical status has been controversial appear in participants' spontaneous responses, offering novel learning-based evidence to support Base-Reduplicant Correspondence Theory (McCarthy and Prince, 1995) as a possible characterization of human learners' hypothesis space.

Chapter 3 examines the abstract properties of surface repetition structures from a formal-language-theoretic view. We revisit the Chomsky Hierarchy, which offers highly abstract characterizations of linguistic processes. Reduplicative patterns with unbounded copying impose a challenge: a model within the classical Chomsky Hierarchy that adequately captures unbounded copying is expected to generate unattested palindrome patterns (e.g., pseudo-Ilokano with reversal, [ŋidlak-kaldíŋ]), which does not match empirical observations. Therefore, we advocate for another language class that cross-cuts the well-known classes in the classical Chomsky Hierarchy. We introduce Finite-state Buffered Machines, an augmentation to the regular class with a primitive copying operation. This is achieved by adding compact memory allocation machinery and an unbounded memory buffer with queue-like storage. We survey the properties of the resulting language class and find this refinement better matches the language typology without sacrificing mathematical rigor.

Chapter 4 proposes a morphophonological learner that extends an expectation-driven maximum entropy lexicon learner proposed by Wang and Hayes (resubmitted) with a component that deals with reduplication learning. Given that the empirical results in Chapter 2 support Base-Reduplicant Correspondence Theory (McCarthy and Prince, 1995), the learner adopts constraints proposed by this theory and learns "hidden structures" (Tesar and Smolensky, 1998), i.e. the prosodic templates. We demonstrate that in this way, the learner can learn different types of reduplication-phonology interactions and capture the population-level results observed in the learning experiments.

The dissertation of Yang Wang is approved.

Kie Ross Zuraw

Colin Wilson

Claire Moore-Cantwell

Bruce P. Hayes, Committee Co-Chair

Timothy Hunter, Committee Co-Chair

University of California, Los Angeles

2024

献给王素贞与帅振全

For Suzhen Wang and Zhenquan Shuai

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support I received from my advisors, teachers, and colleagues. I would like to start by thanking my whole committee for the countless hours they dedicated to helping me refine and polish my ideas.

My deepest gratitude belongs to my co-chairs, Tim and Bruce. Tim pretty much shaped me as a researcher and a linguist. I first met Tim during his undergraduate class in computational linguistics. Once, I dropped in his office hours and never thought a question of currying and uncurrying could lead to a two-hour discussion – now I realize Tim is so knowledgeable that anyone can have a conversation with him for hours on anything. Tim has a knack for presenting questions in their most exciting and comprehensible format, and I quickly became one of the frequent attendees of his office hours. One time, Tim sketched the Chomsky Hierarchy on a piece of paper, explained the motivations behind it in clear, layman-friendly terms, and asked me with great enthusiasm, "Isn't this cool?" I can't recall how I responded back then, but Tim, yes, it still amazes me! Tim recommended Stabler (2004) as a fun read, which turned out to be the first linguistic paper I read closely and, in turn, stimulated my writing sample for grad school application, MA thesis, and the early ideas that underpin this dissertation. I cannot say enough good words about Tim: intelligent (proof: learned phonology remarkably fast), open-minded, fun, supportive, caring, and patient. He was always there to listen to my incoherent, unintelligible thoughts and never commented harshly, even after I asked him to be my Reviewer #2. He was incredibly supportive during the chaotic times and respected my working habits (i.e., long-time procrastination). Thank you, Tim, for your faith in me. Its origin remains a great mystery, but its impact is undeniable.

Without Bruce, I probably would never have stepped into the field in the first place. Knowing nothing about linguistics, I took his introductory class during my first quarter at UCLA. It was Fall 2015 (time went by that fast?!). In the last lecture, Bruce drew the "boxology" as a summary: from forming an intent to the box of syntax, the box of morphology, the box of phonology, and then the execution of phonetics. "This is how we

humans communicate. It all starts with the S". Boom! Goosebumps. Afterward, I sat by one of the pillars in front of Royce Hall, sipping a hazelnut latte from a vending machine and chewing on his words. It was at that moment I realized I needed to take more linguistic classes. Bruce is not only a charming instructor but also an open-minded, diligent, and generous researcher with so much wisdom. I feel lucky and privileged to have collaborated with him, from which I learned a great deal by simply observing his research pipelines. Bruce always encouraged me whenever I doubted myself, and always offered sharp and sincere advice when I needed it most. It was also Bruce who told me to practice a principled mindset to divide work from life and maintain sustainable habits for a long career. I remember one time our scheduled meeting lasted only ten minutes because Bruce found out I was in Amsterdam and insisted I go out and enjoy the city. Without Bruce, my PhD life would never be so intellectually stimulating with that much fun. I deeply appreciate all the opportunities, guidance, critiques, and lessons Bruce offered me (sprinkled with mid-century English slang and metaphors).

I would also like to express my gratitude to my other committee members. I want to thank Colin for his expertise and support. I am constantly amazed by how knowledgeable Colin is. If academia is like movie production, Colin could win multiple Oscars for playing all the roles himself: phonologist, phonetician, experimentalist, statistician – you name it! I have always been a big fan of Colin's work, and it felt like a dream come true when he collaborated with me and kindly agreed to be on my committee. Colin treated me as an individual researcher from the beginning, offering just the right amount of help while leaving room for my growth. I also want to thank him for the hours he spent listening to my ramblings and his help in formulating my questions in a clearer way. Chapter 2 would never have existed if I hadn't had Colin on my committee.

One aspect that I appreciate about the UCLA linguistics department is the many women role models who have shown me what passion, power, and confidence really mean. Kie and Claire are two of these inspiring figures. After my first meeting with Kie, I was struck by how much information she could unpack in just 30 minutes. Kie always knew exactly the kind of help I needed, from offering empirical patterns to helping me navigate different pieces of literature. I especially want to thank Kie for teaching me how to write and how to add my personality to various statements and academic pieces, which helped shape the tone for this dissertation.

Claire, on the other hand, knows so much about English stress and modeling. Deriving linear programming equations on the transparent tabletop in her office was a lot of fun! Claire always reminds me to think more deeply about the cognitive implications of formal work. I am grateful for the questions she raised, many of which continue to make me think more and believe that there is still so much to explore in reduplication. I also want to thank Claire for all the encouragement and moral support she has provided throughout this dissertation.

Speaking of women role models, it is impossible to skip Megha Sundara and Laurel Perkins, whom I consider to be unofficial committee members. I cannot thank them enough for their advice on this dissertation, their guidance on my career, and their help during my job application process. I appreciate all the thought-provoking questions they have asked, especially the fact that these questions came *exactly* when I needed them at various stages of this dissertation. Their questions pushed me to think more precisely about what I mean, and why I am asking my research questions. I want to thank Megha for the many valuable conversations we've had over drinks and coffee, both academic and non-academic. Thanks to Laurel for her insightful comments on my experimental design and for encouraging me through some of the most challenging times.

I also want to thank the following individuals for their help in developing this dissertation. To Dylan Bumford, for serving on my MA committee, spotting problems in my proofs and reasoning and raising questions that I still don't quite have good answers for. To Ed Keenan, for introducing me to Malagasy reduplication. To Hossep Dolatian for producing excellent work with deep empirical grounding and being a constant source of inspiration. To Jeff Heinz for directing me to Dolatian and Heinz (2020), and to Jeff, Jon Rawski, Scott Nelson for inviting me to present at the LSA, where I received many valuable feedback. To Sam Zukoff for the intellectually rich proseminar on reduplication and the many inspiring discussions. To Canaan Breiss and Sam Zukoff for helping me better organize my thoughts on the faithfulness structures. To Iza Sola-Llonch for being the phonetics consultant. To my RAs, Mariana Cui, Jacob Hanna, Jenessa Lathrop, Shawdi Sani, Alexandria Zarko, and Boyi Zheng, for bearing with me during the chaotic phase of setting up my experiments, making this dissertation possible, and giving me this opportunity to learn and grow as a mentor.

I want to thank all of the people who helped me grow into the linguist I am today. Thanks to all the great teachers I have learned from, especially Sun-Ah Jun and Pat Keating. They made phonetics so approachable and intriguing to study. Their immense efforts in pedagogy and undergraduate research gave me the confidence to explore being a phonetician. Although I took a different path, their approaches to answering questions and exploring ideas continue to inspire me. I also want to thank my other linguistics professors at UCLA and beyond – Brice Roberts, Jos Tellings, Shu-hao Shih, Dylan Bumford, Ethan Poole, Jesse Harris, Eric Bakovic, Richard Futrell, Giorgio Magri, Canaan Breiss, Bruno Nicenboim – they gave me the opportunity to peek at the meatiest part of each subfield. Additionally, I want to thank my TAs Iara Mantenuto and Marju Kaps, whose teaching practices greatly influenced me and motivated me to become the kind of teacher I aspire to be. I "stole" Marju's way to structure a section and Iara's strategies for scaffolding tasks. A special shoutout goes to Iara – I must have been such a clueless kid in the undergrad Syntax II class. Iara patiently helped me organize my thoughts, and she was the very first person who taught me the nuts and bolts of making an argument.

I am proud to be a part of the UCLA linguistics community. Thanks to the first linguist I interacted closely with, Jeremy Steffman, for showing me how to be an inquisitive, calm, and diligent researcher through his actions. I also want to thank Abeer Abbas, Anissa Gladney, Tom Trigg, Kerri Devlin, Andy Xu, Connor Mayer, Canaan Breiss, Hiro Katsuda, Jinyoung Jo, Jennifer Kuo, Jahnavi Narkar, Christian Muxica, Boram Kim, Jian-Leat Siah and many others for their support and help in various aspects. Thanks to my linguist friends outside UCLA. Shout out to Hossep Dolatian for being my first conference buddy – even though we met online – the first crash space person, and my life mentor. Thanks also to Suyuan Liu, Magda Markowska, Jian Cui, Scott Nelson, Richard Wang, and Huteng Dai for making conferences something to look forward to.

Thanks to Katya Kholystova, the VP queen, the voice of infant studies, and the glue that holds us together. Katya always offers solid hugs to celebrate every milestone we have accomplished, big or small. I will always remember our walks through sculpture gardens, around campus, and to Trader Joe's, as well as the delicious coffee she treated me to. Katya, you've heard me say this so many times, but a heartfelt thank you – this dissertation would have been so much harder to write without you (and Scout!). Lily Xu, my procrastination comrade, the boba connoisseur and the greatest bartender, thank you for always offering valuable input and perspectives, for our academic and non-academic conversations, and for getting us tickets to Allianz Arena and currywurst, which checked one item off my bucket list (even though it was, sadly, a bit too late – you know what I mean). Iza Sola-Llonch, the Mandarin near-native speaker and talented designer with the most sass – chatting with you is always such a joy. I especially appreciate the surprise handmade tote bag, and I am going to miss your flan. Kiki Liu, whose vibe most resembles Faye Wong (and a soft, floating cloud) among all the people I know – thank you for walking me home, dragging me out of my apartment for food and hikes, and helping me regain my confidence when I felt lost. Huilei Wang, our beloved artistic photographer and poet, thank you for being such a calm, hardworking, disciplined, and wise figure in my life – someone I truly look up to. Thank you all, for tacos, for exploring LA, for driving me home all the time, for musicals, movies and theater, for cocktails, potlucks, Izakaya, dancing&cheesy reality show watching&nail-painting nights, and most importantly, for being my swim ring. Without any of you, I would have never survived.

Thank you to my non-linguist friends who have been with me through all these years. Jialan Ma and Ziyi Wang, for all the memories we share – night walks, video chats, clouds, trees and sunsets. I feel so fortunate to have grown alongside you and to have witnessed the hard work that has brought you both this far. I am so proud of you, and hope I can make you proud as well! Ziyan Fu and Xuqing Zhang, thank you for giving me strength from different parts of the world. Yaying Shen, thank you for bringing me your dad's tasty dishes and for singing Mongolian folk songs to me. Lu Feng, my friend of over 15 years, I'm so excited that we will finally live in the same city again and get to explore Salt Lake City together! Tianyi Zhao, thank you for the hotpots and daily joke reposts. Angel Dong and Soros Chen, thank you for the biweekly fun times in Ktown and for showing me different lifestyles.

To the greatest game designer, Keyan Zhang, who never complained when I pulled out my laptop in the foreign lands we explored, who was always ready to be "PoS"-tested on random reduplication patterns, who told silly jokes when I was down and who made sure I had food to eat when I was hungry. When I was deciding among different career paths, it was you who empowered me to choose to become a linguist, and it is your companionship that has helped me get this far. bibobobibobi.

Finally, to my grandparents, Suzhen Wang and Zhenquan Shuai, who taught me how to walk and run, how to speak and write, and how to receive and give. To my parents, Yanhua Shuai and Zhetao Wang, who taught me the power of knowledge. This dissertation is dedicated to you. 最后再次感谢姥姥姥爷，您们用朴素的行为教育我人生的是非对错，教会我阅读和学习的力量。面对得意和失意时，孙女都会牢记你们的教诲。谨以这篇毕业论文献给你们。

VITA

2021	M.A. (Linguistics), University of California, Los Angeles
2019	B.A. (Linguistics) and B.S. (Mathematics of computation) University of California, Los Angeles

PUBLICATIONS

Wang, Yang and Tim Hunter. 2023. On Regular Copying Languages. *Journal of Language Modelling*, 11 (1):1–66. https://doi.org/10.15398/jlm.v11i1.342.

Wang, Yang. 2021. Recognizing reduplicated forms: Finite-State Buffered Machines. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 177–187, Online. Association for Computational Linguistics

CHAPTER 1

Introduction

Human languages exhibit diverse ways of forming morphologically complex words. For example, English pluralization concatenates /-z/ to the end of singulars. This plural marker is realized differently in different phonological contexts, as [-z] in [dɑg-z] *'dogs'* but as [-s] in [kæt-s] *'cats'* and [-ɨz] in [glæs-ɨz] *'glasses'*. Theoretical linguists hypothesize a **concatenative** account for this process. As illustrated in Figure 1.1, the morphological grammar concatenates a unifying underlying mental representation (UR) of a stem with the UR of the plural marker, and then the phonological grammar derives the surface variants [-s], [-z] and [-ɨz] in different contexts. Many empirical studies (Berko, 1958, *et seq.*) and computational work (e.g., Cotterell et al., 2015 for English pluralization) have focused on concatenative processes.

Dog-pl	Cat-pl	Glass-pl	
/dɑg-z/	/kæt-z/	/glæs-z/	**Suffixation**
↓	↓	↓	**Phonology**
dɑg-z	kæt-s	glæs-ɨz	
	Devoicing	*Epenthesis*	

Figure 1.1: *English pluralization as a concatenative morphophonological process*

Unlike English pluralization, plural formation in many languages involves **reduplication**, the phonological realization of which relies on active copying. (1) provides illustrative examples. Dyirbal (Pama-Nyungan, Australia) exhibits *total reduplication*, with the plural form of a nominal comprised of two perfect copies of the full singular stem. In contrast, *partial reduplication* is exemplified by Ilokano (Austronesian, Philippines), where plural forms only copy up to the first post-vocalic consonant of the corresponding singular forms.

(1) Pluralization realized by reduplication

a. Dyirbal (Pama–Nyungan, Australia; Dixon, 1972, p. 242)

[midi]	‘little, small’	[midi-midi]	‘lots of little ones’
[gulgiɽi]	‘prettily painted men’	[gulgiɽi-gulgiɽi]	‘lots of prettily painted men’

b. Ilokano (Austronesian, Philippines; Hayes and Abad, 1989, p. 357)

[kaldíŋ]	‘goat’	[kal-kaldíŋ]	‘goats’
[púsa]	‘cat’	[pus-púsa]	‘cats’

In these examples, the copying operation bears a semantic meaning, namely pluralization. However, there are cases in natural languages where surface reduplicative structures appear to be *semantics-free*. That is, the identical substrings within a word form do not necessarily correspond to a mapping between lexical items while carrying a non-trivial meaning. For example, as shown in (2a), many words in Warlpiri (Pama-Nyungan, Australia) are lexically reduplicated. These word forms contain internal identical substrings, but the lexicon lacks unreduplicated counterparts.

(2) Semantics-free phonological reduplication

a. Warlpiri (Pama–Nyungan, Australia; Nash, 1980, pp. 118-119)

✗[warnpi]	[warnpiwarnpi]	‘long and slender’
✗[japarla]	[japarlajaparla]	‘chest (bone)’
✗[ngunju]	[ngunjungunju]	‘white clay – used for mourning’

b. Tagalog (Austronesian, Philippines; Zuraw, 2002, p. 398)

*[pat]	[patpát]	‘stick, piece of split bamboo (N)’
*[sag]	[sagság]	‘split, blunt, sagging, at the peak of success (A)’
*[tal], *[taː]	[táːtal]	‘wood chips, splinters, shavings (N)’

Another compelling example is Tagalog (Austronesian, Philippines) *pseudo-reduplication*, as illustrated in (2b). These pseudo-reduplicated words exhibit internal identity as well, following either the pattern of $\mathrm{C_1V_2C_3}$-$\mathrm{C_1V_2C_3}$ (*total*) or the pattern of $\mathrm{C_1V_2}$ː-$\mathrm{C_1V_2C_3}$ (*partial*). The decomposed forms are not only absent in the lexicon, but also phonotactically illegal, as Tagalog roots must be minimally disyllabic. Evidence from phonological alternation suggests that these pseudo-reduplicated words also have copying structures similar to reduplicated words that are morphologically complex (Zuraw, 2002). A recent MEG study on visual inputs (Wray et al., 2022) further supports this claim: Tagalog speakers show early decomposition of these morphologically simple pseudo-reduplicated words, similar to how they process morphologically complex reduplicative structures.

1.1 Why reduplication: A summary of issues

Reduplication and surface repetitions have been attracting attention from many fields of scientific investigations. Robust evidence suggests that reduplication and surface repetitions are indispensable components of human linguistic and cognitive systems. Within natural languages, reduplication is commonly attested with many subtypes (Moravcsik, 1978; Rubino, 2005; Hurch and Mattes, 2009; Dolatian and Heinz, 2019). Previous studies on these attested reduplicative patterns have identified crucial typological generalizations. For example, total reduplication seems to be more frequent than partial reduplication. Moravcsik (1978, p. 328) hypothesized that all languages with attested partial reduplication also use total reduplication, but not vice versa. This is later supported by a reported survey in the World Atlas of Language Structure Database (WALS; Rubino, 2013),[1] 278 languages show both total reduplication and partial reduplication, and 35 languages with only total reduplication. Rubino reported no language that only has partial reduplication.[2] Partial reduplication is said to exhibit more interesting interactions with other phonological phenomena, such as vowel reduction, consonant deletion, cluster simplification, and so on (Steriade, 1988; Inkelas and Downing, 2015; Zimmermann, 2021b). Rubino also made the remark that partially reduplicated material is most commonly found as prefixes, though it is possible to surface as suffixes and infixes. Typological generalizations of this sort have laid the foundations for numerous theoretical proposals of the phonological grammar and morphology-phonology interface (e.g., Wilbur, 1973; Aronoff, 1976; Moravcsik, 1978; Carrier, 1979; Marantz, 1982;

[1] https://wals.info/chapter/27

[2] Palauan (Austronesian) seems to be an exception to this generalization. See Zuraw (2003).

Levin, 1985; Steriade, 1988; Hayes and Abad, 1989; McCarthy and Prince, 1986, 1995; Hendricks, 1999; Raimy, 2000; Zuraw, 2002; Inkelas and Zoll, 2005; Yu, 2005; Frampton, 2009; Saba Kirchner, 2010, 2013; Kiparsky, 2010; McCarthy et al., 2012; Zukoff, 2017; Stanton and Zukoff, 2018; Wei and Walker, 2020; Zimmermann, 2021b; Lamont, 2023; Yang, 2023), which in turn motivated the experimental studies in this dissertation (see Chapter 2).

Besides its typological prevalence and theoretical significance, reduplication is found to have an important function in language acquisition: reduplicated words have been shown to aid speech segmentation (Ota and Skarabela, 2018) and facilitate lexical learning (Ota and Skarabela, 2016). Outside of natural languages, humans are found to be highly sensitive to surface repetitions and identity relations in artificial languages – this holds for newborns (Gervain et al., 2012), infants (Marcus et al., 1999; Gerken, 2006; Marcus et al., 2007; Kovács and Mehler, 2009a,b; Gerken, 2010),[3] and adults (Gallagher, 2013; Berent et al., 2016, 2017; Moreton et al., 2021; Gow Jr et al., 2023), in both speech and signs. Evidence supporting the prominence of surface repetitions has also been found in other cognitive domains (e.g., Endress et al., 2007; Saffran et al., 2007; Frank et al., 2009; Dawson and Gerken, 2009; Finley and Christiansen, 2011; Thiessen, 2012; Ferguson and Lew-Williams, 2016).

There are few experimental studies that have investigated reduplication for specific questions rooted in linguistic theories. In four experiments of Berent et al. (2016), English speakers were prompted to choose between a reduplicated string (e.g. [slaflaf]) and an unreduplicated string (e.g. [slafmak]) orthographically presented on the screen. They found that English speakers disliked reduplicated strings if they were presented as surface word forms but showed a preference if reduplicated strings were presented together with a base (e.g. [slaf]) and paired with a morphological operation to mark pluralization. Berent et al. (2017) further confirmed a *contiguity preference*: in a morphological setting, words showing contiguous copying as in [traf-raf] were rated as a better word than words with segmental skipping as in [traf-taf]. However, [traf-raf] was rated equally bad as [traf-taf] when presented as the surface forms only. Based on the significant differences when stimuli were

[3] See a meta-analysis in Rabagliati et al. (2019) for a more complete review on sensitivity to repetitions in infancy.

presented in different contexts, Berent et al. (2016) and Berent et al. (2017) argued for a phonology-morphology split.

Given these experimental results, we could also conclude that participants in both experiments might have constructed reduplicated structures for the surface phonological forms, as they have exhibited a clear *dispreference* or *preference*, not *insensitivity*. In two-alternative forced choice tasks, to form an 'aversion', the difference between the reduplicated string ([slaflaf]) and the non-reduplicated string ([slafmak]), must be systematically detected. Otherwise, participants would show chance-level random guessing. One key aspect of the copying operation in phonology is the ability to recognize copies, but not necessarily always prefer copied structures. Once the special status of a reduplicated string is acknowledged, it suggests that purely looking at the phonological forms, participants were able to parse novel reduplicated strings.

Aside from Berent et al., to our knowledge, there is only one other study directly examining variable shapes, with the goal of studying the relationship between reduplication learning and its typology. Haugen et al. (2022) prompted native speakers of English to learn one of the three different partial reduplication patterns for the augmentative, together with a total reduplication pattern for pluralization as a control. The three partial reduplication patterns were (1). the base-independent fixed light syllable copying (e.g., [va.vam.se.ta] and [ne.ne.ko.la]), (2). the base-independent fixed heavy syllable copying (e.g., [vam.vam.se.ta] and [nek.ne.ko.la]) and (3). the base-dependent syllable copying (e.g., [vam.vam.se.ta] and [ne.ne.ko.la]). The previous two patterns are well-attested across world languages yet the third pattern is quite rare (Moravcsik, 1978; Marantz, 1982; McCarthy and Prince, 1995). Their experimental design followed an *ease of learning* paradigm. There was only one testing phase but not a training phase. On each trial, participants were prompted with three choices, including total reduplication (e.g., [vam.se.ta.vam.se.ta]), CV reduplication (e.g., [va.vam.se.ta]), and CVC reduplication (e.g., [vam.vam.se.ta]), with orthographic input. They received immediate feedback on each trial. Participants were assessed by the final performance and their learning trajectories. Compared to the learning of base-independent patterns, participants in the base-dependent syllable copying condition showed some amount of learning in the end, but their learning trajectories were much slower with more errors. The authors concluded that the typological rarity of the base-dependent reduplicative patterns might be related to their learning difficulty.

Surprisingly, despite its rich typology and cognitive saliency, reduplication imposes a long-standing challenge for computational modeling work. Scholars have argued that to represent the identity-based relations within repetitions, it is necessary to use abstract rules in models of acquisition, as well as variables in models of cognition (e.g., Marcus, 2003). The argument on the status of a variable was initiated after the experimental work by Marcus et al. (1999). In this experiment, infants detected and generalized the familiarized repetition rule to novel syllables, but the authors reported that variable-free connectionist models (back then, simple recurrent network; Elman, 1990) failed to generalize as infants did. Since then, computational modeling literature for identity-based patterns have kept growing (e.g., Frank and Tenenbaum, 2011; Berent et al., 2012; Ellis et al., 2022). For a more comprehensive overview, see Alhama and Zuidema (2019). It remains an active area of research, and some of the most recent efforts include testing the architectural advancements of (deep) neural networks against the reduplicative patterns to assess their capabilities (e.g., Wilson, 2019; Nelson et al., 2020; Haley and Wilson, 2021; Beguš, 2021; Beguš and Zhou, 2022; Prickett et al., 2022).

Within this extensive literature, two fundamental questions remain unresolved. First, what are all possible natural language word forms with reduplication? Second, how can they be characterized and learned as a part of morphophonology? We think three types of missing evidence might help to improve our understanding of these questions.

First, there needs to be more investigation into the empirical landscape of reduplication. In theoretical work, the (un)attestedness of certain types of reduplication often motivates arguments for some theories over others. However, it is unclear how speakers encode these patterns, for which experimental methods will be valuable. The majority of previous experimental studies have focused on the most canonical cases of reduplication, leaving (slightly) more complicated patterns with intricate theoretical consequences less frequently explored. Therefore, experiments focusing on more diverse reduplicative patterns are necessary at the current stage to test the psychological plausibility of various descriptive patterns and to provide evidence, either converging or conflicting, to relate to typological generalizations.

Second, despite its empirical ordinariness and learning ease, reduplication has been treated as a long-standing challenge made in formal language theory (Sproat, 1992; Dolatian and Heinz, 2020; Wang and Hunter, 2023), even when we limit ourselves to the most canonical cases. It demands more computational power than other phonological and morphophonological structures in the classical Chomsky Hierarchy, resulting in a *misalignment* between current formal language classes and the natural language typology. This requires reconsideration of formal mathematical analyses of the copying operation in the context of morphophonology.

Third, a computational learning model that concretizes the procedures of reduplication learning will provide insights into its learning, particularly on how reduplicative patterns can be learned together with other (morpho)phonological processes.

1.2 The goal and structure of this dissertation

A major goal of this dissertation is to unify the phenomenon of reduplication and its learning with other (morpho)phonological structures. We address these issues by integrating phonological theory with artificial grammar learning experiments, formal mathematical analyses, and computational modeling. Our studies build on phonological theories of reduplication. Thus, before presenting our own work, we provide an overview of the previous theories of reduplication in Section 1.3, with a focus on Base-Reduplicant Correspondence Theory (McCarthy and Prince, 1995) to set the stage.

The remaining chapters are structured as follows. Chapter 2 presents empirical evidence on how human learners tend to generalize different reduplicative patterns. We conducted few-shot artificial grammar learning experiments following *the poverty of the stimulus paradigm* (Wilson, 2006). We provided participants with only a few familiarized forms, consisting of pairs of stems and their reduplicated forms, and then asked them to make generalizations for novel forms. The input provided to the participants was of limited variety, allowing them to be compatible with multiple hypotheses at different granularities of phonological abstraction. The first series of experiments, consisting of three experiments, focused on syllable-internal structures. For example, participants were familiarized with [ˈdɔv.ɡə] and [dɔv-ˈdɔv.ɡə], and tested on varying shapes of the target syllable, such as forms with complex onsets [ˈstæb.ɡə] and onsetless ones [ˈɑv.di]. Moreover, we explored how segmental identity affects the encoding of these reduplicative patterns (e.g, familiarize with the reduplicated forms as [dəv-ˈdɔv.ɡə] and [div-ˈdɔv.ɡə] respectively). The second series of experiments extended beyond the level of syllables, familiarizing participants with monosyllabic copying (e.g., [ˈpif] and [ˈpif-pif]) and testing them with longer forms (e.g., [ˌpi.sæ.ˈɡoʊ.bɛ.kʊt]).

Results from the learning experiments in Chapter 2 motivate two directions of research, both aiming to fit reduplication into the rest of (morpho)phonology. First, in Chapter 3, we propose a formal characterization of a more typologically motivated formal language class to serve as the computational-level model (Marr, 1982) of possible natural language word sets. This is achieved by augmenting the regular class with a primitive copying operation. We then prove a pumping lemma and analyze the closure properties of the resulting language class, finding that it preserves some desirable properties of the regular class. Our proposal lays the foundation for future investigations into different variants of the formal model, with the ultimate goal of including all attested reduplicative structures and excluding all unattested ones. We discuss the details of the computational device, the implications it has for phonological theory, and the potential modifications for greater typological coverage.

As for the second line of the investigation, Chapter 4 proposes a computational learner for reduplication learning. The key challenge is to learn the unobserved prosodic template when the underlying representations of the stems and other affixes are also unknown. The proposed learner has two main ingredients. First, it treats the candidate prosodic templates as *hidden structures* (Tesar and Smolensky, 1998), the same status as phonological underlying representations. Second, it attributes the realization of copying to phonology, largely following Base-Reduplicant Correspondence Theory (McCarthy and Prince, 1995) in Maximum Entropy Grammars (Goldwater and Johnson, 2003), a probabilistic framework. We present several simulation results with toy datasets to evaluate the learner's ability to handle the reduplication-phonology interactions, alongside some preliminary investigations of the experimental results in Chapter 2. Our findings show that, under this view, the problem of reduplication learning can be integrated into the overall picture of learning morphophonological processes in intuitive ways.

Chapter 5 concludes the dissertation by summarizing the current studies and discussing directions for future research. Using reduplication as a case study, this dissertation as a whole aims to demonstrate how we can study the grammar architecture and the learner, as well as their interactions, in a more precise manner.

1.3 Theories of reduplication: An overview

In this section, we aim to provide sufficient information on the theories of reduplication to lay the groundwork for our own studies.

Classical theories of reduplication, including Wilbur (1973) and Carrier (1979), have highlighted the role of phonology in deriving the surface reduplicative forms. Subsequent works by Marantz (1982), McCarthy and Prince (1986), and Steriade (1988) built upon the phonological copying account. They argued that reduplication is similar to normal affixation processes, with phonology executing the copying operation triggered by an abstract reduplicative morpheme. Base-Reduplicant Correspondence Theory (hereafter BRCT; McCarthy and Prince, 1995), couched in Optimality Theory (Smolensky and Prince, 1993), stands as one of the most influential theories advocating such a phonological copying approach. Later theories often use BRCT as a baseline, directly comparing and/or modifying its architecture. The later chapters of this dissertation also rely heavily on BRCT. Therefore, we will focus on the workings of BRCT and briefly sketch how later theories develop upon it. For comprehensive literature reviews of other proposals, see Saba Kirchner (2010), Raimy (2011), Inkelas (2014, §5), Inkelas and Zoll (2005), and Downing and Inkelas (2015).

For clarity, we will work with a concrete example of Ilokano partial reduplication presented in (1b), repeated below in (3) but with stress ignored.

(3) Ilokano (Austronesian, Philippines; Hayes and Abad, 1989, p. 357)

[kaldiŋ]	‘goat’	[kal-kaldiŋ]	‘goats’
[pusa]	‘cat’	[pus-pusa]	‘cats’

1.3.1 Base-Reduplicant Correspondence Theory (McCarthy and Prince, 1995)

1.3.1.1 The building blocks

BRCT handles reduplication with three building blocks. First, the underlying representation of the reduplicative morpheme is phonologically empty, denoted as Red. Except for this point, reduplication resembles normal affixation in terms of morphological computation. For example, in the context of Ilokano pluralization, assuming the underlying representation /pusa/ for the stem ‘cat,’ the input to phonology for the form ‘cats’ is the concatenation of the abstract morpheme and the UR of the stem, that is /Red+pusa/.

The surface realization of this abstract morpheme, termed the reduplicant (underlined in the examples below), is regulated by two forces in phonology. In an Optimality-Theoretic view, these are two families of violable constraints that determine the phonological shape and the segmental content of the reduplicant, respectively. The family of templatic constraints, which refer to prosodic units, regulates the phonological shape. For example, Red = $\sigma_{\mu\mu}$[4] penalizes all candidates in which the reduplicant is not a heavy syllable, such as the candidates with a light syllable [pu-pusa]. On the other hand, Red = σ_{μ} penalizes all candidates in which the reduplicant is not a light syllable, such as the observed output [pus-pusa]. To make Ilokano pluralization always surface as a heavy syllable (Hayes and Abad, 1989), the templatic constraint Red = $\sigma_{\mu\mu}$ must be highly ranked, as in (4).

(4) Ilokano: heavy syllable reduplication

/Red+pusa/	Red = $\sigma_{\mu\mu}$	Red = σ_{μ}
☞ a. pus-pusa		$*$
b. pu-pusa	$*!$	

[4] McCarthy and Prince (1995) defines this family of constraints as a kind of Align constraints: Red = $\sigma_{\mu\mu}$ could be re-expressed as Align(Red, L/R; $\sigma_{\mu\mu}$, L/R). For simplicity, we use the Red = X format.

The mechanism that derives segmental identity-based effects requires the following apparatus. First, segments in the reduplicant and the base are hypothesized to be in correspondence with each other, making them "codependent" in phonological computation. The traditional notation uses subscripts to indicate such segmental dependencies, marking two corresponding segments with the same subscript index. For example, in the form $[\underline{p_1u_2s_3}\text{-}p_1u_2s_3a_4]$, each segment in the reduplicant is identical to the corresponding counterpart in the base. The basic model of reduplication consists of both the surface base-reduplicant segmental correspondence and the normal input-output (or input-base) correspondence.[5] Figure 1.2 illustrates the basic model of correspondence relations in reduplication, using the form $[\underline{p_1u_2s_3}\text{-}p_1u_2s_3a_4]$.[6]

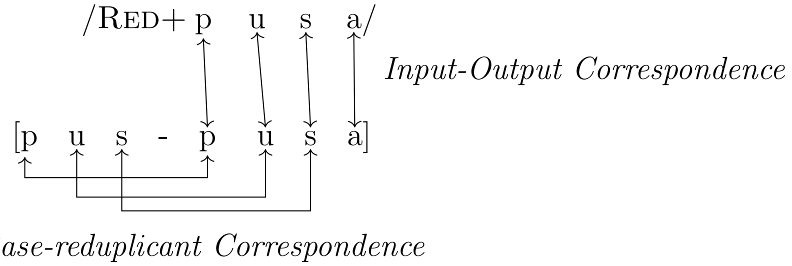


Figure 1.2: *An illustration of correspondence theory with [pus-pusa]*

Consider another possible candidate $[\underline{b_1u_2s_3}\text{-}p_1u_2s_3a_4]$. The onset in the reduplicant is not identical to the corresponding onset in the base. To select the winner over this candidate, we need to rely on faithfulness constraints, which penalize discrepancies between corresponding segments. The commonly adopted faithfulness constraints and their definitions can be found in Appendix A. In the current context, the crucial constraint that differentiates these two candidates is Ident-BR(Voice), which bans any voice mismatches between the reduplicant and the base. The observed output $[\underline{p_1u_2s_3}\text{-}p_1u_2s_3a_4]$ does not violate this constraint. On the other hand, the imperfect copying candidate $[\underline{b_1u_2s_3}\text{-}p_1u_2s_3a_4]$ does incur one violation.

[5] Besides IO and BR correspondence, the full model of reduplication also includes the segmental correspondence between the input and the reduplicant in the output (IR).

[6] We have only depicted one-to-one correspondence, omitting possible many-to-one correspondences, such as those found in coalescence ($/x_1y_2/ \rightarrow [z_{12}]$) and diphthongization ($/z_{12}/ \rightarrow [x_1y_2]$). The Integrity and Uniformity constraint families are responsible for these many-to-one correspondence relations. Likewise, we did not include discussions of metathesis, which is governed by Linearity constraints. See the original discussion in McCarthy and Prince (1995, pp. 123-124).

This Ident-BR(Voice) must outrank any markedness constraints motivating the voicing mismatch (e.g., *[p). The required constraint ranking is as in (5).

(5) Ilokano: segmental identity

/Red+pusa/	Ident-BR(Voice)	*[p
☞ a. $\underline{p_1u_2s_3}$-$p_1u_2s_3a_4$		*
b. $\underline{b_1u_2s_3}$-$p_1u_2s_3a_4$	*!	

1.3.1.2 The typology

In OT, the set of violable constraints is said to be universal. For each language, the optimal candidates are selected based on constraint ranking. Different languages are the result of different constraint rankings. To see this, let us consider the five constraints in (6), some of which were already described above. For simplicity, let us assume IO-faithfulness constraints are undominated. In other words, the base always surfaces faithfully as the UR of the stem.

(6) a. Templatic constraints

1. Red = $\sigma_{\mu\mu}$: penalize any reduplicant that is not a heavy syllable
2. Red = σ_{μ}: penalize any reduplicant that is not a light syllable

b. BR-faithfulness constraints

3. Ident-BR(Voice): assign one violation for each pair of BR-corresponded segment that differ on [Voice] feature
4. Max-BR: assign one violation for each segment in the base without a correspondent in the reduplicant

c. Markedness constraint

5. *[p: assigns one violation to any word-initial voiceless labial stop.

This constraint set will yield six possible languages, varying along two linguistic dimensions: the shape of the reduplicant, demonstrated in (7) and the degree of phonological identity on the surface, demonstrated in (8). For the shape of the reduplicant, there are three possibilities: (a). surfacing as a heavy syllable, which occurs when Red $= \sigma_{\mu\mu}$ dominates Red $= \sigma_{\mu}$ and Max-BR; (b). surfacing as a light syllable, which occurs when Red $= \sigma_{\mu}$ dominates Red $= \sigma_{\mu\mu}$ and Max-BR; (c). total reduplication, which occurs when Max-BR dominates the other two templatic constraints.

(7) a. Ilokano: heavy-syllable reduplication with perfect identity

/Red+pusa/	Ident-BR(Voice)	Red $= \sigma_{\mu\mu}$	Max-BR	Red $= \sigma_{\mu}$	*[p
☞ a. $\underline{p_1u_2s_3}$-$p_1u_2s_3a_4$			*	*	*
b. $\underline{p_1u_2}$-$p_1u_2s_3a_4$		*!	**		*
c. $\underline{p_1u_2s_3a_4}$-$p_1u_2s_3a_4$		*!		*	*
d. $\underline{b_1u_2s_3}$-$p_1u_2s_3a_4$	*!		*	*	
e. $\underline{b_1u_2}$-$p_1u_2s_3a_4$	*!	*	**		
f. $\underline{b_1u_2s_3a_4}$-$p_1u_2s_3a_4$	*!	*		*	

b. Ilokano': light-syllable reduplication with perfect identity

/Red+pusa/	Ident-BR(Voice)	Red $= \sigma_{\mu}$	Max-BR	Red $= \sigma_{\mu\mu}$	*[p
a. $\underline{p_1u_2s_3}$-$p_1u_2s_3a_4$		*!	*		*
☞ b. $\underline{p_1u_2}$-$p_1u_2s_3a_4$			**	*	*
c. $\underline{p_1u_2s_3a_4}$-$p_1u_2s_3a_4$		*!		*	*
d. $\underline{b_1u_2s_3}$-$p_1u_2s_3a_4$	*!	*	*		
e. $\underline{b_1u_2}$-$p_1u_2s_3a_4$	*!		**	*	
f. $\underline{b_1u_2s_3a_4}$-$p_1u_2s_3a_4$	*!	*		*	

c. Ilokano'': total reduplication with perfect identity

/Red+pusa/	Ident-BR(Voice)	Max-BR	Red = σ_{μ}	Red = $\sigma_{\mu\mu}$	*[p
a. $\underline{p_1u_2s_3}$-$p_1u_2s_3a_4$		*!	*		*
b. $\underline{p_1u_2}$-$p_1u_2s_3a_4$		*!*		*	*
☞ c. $\underline{p_1u_2s_3a_4}$-$p_1u_2s_3a_4$			*	*	*
d. $\underline{b_1u_2s_3}$-$p_1u_2s_3a_4$	*!	*	*		
e. $\underline{b_1u_2}$-$p_1u_2s_3a_4$	*!	**		*	
f. $\underline{b_1u_2s_3a_4}$-$p_1u_2s_3a_4$	*!		*	*	

As for the degree of phonological identity, two relevant constraints are the markedness constraint [*p and the other BR-faithfulness constraint Ident-BR(Voice). When Ident-BR(Voice) dominates *[p, we expect perfect identity between reduplicant and the base, as in all three languages above. However, when the urge to ban word-initial [p] is strong enough to dominate Ident-BR(Voice), instead of surfacing faithfully, the winner is expected to show markedness repair and favor the word-initial [b], as illustrated below.

(8) a. Ilokano''': heavy-syllable reduplication with initial voicing

/Red+pusa/	*[p	Red = $\sigma_{\mu\mu}$	Max-BR	Red = σ_{μ}	Ident-BR(Voice)
a. $\underline{p_1u_2s_3}$-$p_1u_2s_3a_4$	*!		*	*	
b. $\underline{p_1u_2}$-$p_1u_2s_3a_4$	*!	*	**		
c. $\underline{p_1u_2s_3a_4}$-$p_1u_2s_3a_4$	*!	*		*	
☞ d. $\underline{b_1u_2s_3}$-$p_1u_2s_3a_4$			*	*	*
e. $\underline{b_1u_2}$-$p_1u_2s_3a_4$		*!	**		*
f. $\underline{b_1u_2s_3a_4}$-$p_1u_2s_3a_4$		*!		*	*

b. Ilokano'''': light-syllable reduplication with initial voicing

/Red+pusa/	*[p	Red = σ_μ	Max-BR	Red = $\sigma_{\mu\mu}$	Ident-BR(Voice)
a. $\underline{p_1u_2s_3}$-$p_1u_2s_3a_4$	*!	*	*		
b. $\underline{p_1u_2}$-$p_1u_2s_3a_4$	*!		**	*	
c. $\underline{p_1u_2s_3a_4}$-$p_1u_2s_3a_4$	*!	*		*	
d. $\underline{b_1u_2s_3}$-$p_1u_2s_3a_4$		*!	*		*
☞ e. $\underline{b_1u_2}$-$p_1u_2s_3a_4$			**	*	*
f. $\underline{b_1u_2s_3a_4}$-$p_1u_2s_3a_4$		*		*	*

c. Ilokano''''': total reduplication with initial voicing

/Red+pusa/	*[p	Max-BR	Red = σ_μ	Red = $\sigma_{\mu\mu}$	Ident-BR(Voice)
a. $\underline{p_1u_2s_3}$-$p_1u_2s_3a_4$	*!	*	*		
b. $\underline{p_1u_2}$-$p_1u_2s_3a_4$	*!	**		*	
c. $\underline{p_1u_2s_3a_4}$-$p_1u_2s_3a_4$	*!		*	*	
d. $\underline{b_1u_2s_3}$-$p_1u_2s_3a_4$		*!	*		*
e. $\underline{b_1u_2}$-$p_1u_2s_3a_4$		*!*	**	*	*
☞ f. $\underline{b_1u_2s_3a_4}$-$p_1u_2s_3a_4$			*	*	*

1.3.2 The development of later theories

Later theories have raised doubts about the nature of these varying dimensions discussed above and challenged the necessity of corresponding components of BRCT to varying degrees. We will briefly review these threads in the following sections.

1.3.2.1 Templatic effects in reduplication

One commonly debated thread is the nature of templatic effects in reduplication. Proposals reconsidering this aspect can be broadly categorized as three positions along a spectrum.

The first group is pro-templates, including Minimal Reduplication (Saba Kirchner, 2010, 2013), Serial Template Satisfaction (McCarthy et al., 2012) and Reduplication as Distribution of Underlying Activity (Zimmermann, 2021b). These theories suggest that grammar should still explicitly specify prosodic templates for reduplicative morphemes, not through templatic constraints but via the UR. In other words, instead of being phonologically empty, the UR for the reduplicative morpheme is a segmentless prosodic unit. The copying operation is triggered purely by the drive to satisfy the templatic UR. For Ilokano reduplication, the UR for 'cats' is simply /$\sigma_{\mu\mu}$+pusa/.

For the sake of parsimony, some theories question whether *reduplication-specific* prosodic templates, including both templatic constraints and URs, are needed. McCarthy and Prince (1994), and subsequently Urbancyzk (1996, 2001) propose that reduplication should not be singled out as its own type but be considered as a particular kind of morpheme (e.g., root, affix, stem). In this theory, often referred to as Generalized Template Theory, different types of morphemes have specific prosodic requirements, implemented as different types of templatic constraints, such as STEM = PRWD and AFFIX = σ. For Ilokano reduplication, there is no RED = $\sigma_{\mu\mu}$, but rather the more general AFFIX = σ.

Going further, a-templatic approaches (Spaelti, 1997; Gafos, 1998; Hendricks, 1999; Riggle, 2006) abandon the special mechanism to derive the templatic effects altogether. Instead, they suggest that these effects are byproducts of the interaction among independently motivated constraints. For example, partial reduplication might result from the *STRUCT family (Zoll, 1993, 1994), which penalizes any phonological material in the output form. The constraint ranking *happens to* derive the appropriate size of the reduplicant.

1.3.2.2 Surface BR correspondence relations

The strongest empirical motivation for surface BR correspondence relations comes from different types of reduplication-phonology interactions, including overapplication and underapplication, identified by Wilbur (1973) and Carrier (1979).

Overapplication describes the scenario in which a phonological process applies in an unmotivated phonological context due to the need to satisfy BR faithfulness. An example comes from Malay (Austronesian, Malaysia). Malay phonotactics forbids the sequence of a nasal followed by a voiceless stop, denoted as *NC̥. This restriction leads to consonant coalescence, as illustrated in (9a). However, in the case of reduplication, consonant coalescence applies to unmotivated environments. Consider the possible form *[mə**m**ukul-**p**ukul]. The voiceless stop [p] in the second copy does not violate this phonotactic restriction to motivate coalesence. In actual fact, coalescence still applies, making [mə**m**ukul-**m**ukul] the output.

(9) Overapplication in Malay (Austronesian; Yang and Wong, 2020, p. 120)

a. Malay nasal substitution

UR	*SR*	*gloss*	*UR*	*SR*	*gloss*
/mə**N**+**p**ukul/	[mə**m**ukul]	‘hit’	/mə**N**+bunʊh/	[mə**m**bunʊh]	‘kill’
/mə**N**+**t**ari/	[mə**n**ari]	‘dance’	/mə**N**+duga/	[mə**n**duga]	‘suspect’
/mə**N**+**k**əʤar/	[mə**ŋ**əʤar]	‘chase’	/mə**N**+ganti/	[mə**ŋ**ganti]	‘change’

b. Overapplication with /məN/-prefix

Stem UR	*Reduplicated*	*gloss*	*Stem UR*	*Reduplicated*	*gloss*
/pukul/	[mə**m**ukul-**m**ukul]	‘hit’	/bunʊh/	[mə**m**bunʊh-bunʊh]	‘kill’
/tari/	[mə**n**ari-**n**ari]	‘dance’	/duga/	[mə**n**duga-duga]	‘suspect’
/kəʤar/	[mə**ŋ**əʤar-**ŋ**əʤar]	‘chase’	/ganti/	[mə**ŋ**ganti-ganti]	‘change’

The constraint ranking scheme that gives rise to overapplication is as in (10a), which involves the participation of IO-faithfulness constraints. A tableau with the Malay example is given in (10b).

(10) a. Constraint ranking scheme for overapplication

BR-faithfulness, Markedness constraint >> IO-faithfulness constraint

b. Malay overapplication

/məN+Red+pukul/	BR-Faith	*NC̥	IO-Faith
☞ a. mə**m**ukul-**m**ukul			*
b. mə**m**ukul-**p**ukul	*!		
c. mə**mp**ukul-**p**ukul		*!	
d. mə**mp**ukul-**m**ukul	*!		*

BRCT predicts a special kind of overapplication process, known as *templatic backcopying* (or the Kager-Hamilton Problem; McCarthy and Prince, 2004, pp. 258-267). This refers to the cases when the phonological shape of the reduplicant affects the shape realization of the base, resulting in a reduction in size for both copies. This pattern is derived using the same constraint ranking schema as above, specifically with the faithfulness constraint Max-BR and the templatic constraint dominating the IO-faithfulness constraint Max-IO. Templatic backcopying is often regarded as a "conundrum" for BRCT because it does not appear to be robustly attested in natural languages. To our knowledge, there are only a few real language examples: Hausa (Chadic; Downing, 2000), Tonkawa (Coahuiltecan; Gouskova, 2007), and Guarijio (Uto-Aztecan; Caballero, 2006). (11) presents the case of Guarijio abbreviated reduplication, in which both copies are truncated to light syllables.

(11) Guarijio (Uto-Aztecan; Caballero, 2006, p. 278)

[toní]	'to boil'	[to-tó]	'to start boiling'
[kusú]	'to sing (animals)'	[ku-kú]	'to start singing'
[muhíba]	'to throw'	[mu-mú]	'to start throwing'

Underapplication works the opposite way, as illustrated by an example from Akan reduplication (Niger–Congo, Ghana; McCarthy and Prince, 1995, pp. 93-97) in (12). In Akan, velars ({k, ɡ, w, ŋ^{w}}) and [h] typically do not precede non-low front vowels ({i, ɪ, ɛ, e}). However, this restriction is violated in the context of reduplication. In Akan CV reduplication, the reduplicant is prespecified with a high vowel, following either the pattern of *Ci*- or *Cɪ*-. When the base starts with a velar or a [h], the reduplicant would contain the illegal sequence, such as [kɪ] and [hɪ]. The velar [k] and [h], thus, are expected to palatalize to [tɕ] and [ç] respectively. However, the reduplicated form surfaces with the otherwise forbidden sequence, due to the urge to maintain base-reduplicant faithfulness. Yoruba (Niger-Congo, Nigeria) gerundive reduplication shows a similar pattern: laterals nasalize to [n] before high front vowels, but laterals do appear before high front vowels in reduplicants; see Pulleyblank (2008).

(12) Underapplication in Akan (Niger–Congo, Ghana; Pulleyblank, 2008, p. 351)

a. Palatalization in Akan

Attested	*gloss*	*Unattested*
[tɕɛ]	‘divide’	*[kɛ]
[çɪ]	‘border’	*[hɪ]

b. Underapplication of palatalization in the reduplicant

Stem UR	*Reduplicated*		*gloss*
/kaʔ/	[kɪ-kaʔ]	*[tɕɪ-kaʔ]	‘bite’
/hawʔ/	[hɪ-hawʔ]	*[çɪ-hawʔ]	‘trouble’

At a conceptual level, the difference between underapplication and overapplication seems intuitive. However, BRCT makes the interesting prediction that underapplication should be not found without other independently motivated phonological factors (McCarthy and Prince, 1995, p. 5). Underapplication cannot be derived from the interactions of the constraints that are sufficient for overapplication. Note that markedness must dominate IO-faithfulness for markedness repair in non-reduplicated forms, as in (13). Due to the symmetric nature of the BR faithfulness constraints, the BR-faithfulness constraint cannot distinguish the overapplication candidate *[tɕɪ-tɕaʔ] and the underapplication candidate [kɪ-kaʔ]. Without other effective constraints, as illustrated in (14), the grammar would entertain the overapplication candidate.

(13) Akan non-reduplicated forms

/hɪ/	Markedness	IO-Faith
a. hɪ	∗!	
☞ b. çɪ		∗!

(14) Akan reduplicated forms without other phonological factors

/Red+kaʔ/	BR-Faith	Markedness	IO-Faith
☞ a. kɪ-kaʔ		∗!	
b. tɕɪ-kaʔ	∗!		
c. kɪ-tɕaʔ	∗!	∗	
☜ d. tɕɪ-tɕaʔ			∗

What makes the underapplication candidate win in Akan is another independently motivated phonological restriction, specifically a ban on cooccurrence of coronals in successive syllables, denoted as OCP(+Cor). As illsutrated in (15b), the overapplication candidate is penalized and loses to the underapplication candidate. In this way, BRCT predicts that when a language has a categorical underapplication process, there must be another independently motivated phonological factor (denoted as *C* below) that penalizes the overapplication candidate, allowing the underapplication candidate to win. The corresponding constraint ranking scheme is shown in (15a).

(15) a. Constraint ranking scheme for underapplication
C, BR-faithfulness>> Markedness constraint >> IO-faithfulness constraint

b. Akan underapplication with OCP

/Red+kaʔ/	OCP(+Cor)	BR-Faith	Markedness	IO-Faith
☞ a. kɪ-kaʔ			∗!	
b. tɕɪ-kaʔ		∗!		
c. kɪ-tɕaʔ		∗!	∗	
d. tɕɪ-tɕaʔ	∗!			∗

Many have questioned the empirical status of these reduplication-phonology interactions since their discovery (e.g., Kiparsky, 2010; Saba Kirchner, 2010; McCarthy et al., 2012). For example, Saba Kirchner (2010, p. 115) made the remark that templatic backcopying is limited to a small number of verbs in Guarijio, raising doubts about the productivity of this process. If these patterns do not hold, BRCT would be considered too expressive because it overgenerates, particularly in the case of templatic backcopying. While the empirical status of these patterns remains murky in natural language typology, experimental work may provide evidence for a clearer picture. With this in mind, we conclude the overview of the phonological theories and proceed to the learning experiments in the next chapter.

CHAPTER 2

The emergent typology of reduplication: Artificial grammar learning experiments

2.1 Introduction

An extensive literature[1] (e.g, Wilson, 2006; Pater and Moreton, 2012; Becker et al., 2011; White, 2014; Yin and White, 2018; Martin and White, 2021; Wilson, 2022; Moreton and Pertsova, 2024) has investigated analytic biases which guide phonological and morphophonological learning. These learning biases, reflecting how easily a pattern is learned compared to others, are hypothesized to shape the attested linguistic structures across world languages (i.e. typology; e.g., Moreton, 2008; Stanton, 2016), or at least predict the trend of typological universals (Culbertson et al., 2012). Reduplication and surface repetitions have long been the focus of theories of phonology, morphophonology, and language learning. However, few studies have examined the inductive biases that guide learners in reduplication learning. This chapter addresses this gap with artificial grammar learning experiments.

In this chapter, we examine the biases seen in humans' rapid generalization of various reduplicative patterns from highly impoverished input. We followed *the poverty of the stimulus paradigm* (Wilson, 2006). Participants were provided with only a few training forms, consisting of pairs of stems and their reduplicated forms (e.g., [ˈdɔvgə] and [dɔvˈdɔvgə]; [ˈʃæp.mə] and [ʃæpˈʃæp.mə]), with the semantics of reduplication designated as pluralization.

[1] Experiments 1a, 1b, and 1c were collaborated with Colin Wilson and funded by NSF BCS-1941593 to CW. Transcribed responses were made available by Colin Wilson. Experiments 1a and 1c were presented at the Annual Meeting on Phonology 2022. The content of this series of experiments is in preparation for publication, co-authored with Colin Wilson. The second series of experiments were supported by the UCLA Linguistics Department Research Fund. A version of Experiment 2a was presented at the 2024 LSA Annual Meeting.

The familiarized patterns corresponded to some naturally occurring language examples, but the input provided to the participants was of limited variety, allowing them to be compatible with multiple hypotheses at different levels of phonological abstraction.

Our experimental results yield three main findings. First, across two experimental series, we found that participants rapidly generalized and extended copying-based hypotheses to novel forms on the basis of just a few familiarized items. This suggests that reduplication is easy to learn as a morphophonological process. Second, participants' generalizations aligned with the coarse-grained phonological abstractions characterizable by the vocabulary of prosody (e.g., light and heavy syllable, foot, prosodic word), but not by finer-grained phonological specifications like syllabicity or segment count. These findings strongly support high-level phonological abstraction, consistent with the core claims of McCarthy and Prince (1986, *et seq.*). Lastly, we found strong correlations between participants' responses and naturally occurring reduplicative patterns. Beyond the general preferences for the coarse-grained prosodic units, we also discovered that the variations in individually biased grammars reflected the variations attested in natural languages. These results provide a learning-oriented perspective on the attested universals and variations in the context of reduplication and its typology.

We begin this chapter by motivating the research questions and design of our experiments. In Section 2.2, we provide an overview on missing pieces of empirical studies on reduplication. Section 2.3 describes the various dimensions of the reduplicative typology investigated in this chapter and provides corresponding language examples used as the empirical bases for our experiments. We then present the design, the methods, and the results of the three experiments of Experiment Series 1 (Section 2.4) and three experiments of Experiment Series 2 (Section 2.5). This chapter concludes by discussing the implications of these results for phonological theories, biases in morphophonological learning, and human-like computational modeling in Section 2.6.

2.2 Background

2.2.1 The missing piece

For decades, reduplication has been extensively studied by linguists, and the underpinning identity-based dependencies have attracted attention from computer scientists, and psychologists. Yet we still have very limited understanding of how the identity-based patterns are learned as a morphophonological process when we factor in diverse typological variations, let alone their broader implications for models of language, morphophonological learning, and cognition. This is due to three missing lines of evidence.

Firstly, most of the earlier experimental studies have been primarily concerned with whether humans *can* differentiate specific reduplicative structures from non-reduplicative ones. Many other critical questions are left unexplored. For example, how can language learners recognize and generalize reduplicative patterns? On *what levels of phonological abstraction* do learners construct copying-based generalizations? These questions appear to tap into the core components of the grammar architecture and the learning mechanism. However, we currently lack a complete understanding of the answers to these issues.

Secondly, the investigated patterns in previous experiments were mostly canonical and simple cases of surface repetition. In a great number of studies within the speech domain (e.g., Marcus et al., 1999), the repeated sequences are usually CV syllables (e.g., *wo*, *fe*). The involved patterns were also restricted, including ABA (*wofewo*), ABB (*wofefe*), AAB (*wowofe*), XX (*wowo*), and XY (*wofe*). However, cross-linguistically, the reduplicative typology is much richer, providing an empirical basis for possible experimental testing. Their linguistic properties vary along important dimensions, including how much to copy (e.g., partial versus total), which portion to copy, the degree of phonological identities, and so on. We will review the details in Section 2.3. To our knowledge, only a few experimental studies have looked at the more typologically diverse patterns that bear on intricate theoretical consequences (see our discussions of Berent et al., 2016, 2017; Haugen et al., 2022 in Section 1.1). The design of these linguistically motivated experiments was mostly limited to forced-choice tasks, and most of them used orthography instead of auditory inputs. For reduplication patterns that involve phonological identity, it would also be valuable to collect more evidence from different experimental paradigms with ecologically valid design, for example, asking for free production responses and using auditory input as in our experiments here (see discussions in Moreton and Pertsova, 2024).

When it comes to the attested patterns across world languages, it is important to point out that some naturally occurring reduplicative rules, serving as motivations for theoretical proposals, lack a complete assessment of their productivity (see our discussion in Section 1.3). It remains unclear whether the hypothesized reduplicative patterns are actually the kind encoded by native speakers. Experimental investigations are necessary at this stage as another level of verification for the empirical status of certain patterns, thereby allowing us to better evaluate our theory.

2.2.2 The current study

The three major research questions studied in this chapter are summarized in (16).

(16) a. *Can human learners rapidly learn copying-based generalizations?*

When prompted with reduplication as a morphophonological process, can learners recognize the effects of copying and extend copying-based generalizations to novel forms?

b. *What inductive biases based on phonological abstraction do human learners exhibit?*

If multiple generalizations are equally compatible with the familiarized patterns, based on what levels of phonological abstraction do human learners form their generalizations?

c. *Are there learning differences among different attested patterns?*

Do different typologically attested patterns yield different learning results?

Artificial languages have been widely employed to study both surface repetition structures (see references above) and biases in phonological and morphophonological learning (e.g., Wilson, 2006; Moreton, 2008; Finley and Badecker, 2009; White, 2014; Haugen et al., 2022). Our experimental design followed the poverty of the stimulus paradigm (Wilson, 2006; also known as the extrapolation paradigm per Culbertson, 2023). This method has been widely adopted to study humans' analytic biases in language learning (e.g., in phonology, Wilson, 2006; in morphophonology, Wilson, 2022; in syntax, Culbertson and Adger, 2014; in compositionality and concept learning, Lake et al., 2019), best suited for the second research question we ask.

The general design of all experiments is as follows. In the following experiments, reduplication is not merely presented as repetitions within surface sound sequences (e.g., *wofefe* and *wowofe*), but as word-formation processes (e.g., *wofe* ↦ *wofefe* and *wofe* ↦ *wowofe*). English-speaking participants were prompted to learn pluralization in a new language. Each experiment consisted of a familiarization phase and a testing phase. In the familiarization phase, participants listened to a small number of singular-plural pairs. In the first series of experiments, some examples are [ˈdɔv.ɡə] and [dɔv-ˈdɔv.ɡə]; [ˈʃæp.mə] and [ʃæp-ˈʃæp.mə]; in the second series of experiments, some examples include [ˈpif] and [ˈpif-pif], [ˈzæb] and [ˈzæb-zæb]. All familiarized items were homogeneous in terms of their phonological properties. For example, in the first series of experiments, reduplicative patterns were all CVC patterns, and in the second series of experiments, reduplicative patterns were all monosyllabic copying. The familiarized patterns were compatible with multiple phonological generalizations. In the testing phase, participants were asked to apply what they had learned as the pluralization rule for novel singulars. To discover what generalizations participants extrapolated, the testing trials were designed to bear different predictions under different possible generalizations. For example, in the first series of the experiments, some novel singulars were [ˈstæb.ɡə], [ˈav.di] with target syllables showing different shapes from the CVC pattern; in the second series of the experiments, we tested participants on novel stems of greater lengths ([ˈteɪ.pə.ɡæb], [ˌɡɛ.zə.ˈseɪ.kə.dɪv]). In all, the input forms provided no disambiguating information on any of the possible generalizations. Thus, asymmetries in learners' responses likely revealed their learning biases.

The various hypotheses studied here target different granularities of phonological abstractions. To assess the scope of generalizations, Berent (2013) identified different scopes of generalizations based on the phonological properties of involved segments (novel phoneme, novel feature, etc.). These levels of scopes were later adopted by computational linguists as a metric for exploring the generalizing capacities of computational models (see, e.g., Prickett et al., 2022 on the performance of Sequence-to-Sequence networks). Besides ensuring variegated segments, we focused on the phonological properties of the whole copied sequence, seeking insights from theoretical linguistic proposals to characterize possible levels of abstraction. In particular, we studied whether the learned reduplicative structures should be characterized by fine-grained phonological features, consonant/vowel skeleton (Marantz, 1982; e.g. copy CVC for [ˈdɔv.gə] and thus [dɔv-ˈdɔv.gə]; [ˈʃæp.mə] and thus [ʃæp-ˈʃæp.mə]), segment count (Levin, 1985; e.g. copy three segments for [ˈdɔv.gə] and thus [dɔv-ˈdɔv.gə]; [ˈʃæp.mə] and thus [ʃæp-ˈʃæp.mə]), or more abstract prosodic shapes (McCarthy and Prince, 1986; e.g. copy a heavy syllable for [ˈdɔv.gə] and thus [dɔv-ˈdɔv.gə]; [ˈʃæp.mə] and thus [ʃæp-ˈʃæp.mə]). Within each series of experiments, we provide an in-depth discussion of the empirical and theoretical background for the familiarized patterns.

Compared to other studies, the studies here show some methodological advantages. We minimized the possibility of conscious letter-based strategies, by presenting all stimuli auditorily with no orthographic support. Participants were asked to give free spoken responses. This was more demanding than other tasks, such as alternative forced choices with orthographic input as in Haugen et al. (2022) and Berent et al. (2016, 2017). It better approximates morphophonological learning in natural language settings and appears to be more revealing of possible variation in participants' generalizations.

2.3 Dimensions of variations in the reduplicative typology as empirical basis

Reduplicative patterns exhibit cross-linguistic variation along many crucial dimensions; for an overview, see Moravcsik (1978) and Inkelas and Downing (2015). This chapter touches on three of them: the shape of a copy, the copied portion if partially reduplicated, and the degrees of phonological identity between copies. Table 2.1 and Table 2.2 illustrate the relevant variants with examples from natural languages. Following traditional terminology, throughout this chapter, we will use *reduplicant* to refer to the smaller-sized copy if it is applicable, and respectively, the *base* refers to the remaining sequences of segments in the reduplicated form minus the reduplicant. For example, for the form [pus-pusa], [pus] is the reduplicant, and [pusa] is the base. Note when reduplication is total, there is no such distinction between the smaller-sized reduplicant and the base.

Dimensions	*Variants*	*Language and family*	*Examples and glosses*	
			Base/stem	*Reduplicated*
How much to copy? i.e., the phonological shape	Total	Indonesian Austronesian McCarthy and Cohn (1998)	buku *'book'*	bu.ku-bu.ku *'book-*PL*'*
			ma.ša.ra.kat *'society'*	ma.ša.ra.kat-ma.ša.ra.kat *'society-*PL*'*
	Partial A bisyllabic foot	Diyari Pama–Nyungan Austin (1981)	pir.ta *'tree'*	pir.ta-pir.ta 'DIM- *tree'*
			wil.ha.pi.na *'old woman'*	wil.ha-wil.ha.pi.na 'DIM-*old woman'*
	Partial A light syllable	Tonkawa Coahuiltecan Gouskova (2007)	to.poʔs *'I cut it'*	to-to.poʔs 'REP-*I cut it'*
			xej.tsoʔs *'I rub him'*	xe-xej.tsoʔs 'REP-*I rub him'*
	Partial A heavy syllable	Ilokano Austronesian Hayes and Abad (1989)	kut.tóŋ 'thin'	naka-kut-kut.tóŋ 'ADJ-INTENS-thin'
			bu.téŋ 'afraid'	naka-but-bu.téŋ 'ADJ-INTENS-afraid'
	Partial Base-dependent	Hiaki, Uto-Aztecan Haugen and Hicks Kennard (2011)	vu.sa *'awaken'*	vu-vu.sa 'HAB-*awaken'*
			vam.se *'hurry'*	vam-vam.se 'HAB-*hurry'*

Table 2.1: *Typological variation focused in this chapter. Dimension I: the reduplicant shapes. The abbreviated morpheme glosses are as follows.* PL *: plural;* DIM *: diminutive;* REP*: repetitive;* ADJ*:adjective marker;* INTENS*: intensifier;* HAB*: habitual.*

Dimensions	*Variants*	*Language and family*	*Examples and glosses*	
			Base/stem	*Reduplicated*
Which part of the stem is copied if partially reduplicated	Left edge oriented	As illustrated in Table 2.1		
	Right edge oriented	Manam Austronesian Lichtenberk (1983)	salaga *'be long'*	salaga-laga *'long-*SG*'*
			sapara *'branch'*	sapara-para *'having branches'*
	The middle of a word (stress driven)	Samoan Austronesian Broselow and McCarthy (1983)	táa *'strike'*	ta-táa *'strike-*PL*'*
			alófa *'love'*	a-lo-lófa *'love-*PL*'*
			saváli *'walk'*	sa-va-váli *'walk-*PL*'*
Degrees of phonological identity	Perfect identity	As illustrated above and in Table 2.1		
	Imperfect identity with fixed segments Here, copied vowels are always [ə]	Kwak'wala Wakashan Saba Kirchner (2013)	loːq *'hemlock sap'*	lə-loχ-k 'T.M.*-hemlock sap-eat*
			paːs *'flounder'*	pə-paːs-sta-k 'T.M.*-flounder-in water-eat*
			q'əmdzək^w *'salmonberries'*	q'ə-q'əmdzək^w 'T.M.*-salmonberries-eat*
	Imperfect identity with fixed segments Here, copied vowels are always [i]	Doka Timur West Tarangan Austronesian Nivens (1993)	m-ɔn=na *'2s-shoot=it'*	min-m-ɔn=na 'NMLZ-*2s-shoot=it*
			jɛr-para 'NF-*burn*'	jɛr-pir-para 'NF-NMLZ-*burn*'
			jinay *'big'*	jin-jinay 'NMLZ-*big'*
	Imperfect identity with identity avoidance The copied vowels are always non-identical to the base vowels	Javanese Austronesian Yip (1995)	bul *'puff'*	bal-bul 'HAB.REP-*puff'*
			eliŋ *'remember'*	elaŋ-eliŋ 'HAB.REP-*remember'*
			tak *'tap'*	tak-tek 'HAB.REP-*tap'*

Table 2.2: *Typological variation focused in this chapter. Dimension II: which part of the stem is copied if reduplication is partial. Dimension III: degrees of phonological identity. The abbreviated morpheme glosses are as follows.* SG *: singular;* PL *: plural;* T.M.*: "too much";* NF*: nonfinite;* NMLZ*: nominalizer;* HAB.REP*: habitual-repetitive.*

For the phonological shapes of a reduplicant, when reduplication involves total copying, the copy is identical to the stem, which could grow unboundedly long together with the stem, as in Indonesian, [ma.ša.ra.kat] and [ma.ša.ra.kat-ma.ša.ra.kat]. In the case of partial copying, the phonological shape usually remains constant across all possible target bases, as illustrated in foot copying in Diyari (e.g., [wil.ha.pi.na] and [wil.ha-wil.ha.pi.na]), light syllable copying in Tonkawa (e.g., [xe-xej.tsoʔs]), and heavy syllable copying in Ilokano (e.g., [naka-but-bu.téŋ]), as detailed in Table 2.1. Scholars do debate whether natural languages display reduplicants with varying shapes dependent on the target base.[2] A few languages have been argued to exhibit such a base-dependent varying shape, such as Hiaki [vu-vu.sa] but [vam-vam.se]. What we can confidently conclude is that this phenomenon is extremely rare, if not all unattested.

The second dimension concerns the copied portion of the stem if partially reduplicated. Cross-linguistically, the reduplicants are found to appear as prefixes copying the left word edge as illustrated by the examples above, or as suffixes copying the right word edge (e.g, Manam, [salaga-laga]), or as infixes copying the middle part of a word and placed adjacent to the copied portion.[3] When infixes appear in the middle of a word, based on the typological survey in Yu (2003, pp. 8-9), they are either edge-driven or prominence-driven. Here, prominence means stress, as illustrated by Samoan, in which the stressed syllable is copied (e.g., [a-lo-lófa]).

Lastly, in terms of the degree of phonological identity, beyond simple perfect repetitions of a sequence, there are many languages with copies that are not exactly the same. Some cases of imperfect copying regarding feature a first-order hypothesis. That is, the same segments are held as constant across all targeted bases, as exemplified by Kwak'wala [lə-loχ-k] and

[2] For some argument for naturally occurring base dependence, see Haugen and Hicks Kennard (2011); Zukoff (M.S.). For some arguments against such an idea, see Inkelas and Zoll (2005).

[3] Another orthogonal dimension to consider is the relative position between the copies. All presented examples here demonstrate adjacent copying. Yet some languages exhibit non-adjacent copies, as in Creek (Muskogean) [holwak-iː] *'ugly'* and [**ho**lwa**ho**k-iː *'ugly*-PL]. More examples and discussions of non-adjacent copies can be found in Section 3.7 of Chapter 3. In all experiments, the trained patterns always contained adjacent copies. We plan to directly investigate the adjacency of copied portions with similar experiments in the future.

[q’ə-q’əmdzək^{w}]. Other cases reflect identity avoidance, such as Javanese [elaŋ-eliŋ] for the stem [eliŋ]. Here, the fixed vowel is [a]. But when the stem is [tak], the reduplicated form is [tak-tek] but not [tak-tak]. This is a higher-order generalization – the vowel in one copy is *always* different from the vowel in another copy.

2.3.1 The use of surface repetitions in English

In line with previous research on reduplication (Berent et al., 2016, 2017; Haugen et al., 2022), participants in our experiments were English native speakers due to their rare exposure to reduplication as a morphophonological process. It is commonly known that English lacks productive reduplicative morphemes (Rubino, 2013), though copying constructions/surface repetitions do exist in English, as shown in (17). These constructions, never as parts of the inflective system, have been argued to satisfy “expressive,” “aesthetic,” and/or equivalent extra-grammatical purposes (Mattiello, 2013). The involved linguistic constituents are usually bigger than a word. Consequently, they may be “not subject to the same conditions as rules of plain morphology” (Zwicky and Pullum, 1987, p. 338). On the other hand, we are interested in phenomena smaller than a word, making English speakers a perfect group of participants for our purposes.

(17) Surface repetitions in English

a. Echo (Yiddish-derived; Nevins and Vaux, 2003)
Examples: metalinguistic-shmetalinguistic, reduplication-shmeduplication

b. Ablaut (Minkova, 2002)
Examples: chit-chat, dilly-dally, zig-zag

c. Rhyme (Minkova, 2002)
Examples: teenie-weenie, super-duper

d. Contrastive focus (Ghomeshi et al., 2004)
Examples: I will make the tuna salad, and you make the SALAD-salad.

2.4 Experiment series 1: Inside a syllable

2.4.1 Learning reduplication, but at what levels of abstraction within a syllable?

Imagine that a learner encounters two singulars [ˈdɔv.gə], [ˈʃæp.mə] coupled with their pluralized forms [dɔv-ˈdɔv.gə] and [ʃæp-ˈʃæp.mə] respectively. Logically speaking, the learner could attribute these different surface forms to either reduplication or other morphophonological processes, such as /dɔv-/ or /ʃæp-/ prefixation. In other words, in order to learn reduplication, they ought to first recognize the effects of copying, namely the identity-based relations. Then, they need to construct a generalization about the realization of the copying at a certain abstract level. Lastly, they should relate the generalization to the designated operation, here pluralization.

Yet, at what level of abstraction will the learner form a generalization? Thinking about this question from the learner's perspective, we can form a taxonomy of granularity for the phonological abstractions of possible generalizations. Assuming the input [dɔv-ˈdɔv.gə] and [ʃæp-ˈʃæp.mə], some logically possible generalizations are summarized in (3), roughly in an increasing order of coarse-grainedness (and thus, in a decreasing order of phonological specifications).

(18) Logically possible generalizations assuming input [dɔv-ˈdɔv.gə] and [ʃæp-ˈʃæp.mə], among many others.

a. specified with features at each slot (feature-specific) $C_{[\text{coronal}]}V_{[\text{-high}]}C_{[\text{labial}]}$

b. specified with syllabicity (CV skeleton) CVC

c. specified with the number of segment slots (segmental counting) XXX

d. as a prosodic unit with specified weights (base-independent) $\sigma_{\mu\mu}$

e. as a prosodic unit without specified weights (base-dependent) σ_1

Now, let us examine each possible generalization in more detail. For a cautious, bottom-up learner that tries to find a hypothesis that is as tight as possible but allows generalization (e.g., minimal generalization learning; Albright and Hayes, 2003), they might form a rather restrictive template, each slot of which contains some (if not all) shared phonological features. For example, computing [dɔv] and [ʃæp], the learner would end up copying a coronal consonant, followed by a non-high vowel and a labial consonant, i.e., $\mathrm{C}_{\text{[coronal]}}\mathrm{V}_{\text{[-high]}}\mathrm{C}_{\text{[labial]}}$.

The learner can be nonchalant with *some* features and only pay attention to particularly salient ones. One such possibility is syllabicity, or the consonant/vowel status of each realized segment (Marantz, 1982). The psychological plausibility of this level of abstraction is independently supported by a sequence of word priming and computational work on word-formation processes in Semitic languages like Arabic (e.g., Boudelaa and Marslen-Wilson, 2004; Dawdy-Hesterberg and Pierrehumbert, 2014).[4] Under such a generalization, a learner would copy the first CVC sequence.

When moving beyond the sensitivity to syllabicity, one can extract generalizations with a unifying notion of segments, in other words, at a segment-skeleton level (Levin, 1985). A segment-based counting hypothesis could be adopted. In this case, the learner would copy the first three segments, e.g., [avd-avdi]. Last but not least, they might well seek a hypothesis at a level characterized by the vocabulary of abstract prosodic units, yielding a hypothesis of copying a heavy syllable independent of the base (McCarthy and Prince, 1986). If adopting this hypothesis, in the generalization phase, a learner is expected to produce [stæb-ˈstæb.gə] for the novel singular [ˈstæb.gə] and [dɛb-ˈdɛ.beɪ] for the novel singular [ˈdɛ.beɪ]. Or, they can also learn to copy the first syllable of the base (Haugen and Hicks Kennard, 2011). Under this hypothesis, participants are expected to still produce [stæb-ˈstæb.gə] for [ˈstæb.gə] but [dɛ-ˈdɛ.beɪ] for the novel singular [ˈdɛ.beɪ]. [5] For more details, see Table 2.4.

We controlled the shapes of the familiarized stems and their reduplicated forms in the same way as described above, except that in the actual experiments, participants were provided with eight such pairs instead of two. To study what hypotheses participants have formed after the familiarization phase, we designed seven types of novel singulars for test-

[4] See effects of template priming in a meta-analysis of non-concatenative morphology in Semitic languages (Xu et al., 2023).

[5] The space might contain many more hypotheses than what is discussed here. We only selected the most representative ones, which were proposed previously in the literature of phonological theory or were proved cognitively plausible to a certain degree. We revisit this question in the context of the experimental results.

ing, as in Table 2.3. Familiar-type items, such as [ˈzɛv.du] test whether participants are able to extend the generalization to novel stems that share the same restrictions as the familiarized items. Further, they provide a baseline measure of how well the copying-based generalization is learned. To investigate whether the feature-specific generalizations hold, Lab-Cor and High-Vowel-type stimuli create unseen feature combinations. Four additional types alter the prosodic shapes of the stems (Singleton, Rising, Complex, Onsetless), aiming to delineate the level of granularities of the adopted generalizations. Table 2.4 shows the predicted reduplicants under different copying-based hypotheses – Section 2.4.3.1 details how these items were generated.

	Familiar	$C_{[cor]}$	$V_{[-high]}$	$C_{[lab]}$	C	V	ˈzɛvdu
Segment manipulations	Lab-Cor	$C_{[lab]}$	$V_{[-high]}$	$C_{[cor]}$	C	V	ˈfædno͡ʊ
	High-V	$C_{[cor]}$	$V_{[+high]}$	$C_{[lab]}$	C	V	ˈʃipne͡ɪ
Shape manipulations	Singleton	$C_{[cor]}$	$V_{[-high]}$	$C_{[lab]}$	$\varnothing$	V	ˈdɛbe͡ɪ
	Rising	$C_{[cor]}$	$V_{[-high]}$	$C_{[lab]}$	$C_{[son]}$	V	ˈtæpɹe͡ɪ
	Complex	$C_{[cor]}$ C	$V_{[-high]}$	$C_{[lab]}$	C	V	ˈstæbgə
	Onsetless	$\varnothing$	$V_{[-high]}$	$C_{[lab]}$	C	V	ˈavdi

Table 2.3: *A summary of the testing types with example stimuli. The red color indicates how these manipulations were implemented (see Section 2.4.3.1 for more detailed descriptions). In brief,* Familiar *kept the segmental restrictions the same as the familiarized items.* High-V *changed the vowel in the target syllable to [+high].* Lab-Cor *flipped the place of articulation restrictions of the consonants in the target syllable.* Singleton *only allowed one single word-medial consonant, removing the coda from the target syllable.* Rising *made the word-medial cluster a legal onset, thus also removing the coda from the target syllable.* Complex *added another consonant to the onset of the target syllable.* Onsetless *makes the target syllable lack of an onset.*

Testing types	**Example**	$\mathbf{C_{[coronal]}V_{[-high]}C_{[labial]}}$	**CVC**	**XXX**	$\sigma_{\mu\mu}$	$\sigma_{\mathbf{1}}$
Familiar	ˈzɛv.du	zɛv	zɛv	zɛv	zɛv	zɛv
Lab-Cor	ˈfæd.noʊ	–	fæd	fæd	fæd	fæd
High-V	ˈʃip.neɪ	–	ʃip	ʃip	ʃip	ʃip
Singleton	ˈdɛ.beɪ	dɛb	dɛb	dɛb	dɛb	dɛ
Rising	ˈtæ.pɹeɪ	tæp	tæp	tæp	tæp	tæ
Complex	ˈstæb.gə	sæb/tæb	sæb/tæb	stæ	stæb	stæb
Onsetless	ˈav.di	–	–	avd	av	av

Table 2.4: *The predicted reduplicant(s) under each copying-based generalization. Dashes indicate the predicted failure to apply reduplicative rules, but do not indicate that they will provide no realization. Other generalizations are also possible, such as an allomorphy-based analysis (i.e. prefixing one of the affixal forms they have encountered in the familiarization phase). The variations should be freer than what is presented here.*

2.4.2 Different reduplicative patterns in three experiments

The other dimension investigated in this experimental series concerns the degree of identity between the two copies. We are particularly interested in cases where non-identities are created by fixing some segments as constant across different targeted bases. These patterns inherently show a mixture of active copying and fixed phonological materials. We further set the fixed phonological materials to be internal to a copy, which could interrupt the copying operation. The attested structures in natural languages offer the empirical bases for the familiarized patterns tested in our experiments. Copied segments in Experiment 1a (Section 2.4.3) repeat the corresponding base segments perfectly. Experiment 1b (Section 2.4.4) changes the vocalism in the reduplicants to a fixed mid-central [ə]. Experiment 1c (Section 2.4.5) changes the vocalism in the reduplicants to a fixed high front vowel [i]. Figure 2.1 provides intuitions of the three patterns.

The learning question in Experiment 1b and 1c differs from that in Experiment 1a. Again, the copied vowels are not identical to the base vowels but are fixed across different bases. This requires the learner to balance the non-identity between copies of the same stem and the invariant vowel quality across different stems, which consequently creates ambiguity between a first-order generalization (i.e. always has a fixed vowel regardless of the base) and a higher-order one (i.e. always ensure non-identity between the reduplicant and the base).

Using Experiment 1c as a concrete example, a first-order generalization would predict the vowel to always be a fixed [i]. Hence, for a novel singular [ʃip.ne͡ɪ], participants would be predicted to produce [ʃip-ʃip.ne͡ɪ], without worrying about the created perfect repetitions absent in the input. Another possibility is when learners notice that the vowels in the base and the vowels in the copy are always non-identical and are different in vowel heights. The logically plausible higher-order hypothesis, hence, avoids creating perfect repetitions. Instead, it enforces a difference between the copied vowel and the base vowel, either in the form of more fine-grained restrictions on high feature specifications or a more general segmental non-identity. Such an identity-avoidance tendency is indeed attested (for vowels, see Javanese in Table 2.2; for consonants, see McCarthy, 1986). Then, when encountering the novel singular [ʃip.ne͡ɪ], a learner who learns a higher-order, identity-avoidance hypothesis would never produce [ʃip-ʃip.ne͡ɪ]. For this question, participants' responses for the testing type **High-V** will be relevant.

We picked [ə] and [i] as the fixed vowels because it has been argued that human learners are sensitive to the phonetic substances in (morpho)phonological learning (e.g., Wilson, 2006; White, 2017). Given that the familiarized base vowels are all non-high, [ə] is phonetically closer to the non-high vowels, leading to shorter mappings than [i] and more predictable ones. On the other hand, the mappings between non-high vowels and the high front [i] are of larger phonetic distances, more likely to lead to a hard encoding of vowel overwriting. If any procedure in the learning process bases the computation on the mappings between two copies,

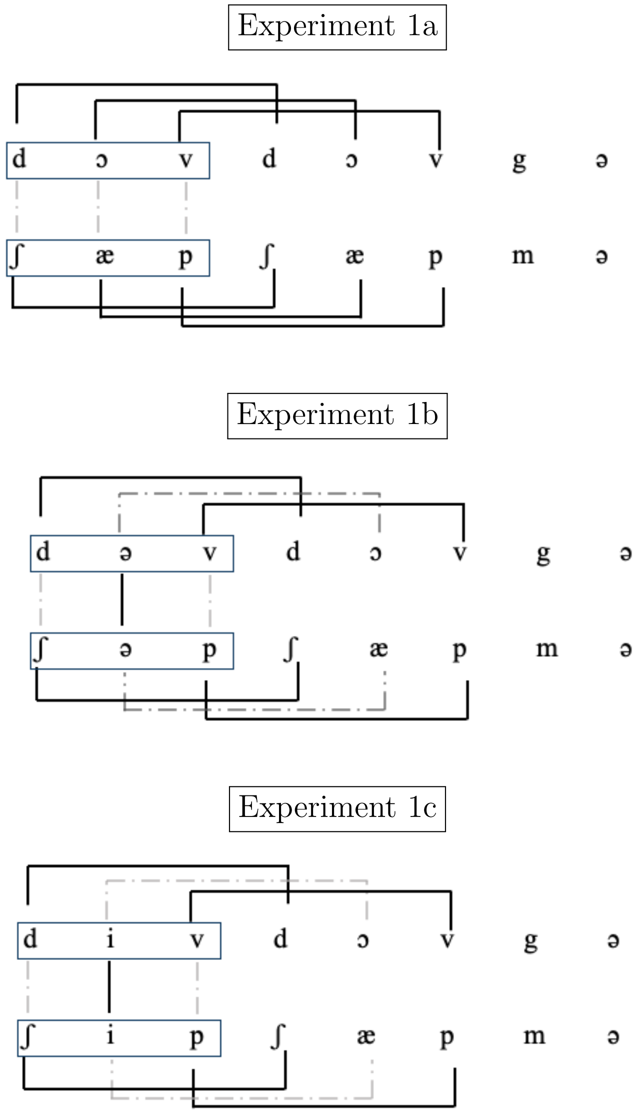


Figure 2.1: *The dependency shapes between the two copies in the reduplicated forms (top: Exp 1a; middle: Exp 1b; bottom: Exp 1c). Boxes represent the affixal parts. Solid black lines represent segment-to-segment identities. Dashed grey lines indicate segment-to-segment non-identities. The darker correspondences between [ə] and ɔ, æ (middle) indicate shorter phonetic distances than mappings between [i] and ɔ, æ (bottom).*

then a learner might be in an easier position to form an analysis grounded by phonological principles when the fixed vowels are [ə] than when the fixed vowel is [i]. On a separate note, participants in our experiments are native English speakers. English phonology often has stressless vowels reduced to [ə] but never to [i]. This might further nudge participants to construct a phonological analysis for the fixed [ə] but presumably, some hard-specification account for [i].

The phenomenon studied here is known as fixed segmentism in theoretical phonology literature. Alderete et al. (1999) argued for two different kinds of fixed phonological materials: one is due to the emergence of phonological unmarked structures and the other is truly morphological. In fact, Experiment 1b with a fixed [ə] could be seen as a case of a phonological fixed segment, and Experiment 1c with a fixed [i] could align with a morphological account. However, for now, we would like to stay agnostic to the question of whether there are hard distinctions between the different types of fixed materials. Likewise, we do not assert that learners deal with fixed vowels in the way we described, since the premise of entertaining the vowel mappings might be simply wrong. After all, we are interested in seeing whether minimal changes in the familiarized patterns, here fixing the vowel qualities internal to the copy, would lead to any differences in the learning outcomes, and if so, in what ways.

2.4.3 Experiment 1a: Perfect identity

2.4.3.1 Methods

Participants 23 participants were recruited through the Amazon Mechanical Turk crowdsourcing platform, and were compensated with $3.50 for completing the experiment (13 females, 8 males, 1 unreported; mean age = 42.36, age range = 26 -72). Data from 20 participants were transcribed and analysed. Two participants were excluded due to missing responses caused by experiment coding errors. One participant was replaced due to failure to understand the task, producing full reduplication suffixed with English plural marker. Among the participants whose data were analyzed, all were self-identified as monolingual English speakers. Three participants reported exposure to other languages (German, Italian, and Spanish), none of which contains productive partial reduplication.

Materials Now, we discuss how stimuli were created. For motivations and rationale, see our previous discussions in Table 2.4 and Section 2.4.1. The familiarization phase consisted of eight pairs of singulars and plurals (Appendix B.1). All singulars were disyllabic $C_1V_2C_3C_4V_5$ non-English words (e.g. dɔvgə, ʃæpmə). The corresponding plurals had the initial $C_1V_2C_3$ repeated locally, thus of the shape $C_1V_2C_3$-$C_1V_2C_3C_4V_5$ (e.g. dɔv-dɔvgə, ʃæp-ʃæpmə). C_1 was drawn from a set of coronal obstruents /t, d, ʃ, z/, V_2 was one of the non-high vowels /ɛ, æ, ɔ, ɑ/, and C_3 was from a set of labial obstruents /p, b, f, v/. For each of the three positions, candidate phonemes occurred an equal number of times, namely twice. C_4 was selected from filler obstruents and nasals /p, b, t, d, k, g, f, v, z, ʃ, tʃ, ʤ, m, n/ and was set to be distinct from C_1 and C_3. V_5 was a legal word-final vowel from /i, u, ə, e͡ɪ, o͡ʊ/. To maximize perceptual saliency and articulatory ease, we prevented C_3C_4 from being fricative-fricative sequences and obstruent clusters disagreeing in place of articulation or voice.

English segmental sequencing restrictions (i.e. phonotactics) disallow obstruent-obstruent and obstruent-nasal sequences to be onsets of a syllable (Hayes and Wilson, 2008, p.397). Thus, the segmental restrictions imposed on C_3 and C_4 guaranteed a syllable boundary in between. In other words, English speakers would parse [dɔvgə] as [dɔv.gə] but never *[dɔ.vgə], or *[dɔvg.ə]. Such a syllable parse produced a highly ambiguous familiarization phase since there were multiple generalizations that participants could adopt.

To tease apart these generalizations effectively and investigate the learning biases over them, we designed testing items that would show different predicted reduplicative forms under each possible generalization. For visualizations of these manipulations with sample stimuli, we refer the readers to Table 2.3. As a baseline, testing items contained novel forms sharing the same restrictions on segments as the familiarized items, hence **Familiar** type. We then took those **Familiar** items and applied one of the six manipulations to generate more forms. The first two types of manipulations focused on segmental restrictions: High-V had the non-high V_2 changed to high vowels /i, u/ instead of non-high /ɛ, æ, ɔ, ɑ/; **Lab-Cor** flipped the place restrictions of C_1 and C_3, and created forms whose target syllable had labial onsets and coronal codas.

We performed four other manipulations targeting the prosodic shape of the first syllable, with two (SINGLETON, RISING) changing the coda, and two (COMPLEX, ONSETLESS) changing the onset. Specifically, SINGLETON removed C_4 and hence, created only a word-internal consonant. This consonant would, in turn, be parsed as the onset of the following syllable. RISING changed C_4 to a liquid /ɹ, l/ such that C_3C_4, with sonority rise, formed licit onsets of a syllable. COMPLEX had its onset selected from /st, dɹ, ʃɹ, sl/ respectively so that the new singular starts with two consonants. ONSETLESS deleted C_1 and formed a target syllable with no onsets. After performing these changes to all FAMILIAR items, we excluded forms attested in the English lexicon, obtaining a set of candidate testing items to choose from.

The testing phase included 56 items, eight for each of the seven testing types, as indicated in Table 2.3. To ensure the robustness and generality of the results, we generated seven such testing lists. Within each list, the unigram and bigram frequencies were balanced across all testing items and within items of each testing type.

Each noun singular was randomly paired with a picture of an individual everyday object. The noun plurality was indicated by showing two instances of the same object, as illustrated in Figure 2.2. To enhance English noun-like-ness (Hayes, 2009, p.245), we placed stress on the initial syllable of these disyllabic singulars, as reflected in Tables 2.4 and Table 2.3. The reduplicated plurals in the familiarization phase had stress on the second syllable (e.g., dɔvˈdɔvgə, ʃæpˈʃæpmə). Stimuli were synthesized through the neural engine of Amazon Polly with Matthew Voice, prompted with their phonetic transcriptions. The synthesized tokens were resampled at a rate of 24 kHz and normalized for intensity to 65 dB.

Procedures The experiment was conducted online. Participants were instructed to participate in the study from a quiet room and encouraged to wear headphones throughout the experiment. After giving their consent, participants were required to complete an audio test for their web microphones and headphones/external speakers. The main experiment started once they successfully passed the audio test. Otherwise, they were kindly asked to return

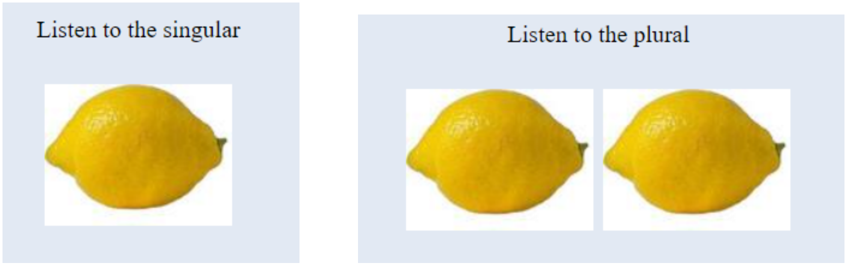


Figure 2.2: *Example familiarization trial*

the study.

The main experiment consisted of two phases: a familiarization phase and a testing phase. Before the familiarization phase, participants were told that they would be learning plural formation in a new language. In each familiarization trial, participants listened to a disyllabic singular, followed by the reduplicated plural, with pictures appearing on the screen. They were asked to repeat the reduplicated plural forms and record their repetitions. After eight pairs of exposure, participants proceeded to the testing phase and were prompted to apply what they learned in the familiarization phase to novel singulars. Each participant was randomly assigned to one of the seven testing lists, and we ensured each list had at least two participants. For each list, 56 items were tested together in a randomized order, and each item was tested once. In each testing trial, participants listened to a novel singular and recorded their production response for the corresponding plural form.

The experiment ended with a questionnaire, asking participants to describe the formed rules if possible. They were also asked whether there were perceptual difficulties for any specific segments or forms during the experiment. The full experiment took roughly 25 minutes to complete.

2.4.3.2 Analyses and results

Though collected online, the data were straightforward and clear to transcribe. In line with traditional linguistic practices, we annotated the participants' plural responses into two parts: a base and an affix.[6] We identified the maximal consecutive segments that match

[6] Here, we labeled them as "affix" to be neutral.

the corresponding singular as the base, with the remaining segments as the affix realization. That is, assuming a response [zav.ˈziv.gi] for [ˈzɔv.gi], we annotated [ˈziv.gi] as the base and [zav] as the affix. To quantify the degree of copying within responses, we first performed string alignments to match the affix and the base segment-by-segment. This was done in two steps: we first aligned the vowels together, and then followed the ALINE algorithm described in Kondrak (2002) to align the remaining segments. We used the average segment similarity as the metric to assess the identity between the reduplicant and the base, where the averaged segment similarity itself was calculated based on the proportion of the shared phonological features, following Pierrehumbert (1993). [7]

Repetition accuracy of the familiarized items The repetitions of the trained items showed high accuracy. The repeated plural forms greatly followed the trained reduplicative patterns. The repeated responses of the plural forms largely maintained the identity relations. For all participants, the mean averaged segmental similarity between the affix and the base exceeded 0.95. That is, for the pairs of segments assessed, they shared approximately 26.6 out of 28 features on average. The sporadic feature mismatches involved voicing, place of articulation, and vowel height.

Are the generalizations copying-based? As discussed before, with the trained items, rather than adopting a copying-based generalization, it is logically possible to form an allomorphy-based generalization where the trained affixes are memorized as fixed forms of the plural realizations. To evaluate whether the generalizations involve some amount of copying, we conducted a Monte Carlo simulation. Specifically, we randomly shuffled the produced affixes and recombined them with the bases for each participant. After each shuffling, we recorded the average segmental similarity between the affixes and the base in the recombined forms. For an intuitive illustration, see Figure 2.3.

The rationale for a Monte Carlo approach is as follows. Considering one of the familiarized

[7] We employed the phonological feature set for English phonemes and allophones from Hayes (2009), which can be accessed via https://linguistics.ucla.edu/people/hayes/120a/Index.htm.

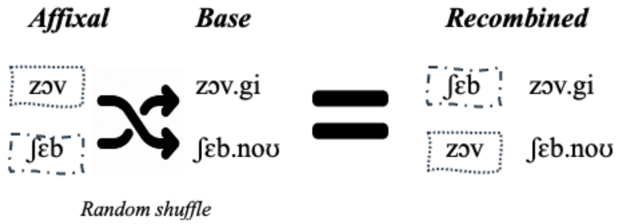


Figure 2.3: *An illustration of the random shuffling of the produced affixal parts and the bases.*

items [ʃæpˈʃæpmə], let us think about the novel singulars [ˈzɔv.gi] and [ˈʃɛb.no͡ʊ]. A copying-based generalization, producing the plural forms [zɔv.ˈzɔv.gi] and [ʃɛb.ˈʃɛb.no͡ʊ], undoubtedly yields high affix-base similarity. Recombining the affixes with the bases leads to forms such as [ʃɛb.ˈzɔv.gi] and [zɔv.ˈʃɛb.no͡ʊ], as illustrated in Figure 2.3. Such a recombination will inevitably lower the average affix-base similarity. Given that the testing target syllables were different in forms, a robust copying generalization is expected to lead to a significant decrease in similarity when forms are recombined. However, an allomorphy-based generalization works oppositely. If the participants ended up learning to prefix /ʃæp-/, the actual produced forms would be [ʃæp.ˈzɔv.gi] and [ʃæp.ˈʃɛb.no͡ʊ]. Random shuffles of the produced affixes within all responses will not decrease the average affix-base similarity.

Indeed, as Figure 2.4 reveals, for the average segment similarity, all participants in Experiment 1 had their observed responses fall greater than the 99% confidence interval of chance obtained by the Monte Carlo procedure, indicating that all participants performed some amount of copying in their actual responses.

To further evaluate whether participants adopted a copying-based generalization, we examined how the base segments were realized in the affix. Figure 2.5, Figure 2.6, Figure 2.7 show the averaged proportion of the segmental realizations in the affixal forms in the onset, nucleus/vowel, and coda respectively. The base segments were predominately realized identically to the affix segments, further supporting the copying-based generalizations. Some systematic changes occurred at a relatively lower rate – these included vowel reduction to [ə], and no realizations of the coda consonants.

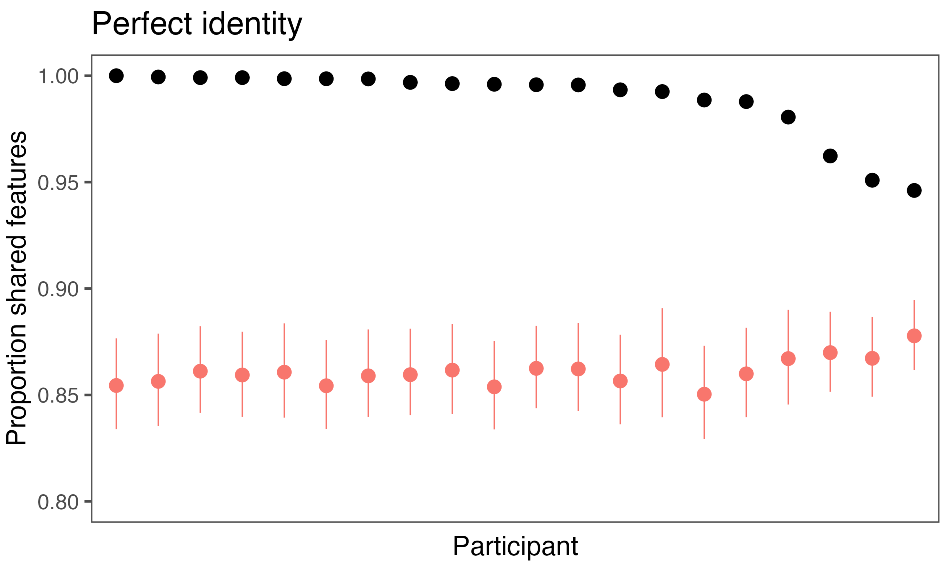
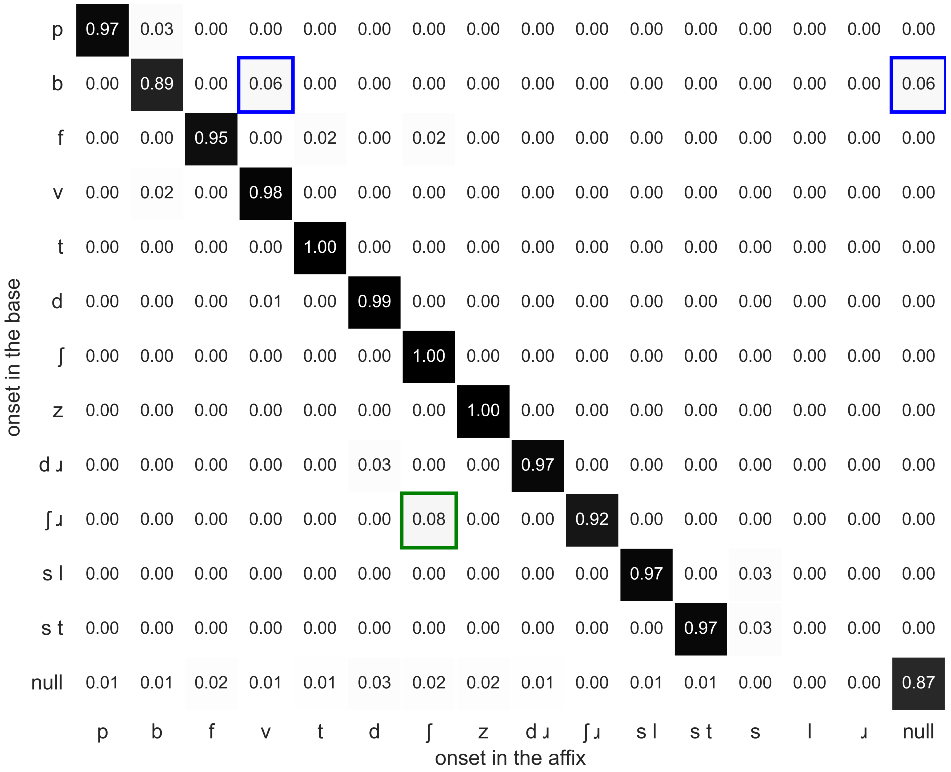


Figure 2.4: *The average segment similarity between the affixes and the novel singular bases. Black dots: the observed responses from each participant. Red dots: the mean calculated in the Monte Carlo procedure (R = 10000) with bars indicating the 99% confidence interval of chance.*



	p	b	f	v	t	d	ʃ	z	d ɹ	ʃ ɹ	s l	s t	s	l	ɹ	null
p	0.97	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
b	0.00	0.89	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06
f	0.00	0.00	0.95	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
v	0.00	0.02	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
t	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d	0.00	0.00	0.00	0.01	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ʃ	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
z	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d ɹ	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ʃ ɹ	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00
s l	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.03	0.00	0.00	0.00
s t	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.03	0.00	0.00	0.00
null	0.01	0.01	0.02	0.01	0.01	0.03	0.02	0.02	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.87

Figure 2.5: *The averaged proportion of the affix onsets conditioned on the base onsets. Average faithful realizations (diagonal): 0.96*

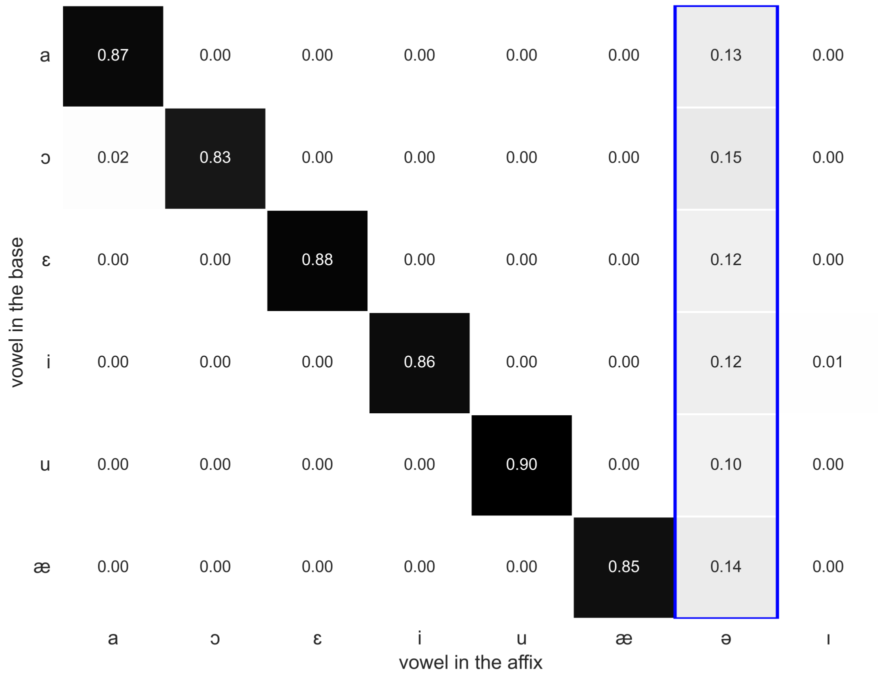


Figure 2.6: *The averaged proportion of the affix nuclei conditioned on the base. Average faithful realizations (diagonal): 0.87; Average reduction (the column ə): 0.13*

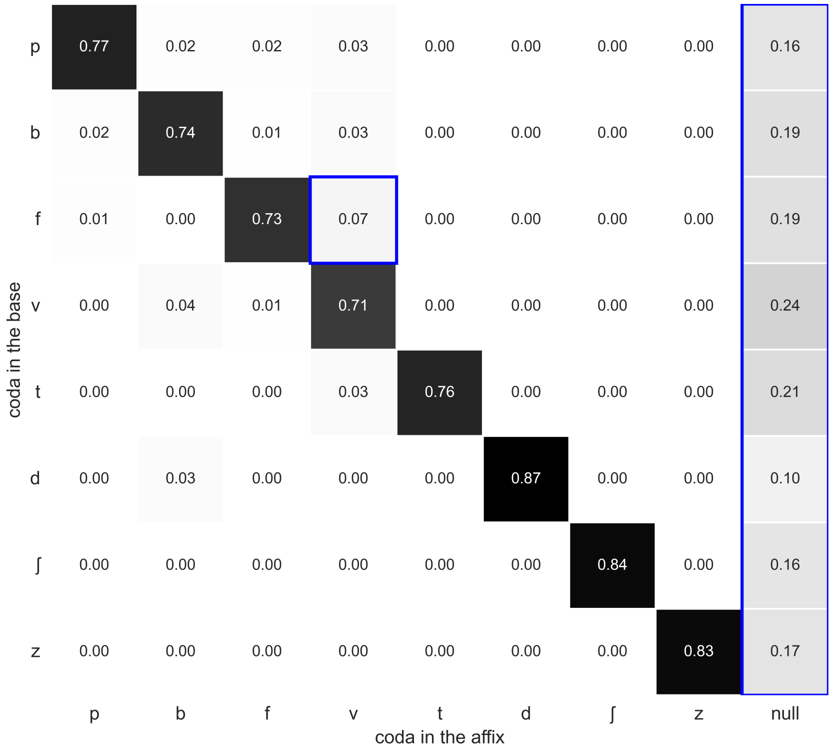


Figure 2.7: *The averaged proportion of the affix codas conditioned on the base codas. Average faithful realization (diagonal): 0.78; Average no coda realization (the column "null"): 0.18*

Generalizing at what levels of abstraction? Having established that all participants formed some copying-based generalizations, we now turn to the question of at what levels of abstraction they had formulated their generalizations.

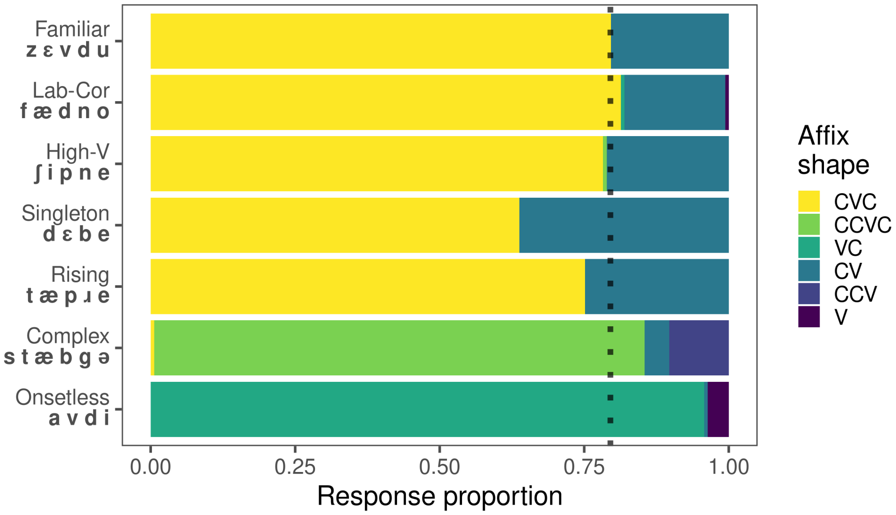


Figure 2.8: *The averaged proportion of affix shape conditioned on testing type; vertical line: the baseline rate of coda incorporation based on the* Familiar *type, namely when the target syllable has a coronal onset, non-high vowel, and a labial coda.*

Figure 2.8 shows the averaged proportion of different affixal shapes conditioned on each testing type. The Familiar type testing items had the CVC (e.g. [zɛv-ˈzɛv.du]) and CV (e.g. [zɛ-ˈzɛv.du]) affix shapes, with CVC as the more frequent one. Manipulating the segment inventory in the target syllable did not change the rate of onset realization and coda incorporation. Lab-Cor and High-V had the CVC shape occurring at the same rate as the Familiar type. On the other hand, manipulating the shapes of the target syllables led to varying affix shapes. Looking at the two types of coda manipulation that resulted in no coda in the targeted syllable, namely, the Rising type and the Singleton type, affixes still largely exhibited CVC shapes (e.g. for Rising [tæp-ˈtæ.pɹe͡ɪ] and for Singleton: [dɛb-ˈdɛ.be͡ɪ] respectively), despite that the incorporated codas were onsets of the following syllable in the base. Notably, for the Singleton type, the CV affix shape (e.g. [dɛ-ˈdɛ.be͡ɪ]) occurred at a higher rate when compared to other testing types. Based on the response proportions in testing types with onset manipulations, participants produced base-dependent onset shapes: responses for the Complex type showed CCVC shapes (e.g. [stæb-ˈstæb.gə]) and the Onsetless type showed VC affixes (e.g. [av-ˈav.di]).

These results were further supported by Bayesian regression modeling. We ran Bayesian mixed-effects multinomial logistic regression models with rStan (Stan Development Team, 2024). The models were run for the onset and the rime separately. The dependent variables were the phonological shapes, which had three levels in the onset model (C, CC, and ∅), and two levels in the rime model (VC and V). We included the testing types as fixed effects and by-subject random slopes. Sampling was done with the `sampling` function in rStan, which uses Monte Carlo Markov Chain (MCMC; specifically, a No-U-Turn sampler) techniques to generate samples from the posterior distributions for each parameter. We ran four chains, each of which had the first 1,000 iterations treated as warm-up and then the following 4,000 iterations as posterior draws, leading to 16,000 samples in total.

Given the sampled population-level parameters, we calculated the sampled posterior probabilities of each possible level based on the phonological shapes for each testing type. Figure 2.9 shows the posterior probabilities of the phonological shapes conditioned on the testing types, from which the same trends described above could be observed. In Table 2.5, we provide the mean of the sampled posterior probabilities for each testing type. Based on these posterior probabilities of the phonological shapes, we performed pairwise comparisons within each testing type and between testing types. For the prior and model specification, see Appendix B. The priors for the parameters assumed a mean-zero normal distribution with varying standard deviations. A summary of within-testing type comparisons can be found in Table 2.11. Table 2.6 provides a summary of between testing type comparisons.

The onset model provides clear evidence that participants produced base-dependent onset shapes. Within all testing types except the Complex type and the Onsetless type, the probability of producing a C is significantly higher than producing a complex onset CC, and producing no onsets (all $p < 0.01$). For the Complex type ([ˈstæb.gə]), the probability of producing a complex onset CC is significantly higher than producing a simple onset C, and producing no onsets (all $p < 0.01$). As for the Onsetless type ([ˈav.di]), the probability of producing onsetless responses is significantly higher than the other two levels (all $p < 0.01$).

Testing type t		$\overline{p(C\|t)}$	$\overline{p(CC\|t)}$	$\overline{p(\varnothing\|t)}$	$\overline{p(VC\|t)}$	$\overline{p(V\|t)}$
Familiar	ˈzɛv.du	$\boxed{\mathbf{0.99}}$	0.00	0.01	$\boxed{\mathbf{0.87}}$	**0.13**
Lab-Cor	ˈfæd.noʊ	$\boxed{\mathbf{0.98}}$	0.00	0.01	$\boxed{\mathbf{0.90}}$	**0.10**
High-V	ˈʃip.neɪ	$\boxed{\mathbf{0.99}}$	0.00	0.01	$\boxed{\mathbf{0.86}}$	**0.14**
Singleton	ˈdɛ.beɪ	$\boxed{\mathbf{0.99}}$	0.00	0.01	$\boxed{\mathbf{0.70}}$	0.30
Rising	ˈtæ.pɹeɪ	$\boxed{\mathbf{0.99}}$	0.00	0.01	$\boxed{\mathbf{0.83}}$	**0.17**
Complex	ˈstæb.gə	0.03	$\boxed{\mathbf{0.96}}$	0.01	$\boxed{\mathbf{0.94}}$	0.06
Onsetless	ˈav.di	0.01	0.01	$\boxed{\mathbf{0.98}}$	$\boxed{\mathbf{0.99}}$	0.01

Table 2.5: *The mean of the sampled posterior probabilities calculated based on the population-level parameter estimates*

Onset: C	All testing types with a C as the onset > Complex, Onsetless	Base dependent onset shapes
Onset: CC	Complex > all other types	
Onset: Ø	Onsetless > all other types	
Rime: VC	Complex, Onsetless > Singleton Onsetless > Familiar, Lab-Cor, High-V, Singleton, Rising	1. Marginal support for copying σ₁ 2. Avoidance of consecutive unreduced vowels, reflected by avoidance of V rime in the *Onsetless* type
Rime: V	Singleton > Complex, Onsetless Familiar, Lab-Cor, High-V, Singleton, Rising > Onsetless	

Table 2.6: *A summary of pairwise comparisons based on the sampled posterior probabilities between testing types in Experiment 1a. ">" indicates that the probability of showing a level is significantly higher than the other ($p < 0.01$).*

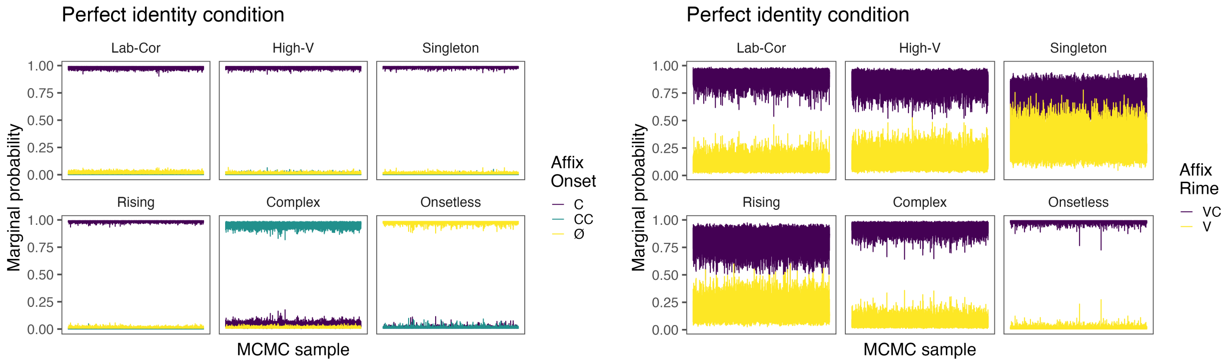


Figure 2.9: *The posterior probabilities of the produced shapes given each testing type.*

The rime model confirms the preference to incorporate the coda, regardless of whether a coda is present in the target base syllable. For all testing types except the Singleton type, the probability of a VC rime is significantly higher than that of a V rime (all $p < 0.01$). As for the Singleton type ([ˈdɛ.beɪ]), the probability of a VC rime ([dɛb-ˈdɛ.beɪ]) is marginally higher than that of a V rime without a coda ([dɛ-ˈdɛ.beɪ]; $p = 0.039 < 0.05$).

Additionally, the rime model shows marginal support for the base-dependent syllable copying generalization. Looking at the between-testing-type comparisons, the probability of producing a V rime for the Singleton testing type ([ˈdɛ.beɪ]) is significantly higher than the Onsetless ([ˈav.di]) and the Complex types ([ˈstæb.gə]; all $p < 0.01$), and marginally higher than the Lab-Cor type ([ˈfæd.noʊ]; $p = 0.026 < 0.05$), though not significantly higher than the Familiar type ([ˈzɛv.du]; $p = 0.07$), the High-V type ([ˈʃip.neɪ]; $p = 0.085$) and the Rising type ([ˈtæ.pɹeɪ]; $p = 0.16$).

On another note, the rime model also identified a phonologically grounded avoidance of marked sequences. The probability of producing a VC rime within the Onsetless type ([av-ˈav.di]) is marginally higher than that of the Complex type ([stæb-ˈstæb.gə]; $p = 0.02 < 0.05$), and significantly higher than that in other testing types (all $p < 0.01$). This is phonologically grounded as avoidance to produce two consecutive identical vowels. The fact that Complex seems to attract VC rimes is consistent with the statistical trend in the English lexicon reported in Kelly (2004): a complex onset might also contribute to the syllable weight and makes the syllable heavier, which led to the tendency of incorporating the coda.

2.4.3.3 Interim summary

Comparing the results here with the predictions in Table 2., we can conclude that participants' responses should be best characterized by copying a heavy syllable ($\sigma_{\mu\mu}$). Base-dependent syllable copying (σ_1) only receives marginal support, consistent with the typological generalization that base-dependent syllable copying patterns are extremely rare. The other plausible generalizations received no support. Participants were as willing to copy unseen labial onsets ([p, b, f, v]) as the familiarized coronal onsets ([t, d, ʃ, z]), to copy unseen coronal codas as the familiarized labial codas, and to copy high vowels that were not copied in the familiarization phase. For the Onsetless type, the XXX hypothesis (i.e., counting three segments) would predict [avd] for [avdi] – this almost never occurred, despite [avdavdi] being well-formed and easy to produce. For the **Complex** type, the predicted forms by the CVC or more fine-grained feature-based template, such as [sæb] or [tæb] for [ˈstæb.gə], rarely happened.

Together, these findings support the hypothesis that human learners are highly sensitive to surface repetitions. Given only eight familiarized items, they rapidly extracted copying-based generalizations and extended them to novel singulars of unseen feature combinations and of unseen shapes. Among all generalizations that were compatible with the data, instead of adhering to more fine-grained phonological specifications, participants exhibited a bias towards higher-level abstractions, namely prosodic units, supporting core claims in McCarthy and Prince (1986, *et seq.*).

2.4.4 Experiment 1b: Vowel reduced to [ə]

2.4.4.1 Methods

24 self-reported English native speakers who did not participate in the previous experiment were recruited through Amazon Mechanical Turk (11 males, 12 females, 1 other; mean age = 39.5, age range = 26 - 56). Six reported exposure to other languages, including Spanish, Portuguese, Italian, French, German, and Japanese, none of which has productive partial reduplication. All training and testing singulars were the same as Experiment 1a, except that the trained reduplicated forms all had vowels reduced to [ə] (see Appendix B.1). Participants followed the same procedure and received $3.50 for completing the experiment.

2.4.4.2 Analyses and results

Repetition accuracy of the trained items Similar to the previous experiment, participants in Experiment 1b showed high repetition accuracy. The repeated responses of the plural forms greatly followed the target reduplicative rule. We replaced the vowels in the repeated bases with the fixed vowel [ə]. Then, we evaluated this by using these newly created forms as the expected affix. The average segmental similarity between the affixes and the expected affixes exceeded 0.95 as well, indicating that participants' repetitions followed the familiarized rule.

Are the generalizations copying-based? Figure 2.10 presents the Monte Carlo simulation of responses in Experiment 1b. The majority of the participants had their observed affix-base similarity exceed the 99% confidence interval of chance. The result of one participant appeared to occur by chance. Further, we examined the average proportions of the base segment realizations in the affix, as illustrated in Figure 2.11 for onsets, Figure 2.12 for nuclei and Figure 2.13 for codas. Notably, for the majority of the onset segments, the most frequent realization was a repetition of themselves, including unseen labial onsets ([p, b, f, v]; ∼ 85%), varying shapes such as onset clusters ([dɹ, ʃɹ, sl, st]; ∼ 60%) and no onset (∅; ∼ 74%). For the vowel slot, [ə] was used for the most of the time, hence indeed fixed (∼ 63%). Active coda copying was observed to occur at a certain rate (∼ 33%), as unseen coronal codas ([t, d, ʃ, z]) were repeated (∼ 29%). These results support the application of generalizations based on more abstract phonological units (a heavy syllable), with a fixed [ə] (∼ 63%).

However, it appeared that generalizations based on allomorphy or fixed shapes computed based on syllabicity were also adopted to some extent. For example, a [d] was sometimes inserted as an onset when the base was onsetless ([dəv-ˈɛvgə]; ∼ 13%). As for the coda, in addition to active copying, across-the-board coda dropping was observed ([tə-ˈtæf.ku]; ∼ 31%). [v] was used as the coda across different base codas ([ʃəv-ˈʃip.neɪ]; ∼ 18%), and a fixed coda [b] at a similar rate ([ʃəb-ˈʃɑf.tu]; ∼ 18%). Beyond the presence of fixed segmental materials, we identified systematic simplifications of onset clusters (i.e. deleting a consonant when there are multiple consonants in the onset). For example, the base onset [dɹ] was frequently realized as [d] ([dəv-ˈdɹav.boʊ]; ∼ 28%). Similarly, the onset [ʃɹ] was frequently realized as [ʃ] ([ʃəf-ˈʃɹaf.pu]; ∼ 48%) and as [s] ([səf-ˈʃɹaf.pu]; ∼ 15%), and both [sl] and [st] were reduced to [s] ([səf-ˈslɛf.tu] and [səb-ˈstæb.nə]; ∼ 20%). It is worth noting that these onset clusters were never simplified to the sonorant: the base onset [dɹ] and [ʃɹ] were never

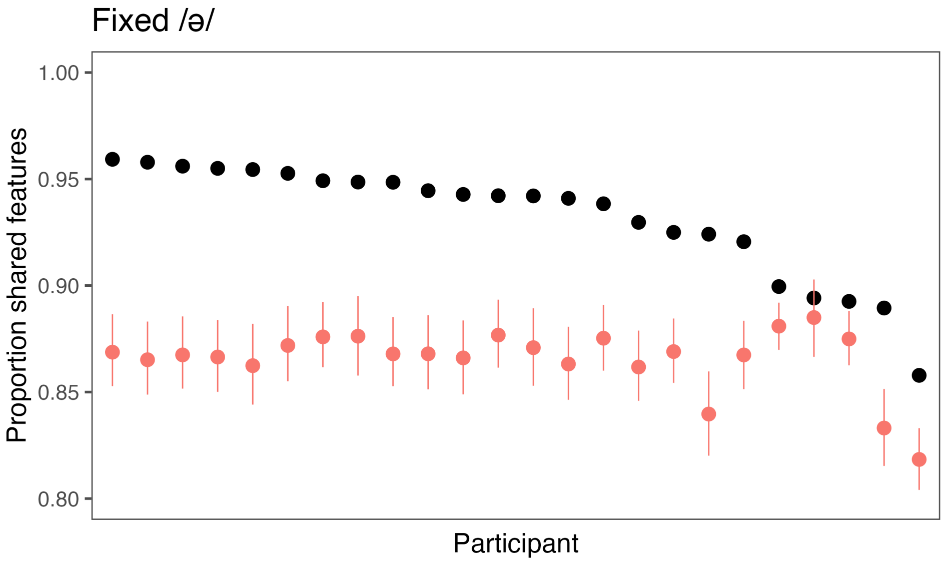
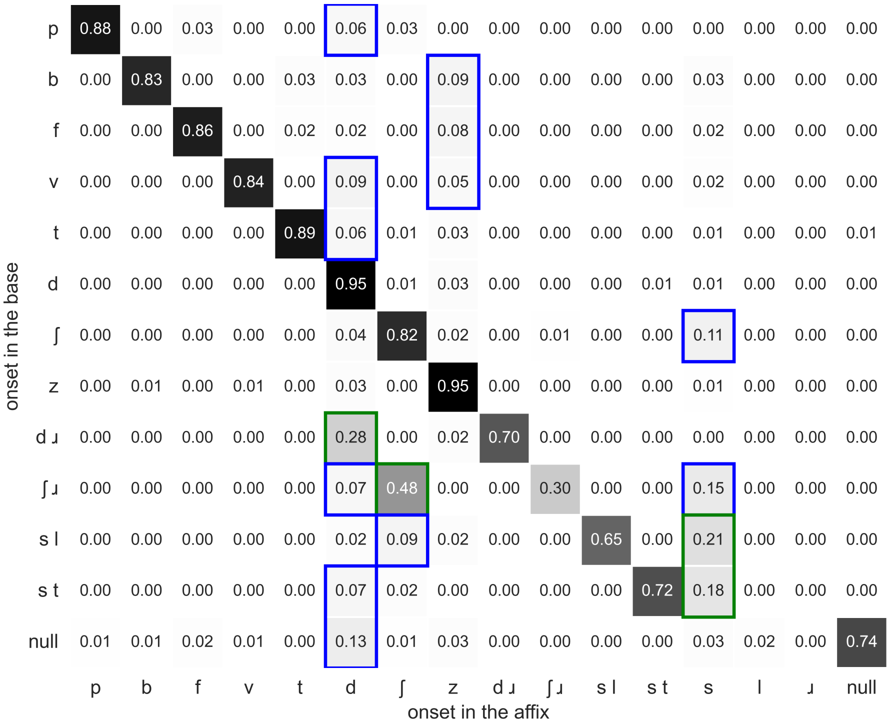


Figure 2.10: *The average segment similarity between the affixes and the novel singular bases. Black dots: the observed responses from each participant. Red dots: the mean calculated in the Monte Carlo procedure (R = 10000) with bars indicating the 99% confidence interval of chance.*



	p	b	f	v	t	d	ʃ	z	d ɹ	ʃ ɹ	s l	s t	s	l	ɹ	null
p	0.88	0.00	0.03	0.00	0.00	0.06	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
b	0.00	0.83	0.00	0.00	0.03	0.03	0.00	0.09	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00
f	0.00	0.00	0.86	0.00	0.02	0.02	0.00	0.08	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
v	0.00	0.00	0.00	0.84	0.00	0.09	0.00	0.05	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
t	0.00	0.00	0.00	0.00	0.89	0.06	0.01	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01
d	0.00	0.00	0.00	0.00	0.00	0.95	0.01	0.03	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
ʃ	0.00	0.00	0.00	0.00	0.00	0.04	0.82	0.02	0.00	0.01	0.00	0.00	0.11	0.00	0.00	0.00
z	0.00	0.01	0.00	0.01	0.00	0.03	0.00	0.95	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
d ɹ	0.00	0.00	0.00	0.00	0.00	0.28	0.00	0.02	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ʃ ɹ	0.00	0.00	0.00	0.00	0.00	0.07	0.48	0.00	0.00	0.30	0.00	0.00	0.15	0.00	0.00	0.00
s l	0.00	0.00	0.00	0.00	0.00	0.02	0.09	0.02	0.00	0.00	0.65	0.00	0.21	0.00	0.00	0.00
s t	0.00	0.00	0.00	0.00	0.00	0.07	0.02	0.00	0.00	0.00	0.00	0.72	0.18	0.00	0.00	0.00
null	0.01	0.01	0.02	0.01	0.00	0.13	0.01	0.03	0.00	0.00	0.00	0.00	0.03	0.02	0.00	0.74

Figure 2.11: *The averaged proportion of the affix onsets conditioned on the base onsets. Average faithful realizations (diagonal): 0.78*

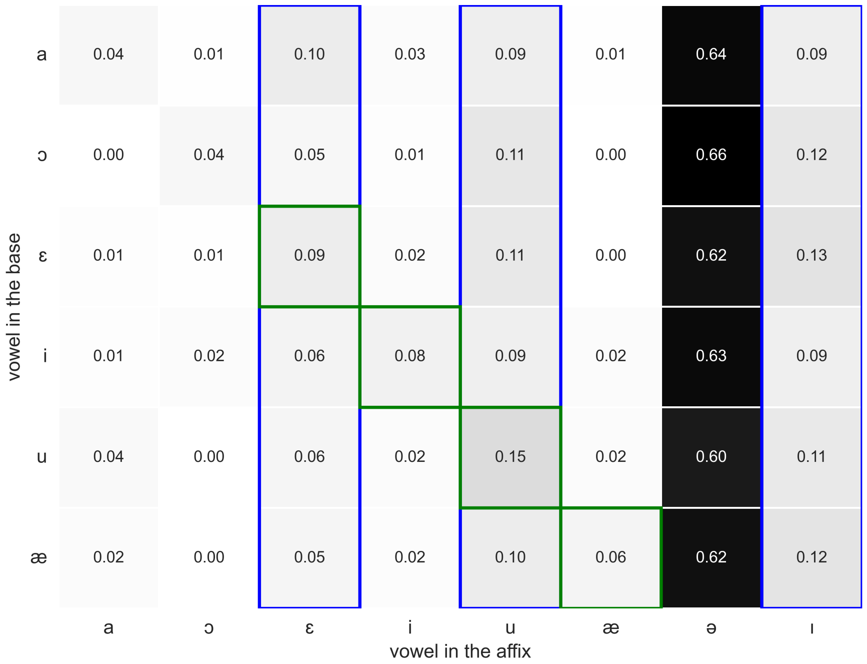


Figure 2.12: *The averaged proportion of the affix nuclei conditioned on the base nuclei. The average of the fixed ə (the column ə): 0.63*

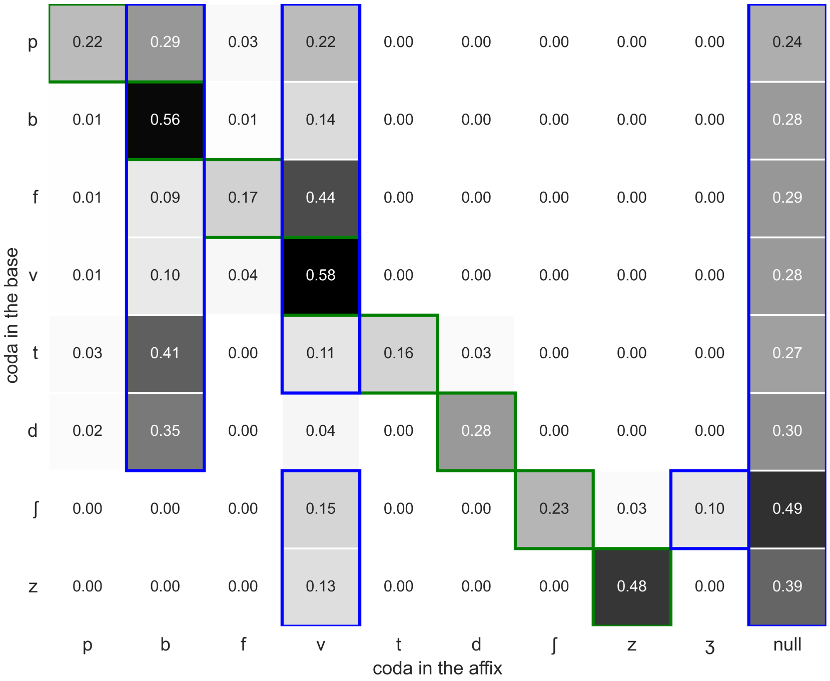


Figure 2.13: *The averaged proportion of the affix codas conditioned on the base codas. Averaged faithful realization (diagonal): 0.33; averaged coda-less realization (the column "null"): 0.31; averaged fixed [v]/[b] (the column [v]/[b] excluding the base [v]/[b]): 0.36.*

realized as [ɹ], and [sl] was never realized as [l]. One explanation for such an asymmetry is that onset simplification in reduplication is grounded in phonological similarity, as suggested by Fleischhacker (2005). *Obstruent + sonorant* clusters such as [dɹ] are more phonologically similar to an obstruent [d] than to a sonorant [ɹ], hence preserving phonological similarities. It is also possible to attribute this asymmetry to the input: the obstruents were copied in the familiarized items, but the sonorants did not appear in the familiarization, though [s] would provide counter evidence since [s] never appeared in the familiarization phase.

Generalizing at what level of abstraction? As mentioned in the Monte Carlo analysis of Experiment 1b, there was one participant whose responses were not consistent with a copying-based generalization. Therefore, we excluded this participant from subsequent analyses on phonological shapes.

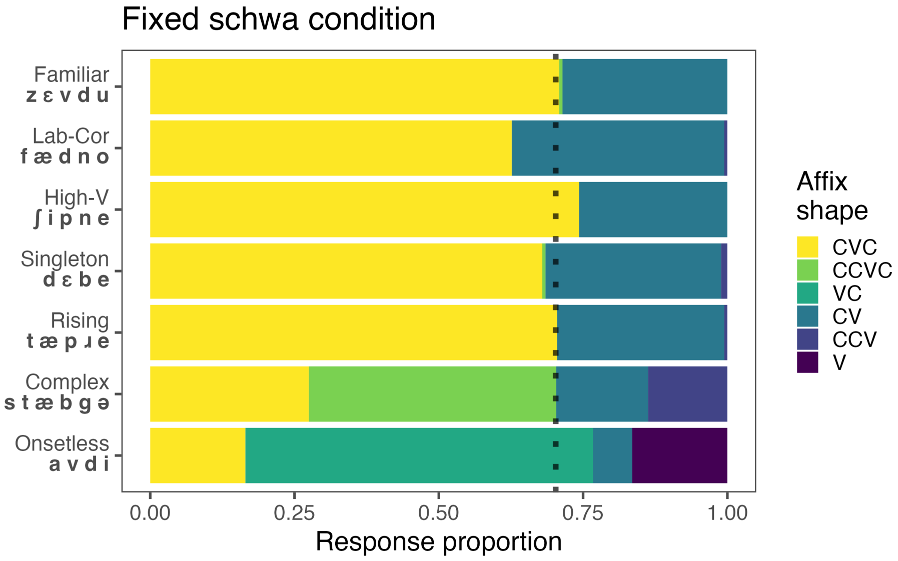


Figure 2.14: *The averaged proportion of affix shape conditioned on each testing type; vertical line: the baseline rate of coda incorporation based on the* Familiar *type, namely when the target syllable has a coronal onset, non-high vowel and a labial coda.*

Figure 2.14 presents the average proportion of affix shapes, conditioned on the testing type for the remaining twenty participants. For the novel singulars of the Familiar type, CVC remained the most frequent affix shape, although CV shape also appeared. We noted a comparable rate of coda incorporation across the other six testing types. However, speaking of onset manipulations, in both the Complex testing type and Onsetless testing type, a significant proportion of responses included a simple onset only with one consonant – this is consistent with the discussions on segmental realizations in the previous section.

One puzzling aspect of the result is the proportion of V in the Onsetless case, which seems to create phonologically ill-formed surface forms, namely having two consecutive vowels (VV/əV).[8] One explanation is that participants have adopted the base-dependent onset copying and maintained a general dispreference of incorporating the coda, either due to an aversion to having a coda after the [ə] or due to a general inclination to perform phonological reduction.

We performed the same Bayesian mixed-effect multinomial logistic regression modeling in rStan, for the onset shape and the rime shape respectively. Figure 2.15 shows the sampled posterior probabilities conditioned on each testing type and the mean of the sampled posterior probabilities are given in Table 2.7. Following the same procedure described in Experiment 1a, we performed pairwise comparisons within each testing types (in Table 2.11) and between testing types (in Table 2.8).

Testing type t		$\overline{p(C\|t)}$	$\overline{p(CC\|t)}$	$\overline{p(\varnothing\|t)}$	$\overline{p(VC\|t)}$	$\overline{p(V\|t)}$
Familiar	ˈzɛv.du	$\boxed{\mathbf{0.99}}$	0.00	0.01	$\boxed{\mathbf{0.80}}$	**0.20**
Lab-Cor	ˈfæd.noʊ	$\boxed{\mathbf{0.98}}$	0.00	0.01	$\boxed{\mathbf{0.67}}$	**0.33**
High-V	ˈʃip.neɪ	$\boxed{\mathbf{0.99}}$	0.00	0.01	$\boxed{\mathbf{0.83}}$	**0.17**
Singleton	ˈdɛ.beɪ	$\boxed{\mathbf{0.99}}$	0.00	0.01	$\boxed{\mathbf{0.74}}$	**0.26**
Rising	ˈtæ.pɹeɪ	$\boxed{\mathbf{0.99}}$	0.00	0.01	$\boxed{\mathbf{0.80}}$	**0.20**
Complex	ˈstæb.gə	**0.42**	$\boxed{\mathbf{0.57}}$	0.01	$\boxed{\mathbf{0.76}}$	**0.24**
Onsetless	ˈav.di	**0.12**	0.01	$\boxed{\mathbf{0.87}}$	$\boxed{\mathbf{0.87}}$	**0.13**

Table 2.7: *The mean of the sampled posterior probabilities calculated based on the population-level parameter estimates*

The onset model establishes that participants predominantly performed base-dependent onset realizations with a fair amount of onset cluster simplifications, consistent with our previous discussions. For the five testing types with one-consonant base onsets (Familiar,

[8] English phonotactics allows VV sequences but not əV.

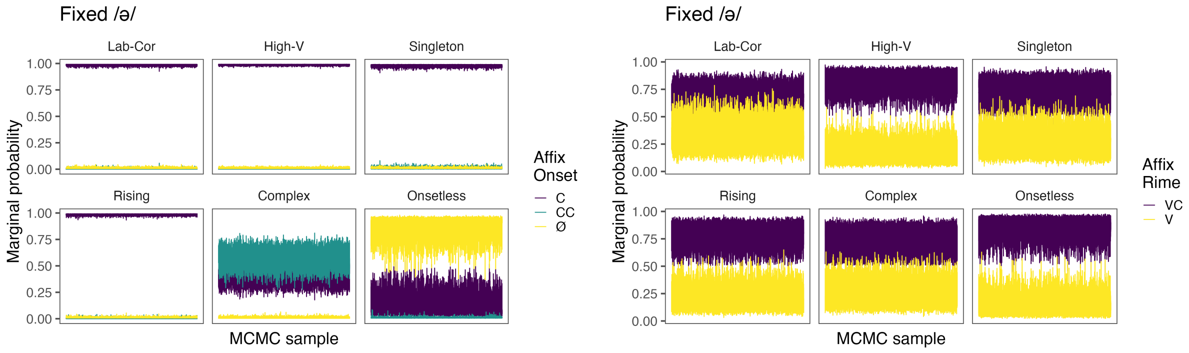


Figure 2.15: *The posterior probabilities of the produced shapes given each testing type.*

Onset: C	All testing types with a C as the onset > Complex > Onsetless	Predominantly base-dependent onset shapes with complex onset simplification
Onset:CC	Complex > all other types	
Onset: Ø	Onsetless > all other types	
Rime: VC		
Rime: V		

Table 2.8: *A summary of pairwise comparisons based on the sampled posterior probabilities between testing types in Experiment 1b. " > " indicates that the probability of showing a level is significantly higher than the other, with a threshold of $p < 0.01$.*

Lab-Cor, High-V, Singleton, Rising), the probability of a C onset was significantly higher than producing CC or no onset (all $p < 0.01$). For the Onsetless testing type, the probability of having an onsetless response ([əv-ˈav.di]) is significantly higher than producing a C ([dəv-ˈav.di]; $p < 0.001$). For the Complex testing type, the probability of producing CC ([stəb-ˈstæb.gə]) is not significantly higher than the probability of producing a C ([səb-ˈstæb.gə]; $p = 0.15$).

Between testing types, compared to the Complex type and the Onsetless type, the probability of having a C onset is significantly higher in the five testing types with a C onset (all $p < 0.01$). The probability of a CC onset is significantly higher in the Complex type ([stəb-ˈstæb.gə]) than in all other testing types ($p < 0.01$). The probability of producing onsetless responses is significantly in the Onsetless type ([əv-ˈav.di]) higher than in all other testing types ($p < 0.01$). Interestingly, the probability of producing a C for the onset is significantly higher in Complex type than in the Onsetless type ($p = 0.007 < 0.01$), which confirms that the urge to simplify the onset could not be reduced to a propensity to produce only one onset across the board.

The rime model confirmed a relatively base-independent preference for coda incorporation. Within testing types, FAMILIAR, HIGH-V, RISING, COMPLEX, ONSETLESS types had the probability of a VC rime ([fəd-ˈfæd.noʊ]) significantly higher than the probability of a V rime ([fə-ˈfæd.noʊ]; $p < 0.01$). For the SINGLETON case, the probability of a VC rime ([dəb-ˈdɛ.beɪ]) is only marginally higher than the probability of a V rime ([də-ˈdɛ.beɪ]; $p = 0.013 < 0.05$). As for the LAB-COR testing type, the probability of producing a VC ([fəd-ˈfæd.noʊ]) is not significantly higher than the probability of producing a V rime ([fə-ˈfæd.noʊ]; $p = 0.07$). For each level, there were no significant differences between testing types.

2.4.4.3 Interim summary

In summary, the results of Experiment 1b demonstrate that most participants were able to extrapolate the familiarized reduplicative rule and apply it to novel stems: they still showed predominantly base-dependent onset copying, with some amount of coda copying, which is compatible with a heavy-syllable-based copying generalization. Moreover, participants successfully extended the fixed [ə] to the high vowels that they had never seen in the familiarization. These findings are consistent with evidence from Experiment 1, indicating that humans are sensitive to identity-based dependencies.

However, the fixed [ə] in this experiment resulted in more variable responses of onsets and codas. In particular, a substantial portion of responses exhibited more phonological reductions. First, when there was more than one consonant present in the onset position, they were simplified to a single consonant. Secondly, in most testing types, the coda was not as strongly incorporated as in Experiment 1a. Thirdly, the affix codas were sometimes fixed. These findings were not in Experiment 1a, which suggests that although participants were able to perform high-level abstractions, to a certain degree, they also conducted computation based on finer-grained phonological materials (e.g., fixed segmental properties in codas and onsets; simplified the onset cluster) when a fixed [ə] was involved.

2.4.5 Experiment 1c: Copy vowel overwritten to [i]

2.4.5.1 Methods

25 self-reported English native speakers who did not participate in the previous experiments were recruited from Amazon Mechanical Turk (13 females, 11 males; mean age = 41.32, age range = 24 - 73). Three reported exposure to other languages, including Spanish and French, none of which have productive reduplication. All training and testing singulars were the same as Experiment 1a and Experiment 1b, except that the reduplicated forms all had vowels rewritten to [i] (see Appendix B.1). Participants followed the same procedures as the previous experiment and received $3.50 for completing the experiment.

2.4.5.2 Analyses and results

Repetition accuracy of the familiarized items Consistent with the previous two experiments, the repetition accuracy of the trained items in this experiment was also high. The repeated responses of the plural forms largely followed the intended familiarized rule. Same as Experiment 1b, we replaced the vowels in the produced bases with the fixed vowel [i] and used these newly created forms as the expected affixes. The average segmental similarity between the actual affix and the expected affix exceeded 0.95 as well, indicating that participants' repetitions followed the familiarized rule.

Are the generalizations copying-based? For Experiment 1c, the Monte Carlo analysis indicates that five participants had their observed affix-base similarity fall into the 99% confidence interval of chance, as in Figure 2.16. We could not confidently conclude that these participants had adopted a copying-based generalization.

Together, Figure 2.17, Figure 2.18 and Figure 2.19 show the average realizations of each base segment in the affix, reflecting similar patterns as in Experiment 1b. First, let us examine the extent of copying: participants were willing to copy unseen labial onsets ([p, b, f, v]; ∼ 54%), the unseen complex onsets ([dɹ, ʃɹ, sl, st]; ∼ 46%) and keep onset absent

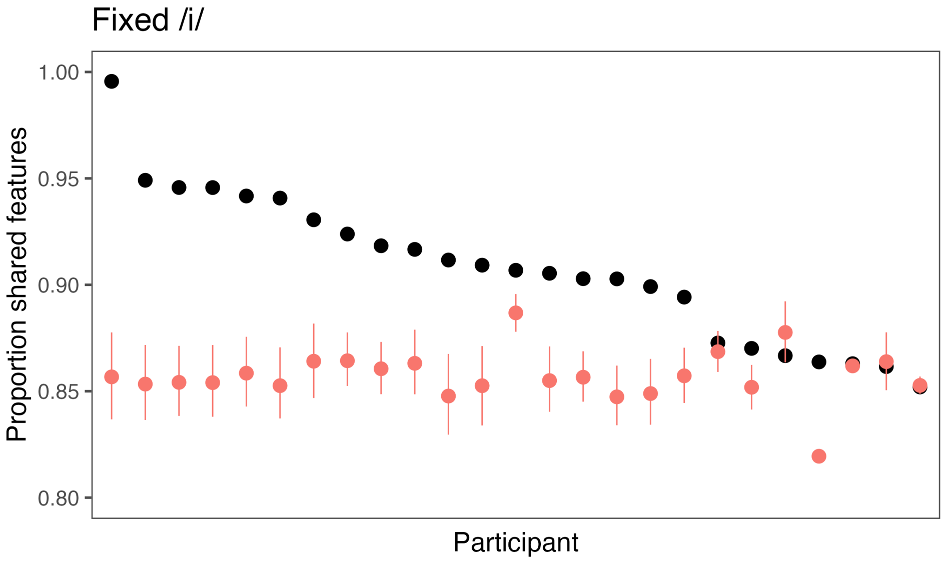
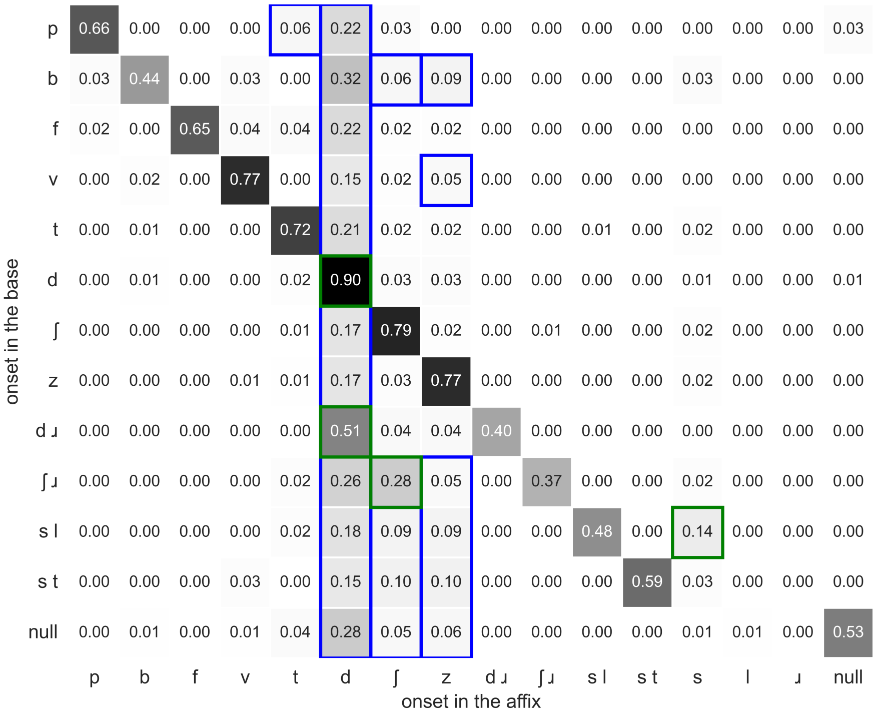
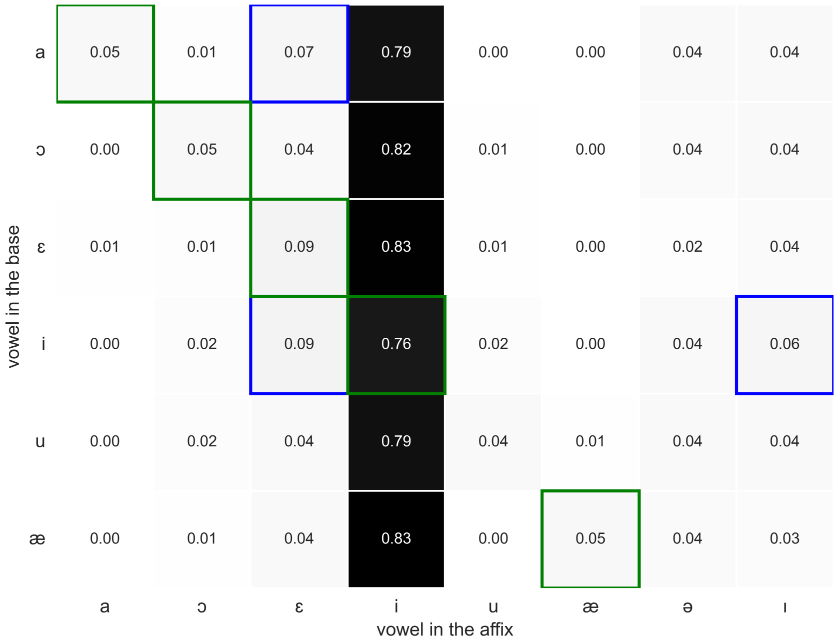


Figure 2.16: *The average segment similarity between the affixes and the novel singular bases. Black dots: the observed responses from each participant. Red dots: the mean calculated in the Monte Carlo procedure (R = 10000) with bars indicating the 99% confidence interval of chance.*



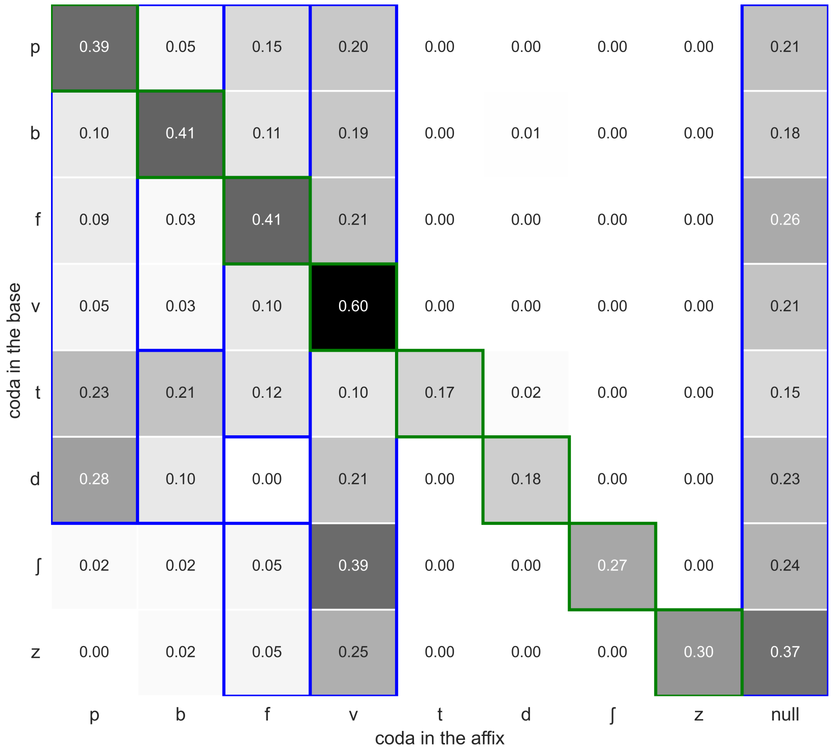
	p	b	f	v	t	d	ʃ	z	d ɹ	ʃ ɹ	s l	s t	s	l	ɹ	null
p	0.66	0.00	0.00	0.00	0.06	0.22	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
b	0.03	0.44	0.00	0.03	0.00	0.32	0.06	0.09	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00
f	0.02	0.00	0.65	0.04	0.04	0.22	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
v	0.00	0.02	0.00	0.77	0.00	0.15	0.02	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
t	0.00	0.01	0.00	0.00	0.72	0.21	0.02	0.02	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00
d	0.00	0.01	0.00	0.00	0.02	0.90	0.03	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01
ʃ	0.00	0.00	0.00	0.00	0.01	0.17	0.79	0.02	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.00
z	0.00	0.00	0.00	0.01	0.01	0.17	0.03	0.77	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
d ɹ	0.00	0.00	0.00	0.00	0.00	0.51	0.04	0.04	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ʃ ɹ	0.00	0.00	0.00	0.00	0.02	0.26	0.28	0.05	0.00	0.37	0.00	0.00	0.02	0.00	0.00	0.00
s l	0.00	0.00	0.00	0.00	0.02	0.18	0.09	0.09	0.00	0.00	0.48	0.00	0.14	0.00	0.00	0.00
s t	0.00	0.00	0.00	0.03	0.00	0.15	0.10	0.10	0.00	0.00	0.00	0.59	0.03	0.00	0.00	0.00
null	0.00	0.01	0.00	0.01	0.04	0.28	0.05	0.06	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.53

Figure 2.17: *The averaged proportion of the affix onsets conditioned on the base onsets. Average faithful realizations (diagonal): 0.62*



	a	ɔ	ɛ	i	u	æ	ə	ɪ
a	0.05	0.01	0.07	0.79	0.00	0.00	0.04	0.04
ɔ	0.00	0.05	0.04	0.82	0.01	0.00	0.04	0.04
ɛ	0.01	0.01	0.09	0.83	0.01	0.00	0.02	0.04
i	0.00	0.02	0.09	0.76	0.02	0.00	0.04	0.06
u	0.00	0.02	0.04	0.79	0.04	0.01	0.04	0.04
æ	0.00	0.01	0.04	0.83	0.00	0.05	0.04	0.03

Figure 2.18: *The averaged proportion of the affix nuclei conditioned on the base nuclei. The average of the fixed i (the column i): 0.80*



	p	b	f	v	t	d	ʃ	z	null
p	0.39	0.05	0.15	0.20	0.00	0.00	0.00	0.00	0.21
b	0.10	0.41	0.11	0.19	0.00	0.01	0.00	0.00	0.18
f	0.09	0.03	0.41	0.21	0.00	0.00	0.00	0.00	0.26
v	0.05	0.03	0.10	0.60	0.00	0.00	0.00	0.00	0.21
t	0.23	0.21	0.12	0.10	0.17	0.02	0.00	0.00	0.15
d	0.28	0.10	0.00	0.21	0.00	0.18	0.00	0.00	0.23
ʃ	0.02	0.02	0.05	0.39	0.00	0.00	0.27	0.00	0.24
z	0.00	0.02	0.05	0.25	0.00	0.00	0.00	0.30	0.37

Figure 2.19: *The averaged proportion of the affix codas conditioned on the base codas. Averaged faithful realization (diagonal): 0.33; averaged coda-less realization (the column "null"): 0.23*

in onsetless case (∅; ∼ 53%). A notable amount of coda copying was observed for coronal coda ([t, d, ʃ, z]; ∼ 23%). However, it is clear that more fixed segments were involved: [d] was frequently used as a fixed onset across various base onsets ([div-ˈɛvɡə]; ∼ 24%). The complex onsets also showed some amount of simplification, with the same asymmetry between obstruents and sonorants in Experiment 2. As for the vowels, [i] was consistently fixed across different base vowels (∼ 80%). For the coda, [v] was systematically used as the fixed coda ([ʃiv-ˈʃip.neɪ]; ∼ 22%), with some instances of fixed [f, p, b]. Remarkably, [p, b, f, v] were the copied codas in the familiarized items.

On another note related to the fixed vowel, subjects extended [i] as the fixed segment to high vowels in the bases, namely [i, u]. This creates perfect repetitions [ʃip.ʃip.ne͡ɪ], which never occurred in the familiarized items. This result does not support a higher-order vowel identity avoidance. Rather, it suggests that participants largely adopted a first-order generalization that involves fixed vowels.

Generalizing at what level of abstraction? As mentioned earlier, we could not confidently conclude that the responses of five participants exhibited copying-based generalizations. Hence, these participants were excluded from the following analyses. Figure 2.20 shows the average proportion of affix shapes, conditioned on the testing type, based on the nineteen remaining participants.

Results from Experiment 1c echoed with results from Experiment 1b. The novel singulars of the FAMILIAR type had CVC ([ziv-ˈzɛv.du]) as the more frequent shape compared to the CV realizations ([zi-ˈzɛv.du]). We observed a similar rate of coda incorporation across the other six testing types. However, in terms of the onset manipulations, it is clear that a noticeable portion of responses was of the CVC shape for both the COMPLEX type ([sib-ˈstæb.ɡə]) and the ONSETLESS testing type ([dib-ˈdɛ.beɪ]).

Bayesian mixed-effect multinomial logistic regression models were performed, on the onset shape and the rime shape respectively. Figure 2.21 shows the sampled posterior probabilities conditioned on each testing type, while Table 2.10 provides the mean of the sampled posterior probabilities. As before, we performed pairwise comparisons based on the sampled posterior

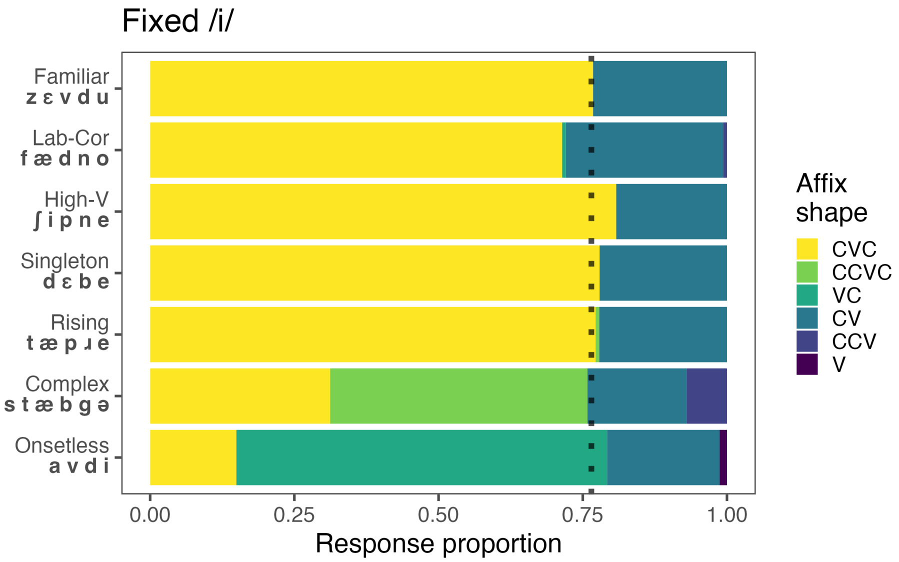
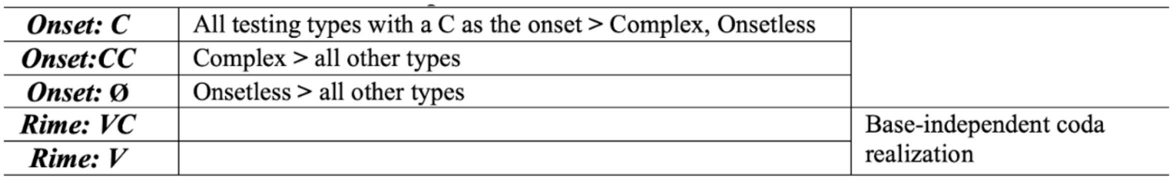


Figure 2.20: *The averaged proportion of affix shape conditioned on each testing type; vertical line: the baseline rate of coda incorporation based on the* Familiar *type, namely when the target syllable has a coronal onset, non-high vowel and a labial coda.*



Onset: C	All testing types with a C as the onset > Complex, Onsetless	
Onset:CC	Complex > all other types	
Onset: Ø	Onsetless > all other types	
Rime: VC		Base-independent coda realization
Rime: V		

Table 2.9: *A summary of pairwise comparisons based on the sampled posterior probabilities between testing types in Experiment 1c. "$>$" indicates that the probability of showing a level is significantly higher than the other, with a threshold of p $<$ 0.01.*

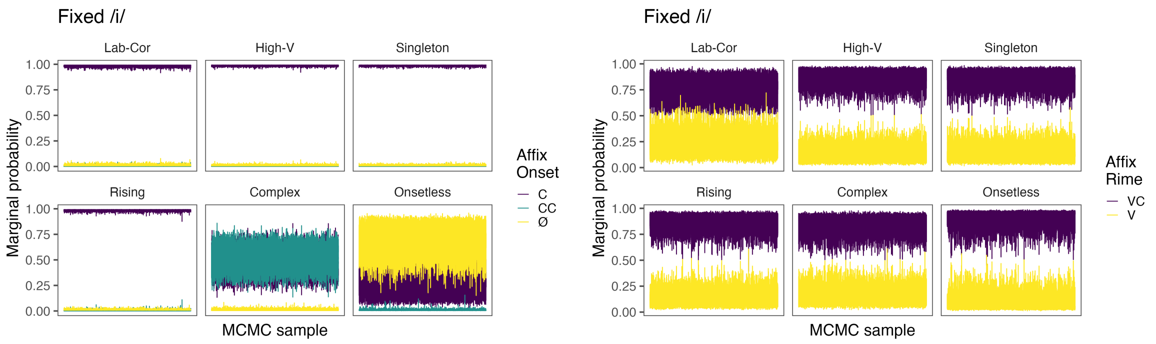


Figure 2.21: *The posterior probabilities of the produced shapes given each testing type.*

Testing type t		$\overline{p(C\|t)}$	$\overline{p(CC\|t)}$	$\overline{p(\varnothing\|t)}$	$\overline{p(VC\|t)}$	$\overline{p(V\|t)}$
Familiar	ˈzɛv.du	$\boxed{\textbf{0.99}}$	0.00	0.01	$\boxed{\textbf{0.87}}$	**0.13**
Lab-Cor	ˈfæd.noʊ	$\boxed{\textbf{0.98}}$	0.00	0.01	$\boxed{\textbf{0.79}}$	**0.21**
High-V	ˈʃip.neɪ	$\boxed{\textbf{0.99}}$	0.00	0.01	$\boxed{\textbf{0.88}}$	**0.12**
Singleton	ˈdɛ.beɪ	$\boxed{\textbf{0.99}}$	0.00	0.01	$\boxed{\textbf{0.87}}$	**0.13**
Rising	ˈtæ.pɹeɪ	$\boxed{\textbf{0.99}}$	0.00	0.01	$\boxed{\textbf{0.87}}$	**0.13**
Complex	ˈstæb.gə	**0.47**	$\boxed{\textbf{0.52}}$	0.01	$\boxed{\textbf{0.85}}$	**0.15**
Onsetless	ˈav.di	**0.27**	0.01	$\boxed{\textbf{0.72}}$	$\boxed{\textbf{0.91}}$	**0.09**

Table 2.10: *The mean of the sampled posterior probabilities calculated based on the population-level parameter estimates*

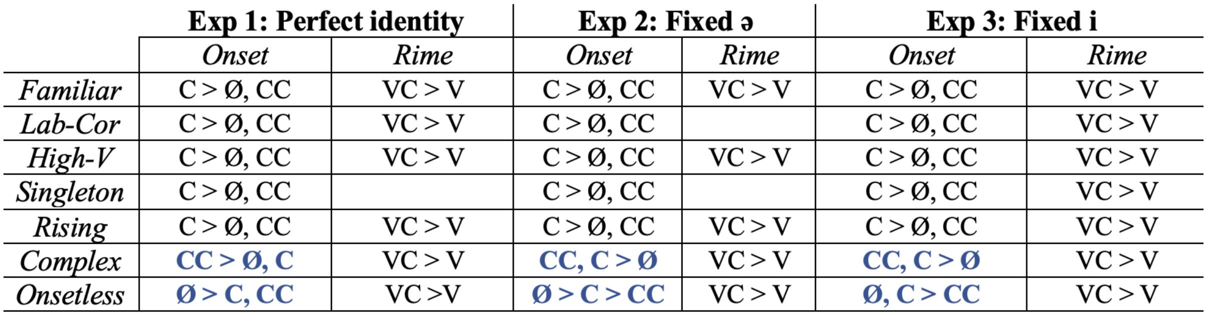
probabilities within each testing type (Table 2.11) and between testing types (Table 2.9).

The analysis of the rime suggests that coda realizations were base-independent. For all testing types, the probability of a VC rime is significantly higher than the probability of a V rime (all $p < 0.01$). For each level, there is no significant difference between testing types.

On the other hand, the onset realization showed mixed results. Within the five testing types with a C as the onset (Familiar, Lab-Cor, High-V, Singleton, Rising), the probability of a C onset is significantly higher than a CC onset and no onset (all $p < 0.01$). Interestingly, in the Complex testing type, the probability of producing CC ([stib-ˈstæb.gə]) is not significantly higher than the probability of a C onset ([sib-ˈstæb.gə]; $p = 0.41$). For the Onsetless testing type, the probability of no onset ([iv-ˈav.di]) is only marginally higher than a C onset ([div-ˈav.di]; $p = 0.057$). Crucially, between testing types, the probability of producing a C is not significantly higher for the Complex type ([sib-ˈstæb.gə]) than for the Onsetless type ([div-ˈav.di]; $p = 0.10$) – this is different from Experiment 1b. We discuss the differences in more detail in Section 2.4.6.

2.4.5.3 Interim summary

In sum, participants showed mixed results on how much to copy and what to copy. On one hand, subjects were able to apply the familiarized rule to novel singulars, especially to those with varying shapes. Hence, convergent with Experiment 1a and 1b, humans are sensitive to



	Exp 1: Perfect identity		Exp 2: Fixed ə		Exp 3: Fixed i	
	Onset	*Rime*	*Onset*	*Rime*	*Onset*	*Rime*
Familiar	C > Ø, CC	VC > V	C > Ø, CC	VC > V	C > Ø, CC	VC > V
Lab-Cor	C > Ø, CC	VC > V	C > Ø, CC		C > Ø, CC	VC > V
High-V	C > Ø, CC	VC > V	C > Ø, CC	VC > V	C > Ø, CC	VC > V
Singleton	C > Ø, CC		C > Ø, CC		C > Ø, CC	VC > V
Rising	C > Ø, CC	VC > V	C > Ø, CC	VC > V	C > Ø, CC	VC > V
Complex	**CC > Ø, C**	VC > V	**CC, C > Ø**	VC > V	**CC, C > Ø**	VC > V
Onsetless	**Ø > C, CC**	VC >V	**Ø > C > CC**	VC > V	**Ø, C > CC**	VC > V

Table 2.11: *A summary of pairwise comparisons based on the sampled posterior probabilities within each testing type. " > " indicates that the probability of showing a level is significantly higher than the other, with a threshold of* $p < 0.01$*.*

identity-based relations. On the other hand, both the Onsetless and the Complex testing type showed significant proportions of simple onsets. These two testing types did not differ significantly in terms of the probability of having a simple onset. Additionally, there was a significant proportion of fixing a coda, on par with active coda copying. Convergent with Experiment 1b, these results on the fixed segments and fixed phonological shapes based on syllabicity (i.e. CVC) diverged from Experiment 1a. Hence, we can conclude that while participants were capable of extrapolating based on more coarse-grained phonological abstractions, they also showed the tendency to fix finer-grained phonological details when the fixed segments are involved.

Regarding the fixed vowels, participants not only recognized that [i] was fixed but also extended this fixed [i] to unseen high base vowels [i] and [u], suggesting a bias towards first-order generalizations of the fixed segment, but not higher-order identity avoidance.

2.4.6 Findings of Experiment Series 1

Rapid generalization of reduplicative patterns with phonological abstractions

In the first series of artificial grammar learning experiments, we studied how speakers learn partial reduplication and focused on the structures inside a syllable. The experimental results presented here are consistent with the previous findings suggesting that humans are highly sensitive to surface repetitions of sound sequences (e.g., Marcus et al., 1999;

Berent et al., 2016, 2017). We add to the previous literature by investigating cases where surface repetitions are results of morphophonological processes, namely, reduplication. We found that participants rapidly extracted copying-based generalizations and extended them to novel forms, suggesting that reduplicative structures are easy to learn as word-formation processes.

We found that participants were able to systematically recognize the effects of copying (i.e. identity-based segmental dependencies), not only when the affixal part only involves copying, but also when there was a mixture of active copying and fixed phonological materials. Moreover, they rapidly generalized reduplicative rules and extended them to novel singulars. The novel singulars tested here involved novel feature combinations as well as novel phonological shapes. When familiarized with perfect phonological identity, participants predominantly adopted the generalizations based on coarse-grained prosodic constituents. Base-dependent syllable copying received marginal support. This is consistent with the previous finding that typologically rare syllable copying is difficult to learn (Haugen et al., 2022). In general, we identified an inductive bias towards more coarse-grained phonological abstractions based on prosodic constituents, supporting McCarthy and Prince (1986).

When familiarized with fixed [i] and [ə] vowels, participants were able to extend the fixed vowels with consonant copying at a more abstract level, though at a lower rate. We also observed a bias towards first-order generalizations of the fixed segment, but not higher-order identity avoidance.

Learning outcomes with different familiarized patterns We indeed identified performance differences between the familiarized pattern showing perfect phonological identity (Expt. 1a) and the familiarized pattern with fixed vowels (Expt. 1b, Expt. 1c). In general, participants in Expt. 1b and Expt. 1c performed phonological computations with respect to the other positions, namely onsets and codas.

Participants in Expt. 1b were more prone to phonological reductions to produce less marked surface forms, including no incorporation of the coda and onset simplifications, the drive to which was irreducible to the need to keep one consonant across the board.

Participants in Expt. 1c produced more fixed materials: copied onsets in the familiarization phase occupied the onset (here [d]), copied codas in the familiarization phase were used as the fixed coda (here, [v]). A proportion of the responses kept the syllabicility-based shapes fixed: among three experiments, only Expt. 1c showed the VC rime occurring at a significantly higher rate within all testing types, and only Expt. 1c had the simple onset occurring at a significantly higher rate within all testing types. Moreover, there is no significant difference in producing a C onset between the Onsetless vowel and the Complex type – this is very different from the findings in Expt.1b. In Expt.1b, the probability of producing a C is higher in the Complex type than in the Onsetless case. The differences between these two experiments seem to suggest that the fixed vowels involved different encodings: the [ə] might often have been learned as the product of phonological reductions, and the [i] might be hard-wired as the fixed materials, providing some support for the predictions by Alderete et al. (1999).

In all, it seems that the fixed [ə] in Expt. 1b led to *more* phonological reduction, and the fixed [i] in Expt. 1c led to *more* detailedly specified fixed materials. These results in Expt. 1b and Expt. 1c point to one additional convergent conclusion to draw. That is, there might be tight bonds among the substructures within a copy. If a phonological computation targets a specific substructure, the effects might be spread to other substructures within the copy. If there were no such tight relations among substructures, the simple fact of having a fixed vowel affects the realizations of other positions, here the onset and the coda slots, would be left unexplained.

2.5 Experiment series 2: From monosyllabic copying to longer forms

2.5.1 Learning reduplication...

At what levels of abstraction beyond a syllable? From Experiment Series 1, we have established that human learners could rapidly extrapolate reduplicative generalizations at the level of the syllable. The eight pairs of familiarized items were limited in their segmental choices and their phonological shapes. However, participants were able to extend copying to novel feature combinations and novel shapes, suggesting that they were insensitive to the shared bottom-up phonological specifications unless the process itself involved some detailed specifications. Instead, participants had generalized in a manner that is sensitive to the coarse-grained prosodic units. In the first series of the experiments, we focused on the properties of prosodic units characterized by syllables and thus, varied the phonological shapes of the testing items at the syllable level. Based on many empirical investigations on word-level stress and tonal patterns, scholars argue that phonological groupings based on the prosodic units go beyond syllables, and form the so-called "Prosodic Hierarchy" (Selkirk, 1980a,b; McCarthy and Prince, 1986), as in (19).

(19) The Prosodic Hierarchy

PrWd (Prosodic word)
|
Ft (Foot)
|
σ (Syllable)
|
μ (Mora)

This naturally leads us to wonder about the status of those coarse-grained units beyond syllables in reduplication learning. If the training items are also compatible with hypotheses at the more abstract levels, at what levels of abstraction will human learners generalize? Do we still observe a systematic preference for syllable-based templates or do we also observe more abstract levels if at all possible?

Unbounded copying from bounded input? Looking at reduplication from a formal language theoretic view, the different levels of abstraction are said to bear qualitatively different formal implications. Starting from Johnson (1972), the module of phonology and the morphology-phonology interface is said to lie in the class of *regular* (i.e. computable by *finite-state* methods; Kaplan and Kay, 1994; Frank and Satta, 1998), or even less than the power of regular (Heinz, 2007, 2018; Chandlee, 2014, 2017), the details of which are more extensively discussed in Chapter 3. To motivate the studies presented below, let us first clarify what distinguishes the finite-state (and strictly less complex) type of computation and the supra-finite-state kind of computation.

One useful perspective to concretize this distinction is to examine the required computational resources for different input sizes, as illustrated in Figure 2.22. For a finite-state kind of computation, the required computational resources are always bounded by a constant. In contrast, the supra-finite-state computation requires the computational resources to grow together with the input size, for example, if the relation between the required computational resources and the input form a linear function, or as any monotonically increasing function (e.g., exponential function, monotonically increasing step functions with infinitely many intervals).

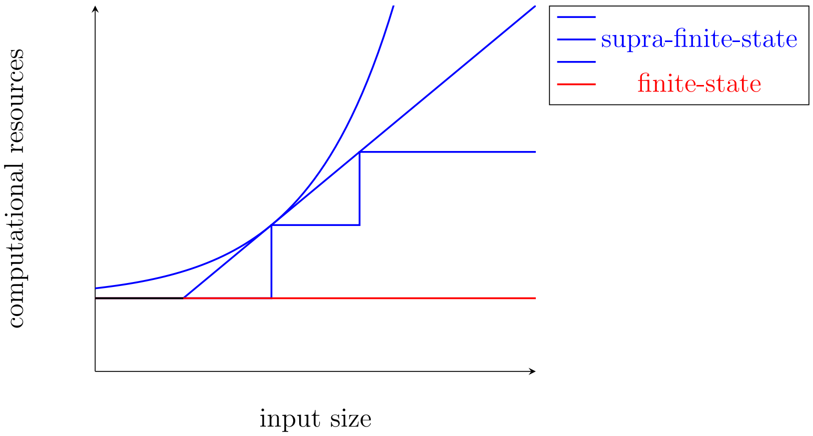


Figure 2.22: *A visualization of different kinds of computation.*

For reduplication, the computational resources naturally correspond to the realization of reduplicant (Chandlee and Heinz, 2012), and the input corresponds to the input stem/the base. Then, situating different reduplicative patterns into this picture, we know that total reduplication is a supra-finite-state computation because there is no upper bound on the reduplicant size, hence an instance of *unbounded copying*. On the other hand, most partial reduplication shows a rather fixed shape and hence is coupled with a possible upper bound on the reduplicant size. In this way, most partial reduplication involves a finite-state kind of computation. If the distinction between finite-state and supra-finite-state is the right sense to single out the space for (morpho)phonological patterns from other modules, say from the morphosyntactic dependencies, then reduplication patterns will inevitably be split into two kinds based on the shape of a copy. In fact, Heinz and Idsardi (2013, p.114) hypothesizes that total reduplication belongs to morphosyntax while partial reduplication belongs to morphophonology. We interpret this hypothesis as a conjecture on the nature of the unboundedness of the copying operation. If total reduplication belongs to morphosynax, the growth of a copy must rely on the unboundedness of possible morphosyntactic compositions (e.g., from monomorphemic to polymorphemic), but not on the compositions of phonological constituents (e.g., from monosyllabic to polysyllabic, while maintaining the same morphosyntactic structure). Granted that our take indeed captures the original hypothesis, here, we ask whether excluding unbounded copying, such as total reduplication, from the space of morphophonology is empirically grounded.

Thus, beyond the other research agenda mentioned at the beginning of this chapter, we hope to address the additional question in (20).

(20) *Is a learner's hypothesis space of morphophonology limited to finite-state kinds of computation?*

When learners try to learn a morphophonological process and are only provided with bounded input, will they interpret the limited familiarized items as an instance of bounded copying and hence adopt a size-restricting partial reduplication generalization, or unbounded copying and hence adopt a non-size-restricting generalization?

Note that the prosodic hierarchy and the formal perspective are not mutually exclusive. In our opinion, they are two sides of the same coin: forming a generalization at a more abstract level on the prosodic hierarchy could lead to unbounded copying.[9] A prosodic word could easily introduce unboundedly many feet, and hence unbounded many syllables and unboundedly many segments. A prosodic word might even have a self-embedded nature:

[9] The prosodic hierarchy could be a natural way to algorithmize the kind of computation involved in unbounded copying; see Section 3.6.3 in Chapter 3 for more discussions.

based on morphophonological patterns of prefixes in Kaqchikel (Mayan), Bennett (2018) argues that a prosodic word may be nested into another prosodic word, and this recursive structure could in principle have an unbounded depth of embedding.

In terms of the foot-syllable level, it seems that a foot usually introduces a limited number of syllables, mostly bimoraic/bisyllabic. The notion of a ternary foot (i.e. have three syllables) was entertained but its status is under debate (Dresher and Lahiri, 1991; Hayes, 1995; van der Hulst, 2000). Though it seems impossible to introduce unboundedly many syllables from one foot (but see arguments for recursive footing in Bennett, 2012), we wonder whether the length-based implications still hold in the process of learning. Namely, when participants are given *one* syllable that is equally consistent with the generalizations of copying a foot, could they extrapolate to this foot-copying analysis and copy *two* syllables?

On a similar note, logically speaking, a syllable might have unbounded many segments. Zec (2007, p.164) made the remark that "if more than one consonant is allowed in a margin, there is in principle no limit to the number permitted," though we do observe that in most languages, each syllable rather contains a restricted number of segments because of the segment sequencing requirements regulated by the principles of sonority (Steriade, 1982; Selkirk, 1984). In reduplication learning, we have seen a clear piece of evidence against imposing a length restriction on the number of segments for extrapolating to syllable-level generalizations. In Experiment Series 1, when familiarized with syllables containing three segments, participants could generalize to new target syllables containing four segments (as well as two segments).

In all, a positive answer to the question of whether human learners can extrapolate unbounded copying from the bounded input will provide a positive answer to whether they construct hypotheses at the more abstract level. We conducted three artificial grammar learning experiments to directly address these questions. Now, we turn to the introduction of the general design of this new experiment series.

The general design Experiments in the second series aim to address phonological abstractions of levels beyond a syllable. In the spirit of the first three experiments, we used the poverty of the stimulus design. We familiarized participants with two patterns based on monosyllabic stems – one pattern was compatible with total reduplication while one was not. The familiarized patterns are summarized in (21) and (22). This time, the exposure was even more limited: participants were only given half of the familiarized items, namely only four pairs of singular and reduplicated plural pairs.

(21) The familiarized pattern in Experiment 2a + 2b

Examples (4)	*Singular*	*Reduplicant*	*Base*
[ˈpif] → [ˈpif-pif]	$C_1V_2C_3$	$C_1V_2C_3$	$C_1V_2C_3$
[ˈzæb] → [ˈzæb-zæb]	3 segs	3 segs	3 segs
	1 syllable	1 syllable	1 syllable

(22) The familiarized pattern in Experiment 2c

Examples (4)	*Singular*	*Reduplicant*	*Base*
[ˈpif] → [ˈpi-pif]	$C_1V_2C_3$	C_1V_2	$C_1V_2C_3$
[ˈzæb] → [ˈzæ-zæb]	3 segs	2 segs	3 segs
	1 syllable	1 syllable	1 syllable

Just as in the first three experiments, the familiarized trials in this experimental series were compatible with multiple hypotheses. Possible generalizations for Experiments 2a and 2b included but were not limited to a generalization of copying the full word (total reduplication), a size-restricting generalization of copying just a heavy syllable, and a similar size-restricting generalization of copying a first syllable. To determine what hypotheses were adopted, the testing items ought to grow in size.

Note that the ambiguity presented in the familiarization phase was not only limited to the shapes of a reduplicant (Dimension I discussed in the context of typological variation; see Table 2.2). If human learners indeed extrapolate to a size-restricting hypothesis and hence copy partially, the familiarized items are not informative at all to determine which part of the stem is copied and the relative position between these two copies. The familiarized patterns were both compatible with copying pivoted at the left word edge, the right word edge, and the primarily stressed syllable. This level of ambiguity concerns Dimension II of the typological variations reviewed in Table 2.2. (23) sketches the logically possible generalizations.[10]

(23) Logically possible generalizations assuming input [ˈpif-pif] and [ˈzæb-zæb].

a. the full word — total reduplication

b. a foot that may encompass two (or three) syllables — Ft

c. always a heavy syllable pivoted to the left word edge — $_{wd}[\sigma_{\mu\mu}$

d. always a heavy syllable pivoted to the right word edge — $\sigma_{\mu\mu}]_{wd}$

e. always a heavy syllable pivoted to the primary stress — $^{\text{ˈ}}\sigma_{\mu\mu}$

f. ...

To effectively tease apart these hypotheses, we designed five testing types, as summarized in Table 2.12. The predicted reduplicants under each possible generalization are given in Table 2.13. Each testing type had a growing number of segments and/or syllables. The PENTASYLLABIC forms had a primary stress on the antepenult and a secondary stress on the word-initial syllable. Within all other testing types, the initial syllables were set to be primarily stressed. Here, every novel singular was presented to be monomorphemic and the only relevant morphosyntactic operation is pluralization. In other words, the morphosyntactic structures of all these stems and their reduplicated forms remain the same but the forms grow by concatenating more phonological constituents, and incorporations of these larger constituents are evidence of a phonological force in unbounded copying.

In terms of Expt. 2c, intact copying of the whole word no longer makes a good hypothesis as it is not compatible with the familiarized items. Participants were expected to adopt a partial reduplicative analysis. The reduplicant shapes for these words could be fixed as a light syllable. But participants could also base their generalizations on more coarse-grained prosodic units and/or learn that the non-incorporation of the coda was due to a phonological

[10] Note that all mentioned generalizations collapsed along the dimension of the relative positions between two copies, and assumed the adjacency comes for free. If one believes in that non-local copying is real (Riggle, 2004a), then, the participants not only need to figure out which to copy, but also need to figure out where to place them.

Testing types	*Shapes*	*Examples*	*# Seg.*	*# σ*
FAMILIAR	ˈ$C_1V_2C_3$	[ˈnoʊg]	3	1
DISYLLABIC CV	ˈ$C_1V_2.C_3V_4C_5$	[ˈti.kɛp]	5	2
DISYLLABIC CVC	ˈ$C_1V_2C_3.C_4V_5C_6$	[ˈdɛb.gɪv]	6	2
TRISYLLABIC	ˈ$C_1V_2.C_3V_4.C_5V_6C_7$	[ˈti.fæ.pəs]	7	3
PENTASYLLABIC	ˌ$C_1V_2.C_3V_4.$ˈ$C_5V_6.C_7V_8.C_9V_{10}C_{11}$	[ˌpi.sæ.ˈgoʊ.bɛ.kʊt]	11	5

Table 2.12: *Novel singulars for Experiment 2*

Hypothesis	DISYLLABIC CV [ˈti.kɛp]	DISYLLABIC CVC [ˈdɛb.gɪv]	TRISYLLABIC [ˈti.fæ.pəs]	PENTASYLLABIC [ˌpi.sæ.ˈgoʊ.bɛ.kʊt]
total	**ˈti.kɛp**-ˈti.kɛp	**ˈdɛb.gɪv**-ˈdɛb.gɪv	**ˈti.fæ.pəs**-ˈti.fæ.pəs	**ˌpi.sæ.ˈgoʊ.bɛ.kʊt**-ˌpi.sæ.ˈgoʊ.bɛ.kʊt
$_{wd}[$FT	**ˈtik**/**ˈtikɛ**-ˈti.kɛp	**ˈdɛb**/**ˈdɛb.gɪ**-ˈdɛb.gɪv	**ˈtif**/**ˈtifæ**-ˈti.fæ.pəs	**ˌpis**/**ˌpisæ**-ˌpi.sæ.ˈgoʊ.bɛ.kʊt
$_{wd}[\sigma_{\mu\mu}$	**tik**-ˈti.kɛp	**dɛb**-ˈdɛb.gɪv	**tif**-ˈti.fæ.pəs	**pis**-ˌpi.sæ.ˈgoʊ.bɛ.kʊt
$\sigma_{\mu\mu}]_{wd}$	ˈti.kɛp-**kɛp**	ˈdɛb.gɪv-**gɪv**	ˈti.fæ.pəs-**pəs**	ˌpi.sæ.ˈgoʊ.bɛ.kʊt-**kʊt**
ˈ$\sigma_{\mu\mu}$	**tik**-ˈti.kɛp	**dɛb**-ˈdɛb.gɪv	**tif**-ˈti.fæ.pəs	ˌpi.sæ.-**goʊb**-ˈgoʊ.bɛ.kʊt

Table 2.13: *The predicted reduplicants under each copying-based generalization for Experiment 2a and 2b. Note other generalizations are also possible, such as an allomorphy-based analysis and non-adjacent copies. The variations should be freer than what is presented here.*

process that employed deletion. Orthogonal to the shape of a partial copy, the ambiguity of which portion to copy persists. A summary of possible generalizations is as in (24).

(24) Logically possible generalizations assuming input [ˈpi-pif] and [ˈzæ-zæb].

a. the full word without the final coda — WD+NOFINALCODA

b. a foot that disprefers codas — FT + NOCODA

c. always a light syllable pivoted to the left word edge — $_{wd}[\sigma_\mu$

d. always a light syllable pivoted to the right word edge — $\sigma_\mu]_{wd}$

e. always a light syllable pivoted to the primarily stressed syllable — ˈσ_μ

f. ...

Of course, all of the previous discussions are based on the assumption that participants can extend copying-based generalizations to new forms. It is worth emphasizing again that participants only received four pairs of familiarized items – this amount of familiarization may make the allomorphy analysis more appealing than in the first experiment series, as their choices of possible allomorphs are much smaller.

Hypothesis	Disyllabic CV [ˈti.kɛp]	Disyllabic CVC [ˈdɛb.gɪv]	Trisyllabic [ˈti.fæ.pəs]	Pentasyllabic [ˌpi.sæ.ˈgoʊ.bɛ.kʊt]
Wd + NoFinalCoda	**ˈti.kɛ**-ˈti.kɛp	**ˈdɛb.gɪ**-ˈdɛb.gɪv	**ˈti.fæ.pə**-ˈti.fæ.pəs	**ˌpi.sæ.ˈgoʊ.bɛ.kʊ**-ˌpi.sæ.ˈgoʊ.bɛ.kʊt
$_{wd}$[Ft+NoCoda	**ˈti.kɛ**-ˈti.kɛp	**ˈdɛb.gɪ**-ˈdɛb.gɪv	**ˈti.fæ**-ˈti.fæ.pəs	**ˌpi.sæ**-ˌpi.sæ.ˈgoʊ.bɛ.kʊt
$_{wd}[\sigma_\mu$	**ti**-ˈti.kɛp	**dɛ**-ˈdɛb.gɪv	**ti**-ˈti.fæ.pəs	**pi**-ˌpi.sæ.ˈgoʊ.bɛ.kʊt
$\sigma_\mu]_{wd}$	ˈti-**kɛ**-kɛp	ˈdɛb-**gɪ**-gɪv	ˈti.fæ.-**pə**-pəs	ˌpi.sæ.ˈgoʊ.bɛ-**kʊ**-kʊt
ˈσ_μ	**ti**-ˈti.kɛp	**dɛ**-ˈdɛb.gɪv	**ti**-ˈti.fæ.pəs	ˌpi.sæ-**goʊ**-ˈgoʊ.bɛkʊt

Table 2.14: *The predicted reduplicants under each copying-based generalization for Experiment 2c. Note other generalizations are also possible, such as an allomorphy-based analysis. The variations should be freer than what is presented here*

2.5.2 Experiment 2a: Monosyllabic CVC copying

2.5.2.1 Methods

Participants 87 participants were recruited through the Prolific crowdsourcing website and were self-identified as monolingual native English speakers. They were compensated with $4.00 for completing the experiment. Data from six participants were excluded due to the experimentors' scripting error.[11] One participant was excluded because of exposure to Serbian from infancy, some dialects of which may involve adjectival reduplication according to Brdar (2013). One participant was replaced because they produced silent responses for 90% of the testing trials. One participant was replaced because they just repeated back the stems for the plural forms. This led to the data from 78 participants being transcribed and analyzed (50 females, 28 males; mean age = 41, age range = 19-70). Nine participants reported exposure to Spanish after age seven, and Spanish does not contain productive reduplication.

Materials The familiarization phase consisted of four pairs of singulars and plurals (see Appendix B.1). All singulars were monosyllabic $C_1V_2C_3$ words (e.g. [pif], [zæb]).[12] The

[11] For these participants, within each trial of the Pentasyllabic testing type, different frames contained different recordings of the stem, which may lead to participants' confusion.

[12] Stimuli were intended to be nonce words. The status of the form being a non-English word was verified against a modified version of the CMU Pronouncing Dictionary, accessed via https://linguistics.ucla.edu/people/hayes/EnglishPhonologySearch/Index.htm. This contains a subset of words from the CMU Pronouncing Dictionary that have a frequency of at least 1 in the CELEX database (Baayen et al., 1995). This set of words was used as the representative lexicon likely to be known to English participants in the previous experimental works, such as Daland et al. (2011). During the experiments, some

corresponding plurals had the initial $C_1V_2C_3$ repeated locally, thus of the shape $C_1V_2C_3$ -$C_1V_2C_3$ (e.g. [pif-pif], [zæb-zæb]). C_1 and C_3 were drawn from a set of obstruents /p, b, t, d, k, g, f, v, z, s/, and were set to be different from each other. V_2 was one of the four vowels /i, u, æ, ɔ/. Each segment appeared exactly once in the familiarization. To maximize perceptual saliency and articulatory ease, C_1 and C_3 were selected such that the C_3C_1 cluster agree in voice and appear as a legal cluster in English.

Testing items were designed to vary in length systematically. Five types of testing items were summarized in Table 2.3 (see Appendix B.1.2.2). As a baseline, four novel testing items shared the same shape $C_1V_2C_3$ as the familiarized items, hence **Familiar** type. Different from the familiarized items, two testing items had the nasal [m] and [n] as the onset, which never appeared in the input. Moreover, V_2 was chosen from the set of vowels that were never used in the familiarized items, /eɪ, ɪ, ʊ, oʊ/, each occurring once. The other consonants were chosen from the same set of consonants /p, b, t, d, k, g, f, v, z, s/ with the same requirement on the C_3C_1 cluster.

For longer words, we first generated candidate CV and CVC syllables, and hope to combine them to form longer words. We generated two sets of CVC syllables, one for word-final position and one for word-initial positions. The vowels in the word-final syllable were realized as one of the vowels /ə, ʌ, ɔ, ɛ, ʊ, ɪ/. The vowels in CV syllables and word-initial CVC syllables were drawn from /ə, i, ɪ, eɪ, ɛ, æ, ɑ, ɔ, oʊ, ʊ, u/. All consonants were drawn from the candidate consonants the same as before, namely /p, b, t, d, k, g, f, v, z, s/.

Disyllabic CV-type items were named as such because they were of the shape $C_1V_2C_3V_4C_5$ (hence disyllabic) and the first syllables were of CV shape. For this testing type, candidate forms were generated by concatenating CV-shape syllables with word-final CVC syllables in a pseudorandom order. For the word-initial CV syllables, V_2 were required to be always tense /oʊ, eɪ, u, i/ to maximally attract stress. **Disyllabic CVC** type was very similar to **Disyllabic CV** type, except that the first syllables were CVC-shaped, resulting in

participants occasionally produced existing English words for some items, perhaps due to misperception or reflecting principles in loanword adaptation. However, this should not affect the experimental results and the conclusion, since the goal was to test *how* participants generalize to longer forms with only monosyllabic copying input, a process that never occurred in English.

$C_1V_2C_3C_4V_5C_6$ forms. Consequently, the first syllables were selected from the set of word-initial CVC syllables, and the second syllables from the word-final CVC syllables, and the vowel was never [ə]. Given C_3 and C_4 were all obstruents, a syllable boundary between C_3 and C_4 was guaranteed. That is, [dɛbgɪv] would be parsed as [dɛb.gɪv] but never *[dɛ.bgɪv] or *[dɛbg.ɪv]. To make both consonants as perceptible as possible, C_3 and C_4 were required to agree in voice and never form a fricative-fricative cluster. Similarly, **Trisyllabic** items were created by concatenating two CV syllables and word-final CVC syllables, and **Pentasyllabic** types concatenating four CV syllables with a word-final CVC syllable. We further excluded items that might lead to unattested consonant clusters in possible reduplicated forms – this is to minimize undesired avoidance of some forms due to independently motivated phonotactics.

The testing phase included 20 items, four for each of the five testing types. To ensure the robustness and generality of the results, we generated five lists of familiarized items and testing items. Within each list, each noun singular was randomly paired with a picture of an individual everyday object. The noun plurality was indicated by showing two instances of the same object. Stimuli were synthesized through the neural engine of Amazon Polly with Matthew Voice, prompted with their phonetic transcriptions. The synthesized tokens were resampled at a rate of 24 kHz and normalized for intensity to 65 dB.

Procedures The experiment was conducted online. Participants were instructed to participate in the study from a quiet room and encouraged to wear headphones throughout the experiment. After giving their consent, participants were required to complete an audio test for their microphones and headphones/external speakers. The main experiment started once they successfully passed the audio test.

Similar to Experiment Series 1, the main experiment consisted of two phases: a familiarization phase and a testing phase. Before the familiarization phase, participants were told that they would be learning how to form plurals in a new language. In each familiarization trial, participants listened to a monosyllabic singular, followed by the reduplicated plural, with pictures appearing on the screen. In this experiment, participants were asked to repeat the singular word and the reduplicated plural forms. After four pairs of exposure, participants proceeded to the testing phase and were prompted to apply what they learned in the familiarization phase to novel singulars. Each participant was randomly assigned to one of the five testing lists, and we ensured each list had at least 15 participants.[13] For each list, 20 items were tested together in a randomized order, and each item was tested once. In each testing trial, participants were expected to listen to a novel singular, record their repetition of the singular form, and give their production response for the corresponding plural form by themselves. The experiment ended with a questionnaire, asking participants to describe the formed rules if possible. The full experiment took roughly 15 - 20 minutes to complete.

2.5.2.2 Analyses and results

The data were transcribed by the phonetically trained author and research assistants. The data were straightforward to transcribe. Similar to Experiment Series 1, we annotated the participants' plural responses into two parts: a base and an affix and their responses were clear and straightforward to identify the two parts. The base part was identified as the longer constituent that maximally aligned with the singular and the affix was the shorter constituent. Sometimes, we observed participants infixed the reduplicant into the base, and the two parts were easy to differentiate. As in Experiment Series 1, participants' repetitions largely followed the reduplicative rule, and the average segmental similarity between the affix and the base all exceeded 0.95. All participants showed the intended CVC copying for at least two trials. [14]

Are generalizations copying-based? The Monte Carlo procedure described in Experiment 1a in Section 2.4.3.1 established that all participants adopted some amount of copying-

[13] There are 15 participants for list 1 and list 2, and each of the other three lists has 16 participants.

[14] There were twelve participants produced CV partial reduplication for one familiarized item (e.g., [ˈfuk-fu]), one participant showed CV partial reduplication for two trials. Two participants produced CV total reduplication for one trial (e.g., [ˈvi-vi]) and they repeated the wrong stems as well (e.g., [ˈvi]). We did not exclude these participants for a rather conservative test. As a spoiler, we found a strong preference for unbounded copying/total reduplication. Logically speaking, including these participants could only make such a trend weaker, as they have seen some input against this analysis.

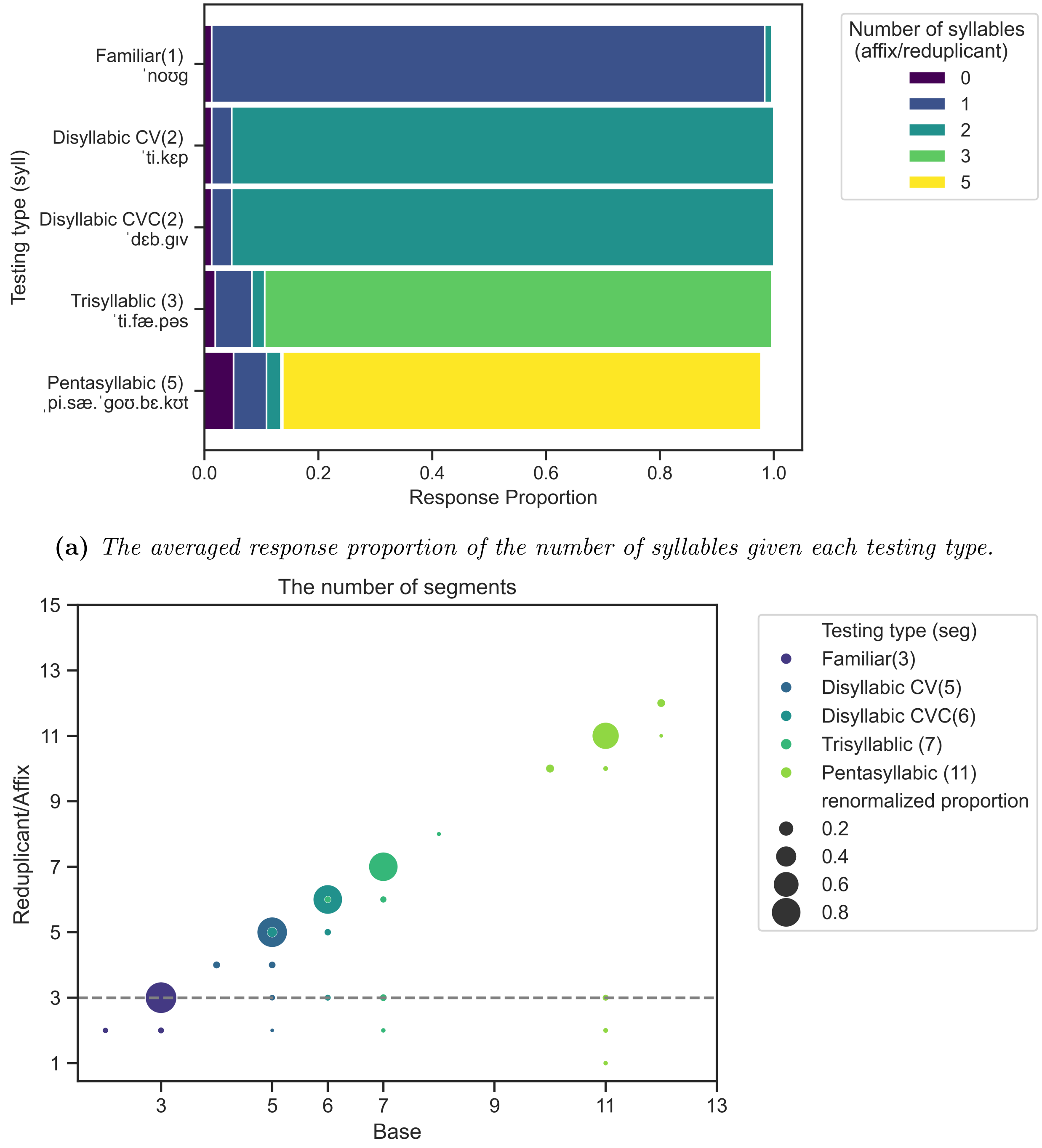
based generalization (see Figure B.3 in the Appendix B). For this experiment series, we also performed entropy analyses based on participants' responses. Entropy is a notion from information theory to quantify uncertainty in the space of possible outcomes of a random variable, the formula of which is given below in (25).

(25)

$$H(X) = -\sum_{x \in X} p(x) lnp(x)$$

Treating the affix responses with segmental substances as a random variable, then, $p(x)$ is the proportion of each possible realization in the actual responses. A larger entropy value indicates more possible outcomes in participants' responses and a relatively flat distribution over each potential outcome. To see this, let us first consider a total reduplication generalization. In this case, each novel singular would result in a different response, thus, leading to 20 different responses for 20 different stems and each occurring exactly once ($p(x) = \frac{1}{20}$). Plugging into the formula, the entropy value for this scenario is 2.99 nats. If a participant uses the familiarized reduplicants as allomorphs and learns their distribution should be equally likely, then the entropy value is 1.6 nats ($p(x) = \frac{1}{4}$). At the other end of the spectrum, if a participant has used a fixed allomorph as the only possible realization for the affix, the corresponding entropy value is 0 nats ($p(x) = \frac{1}{1}$). We performed two entropy analyses of participants' responses: one was based on realizations with segmental substances and the other was based on shapes characterized by syllabicity – the results of both analyses were presented in Figure 2.27 together with results from Expt. 2b and Expt. 2c. For the results computed based on the segmental substances, all participants showed entropy values greater than 2.9 – this excludes the possibility that they have used fixed allomorphy for the affix realization. Together with the Monte Carlo simulation on segmental faithfulness, these results strongly support that all participants have adopted active copying-based generalizations.

Generalizing at what levels of abstraction? If participants' generalizations were copying-based but not allomorphy-based, at what level of phonological abstractions did they generalize? For this experiment, we are most curious about whether participants were able to extrapolate unbounded copying from bounded input.



(a) *The averaged response proportion of the number of syllables given each testing type.*

(b) *The average proportion of the pair of base segment number and affix segment number in Experiment 2a ($p > 0.01$). The horizontal dashed line indicates the number of segments in the familiarized items, namely 3.*

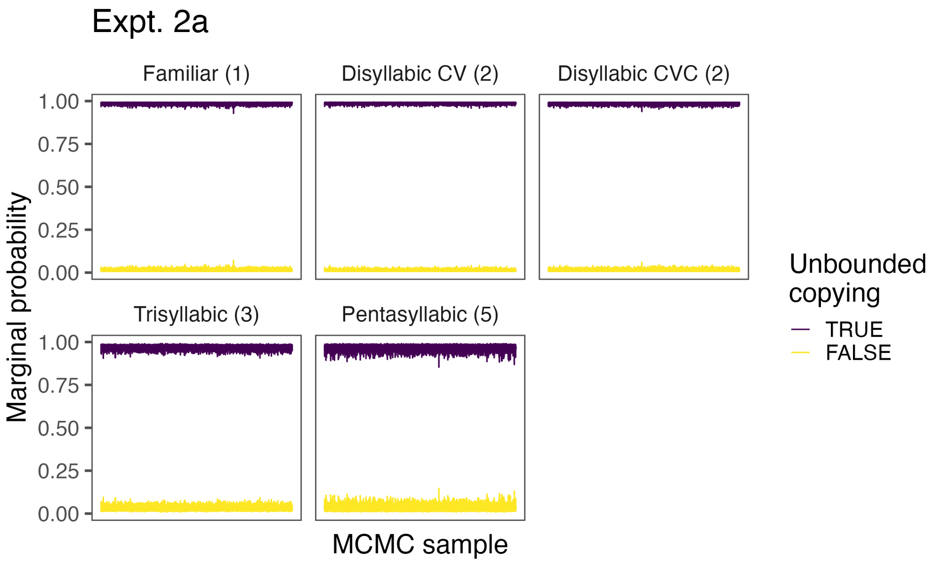
Figure 2.23: *The averaged response proportion of the number of phonological materials in the affix realization conditioned on each testing type in Experiment 2a.*

As a preliminary glance, Figure 2.23 shows the average response proportion of the number of phonological materials in the affix realizations conditioned on each testing type. As demonstrated in Figure 2.23a, for each testing type, instead of having a fixed number of syllables in the affix, the most frequent syllable number in the affix grew together with the syllable number in the base. For **Familiar**, namely monosyllabic stem, the most frequent affix shape (97%) was also monosyllabic, as in [ˈnoʊg-ˈnoʊg]. For both **Disyllabic CV** and **Disyllabic CVC**, participants preferred to give a disyllabic response (95%), as in [ˈti.kɛp-ˈti.kɛp] and [ˈdɛb.gɪv-ˈdɛb.gɪv]. For **Trisyllabic** stems, the most frequent syllable shape was trisyllabic (88%) as in [ˈti.fæ.pəs-ˈti.fæ.pəs] and the most frequent affix shape for **Pentasyllabic** stem was of five syllables (83%) as in [ˌpi.sæ.ˈgoʊ.bɛ.kʊt-ˌpi.sæ.ˈgoʊ.bɛ.kʊt]. Figure 2.23b confirmed the same trend at a segmental level: according to the average proportion of the segment number pair conditioned on each testing type, the segment number of the base and that of the affix formed a linear trend. Compared this to Figure 2.22, this is clear evidence that participants have formed a supra-finite-state kind of computation.

Bayesian mixed-effect multinomial logistic regression modeling established the same results that participants were able to extrapolate unbounded copying from bounded inputs. We performed the same Bayesian logistic regression model described in Experimental Series 1 in `rStan`. The dependent variable was Unbounded copying (Unbounded; TRUE versus FALSE).[15] The models included Testing type as fixed effects and by-subject random effects. For the prior and the model specification, see Appendix B. In general, the priors for these models assumed a mean-zero normal distribution with varying standard deviations. Based on the sampled posterior probabilities of the population-level parameters, we calculated the posterior probabilities of each possible outcome and performed post-hoc pairwise comparisons between testing types and within testing types. Our results confirmed that Unbounded copying was indeed preferred within all testing types. At the same time,

[15] A response was coded to be Unbounded copying if either of these conditions holds: (a). if the realization of the base completely matched exactly with the realization of the copy and there was no syllable-level truncation on both copies; (b). For polysyllabic words, if the number of syllables in the reduplicant equaled the number of syllables in the base, there was no syllable-level truncation on both copies, and the base-reduplicant faithfulness exceeded 0.9.

there was no significant difference between testing types. Table 2.15 shows the mean of the sampled posterior probabilities calculated based on the sampled posterior probabilities of the population-level parameter estimates.



Testing	type t	$\overline{p(\textsc{Unbounded}\|t)}$	$\overline{p(\textsc{Others}\|t)}$
Familiar	ˈnoʊg	$\boxed{\mathbf{0.99}}$	0.01
Disyllabic CV	ˈti.kɛp	$\boxed{\mathbf{0.99}}$	0.01
Disyllabic CVC	ˈdɛb.gɪv	$\boxed{\mathbf{0.99}}$	0.01
Trisyllabic	ˈti.fæ.pəs	$\boxed{\mathbf{0.97}}$	0.03
Pentasyllabic	ˌpi.sæ.ˈgoʊ.bɛ.kʊt	$\boxed{\mathbf{0.97}}$	0.03

Table 2.15: *The sampled posterior probabilities calculated based on the population-level parameter estimates, with the table showing the mean value. For all testing types, producing an unbounded copying response is more probable than others (p < 0.01).*

On the other hand, not all participants had extrapolated unbounded copying: some showed copying a fixed number of syllables (monosyllabic) across all testing types in Figure 2.23a, reflected by smaller dots patterning along the horizontal line (i.e. the reduplicant always has three segments) in Figure 2.23b. Because these patterns were spontaneous outputs, they are also of interests here for a full picture of the learned outcomes. We will discuss these patterns in Section 2.5.5.

2.5.2.3 Interim summary

Participants in Expt. 2a predominantly extended unbounded copying to longer forms – this held for three-syllable and five-syllable novel singulars. This finding further suggests that participants had formed abstractions beyond the level of syllable and were readily able to copy the full novel prosodic word.

Yet the experimental design contained a potential confound, which made the origin of such a preference unclear: in this experiment, the pictures for plural forms were simply the same pictures of the singular form shown twice. Hence, if participants had associated the phonological forms with the visual representations but not the underlying concept, seeing a repetition of the same picture may prompt them to simply repeat the whole stem. On the other hand, the size-restricting hypotheses were not compatible with such a visual correlate.[16] With this in mind, we conducted another version of the same experiment to check whether the preference could be attributed to the visual cue. We changed the picture of the plural forms to indicate true pluralization, with many items of the same concept appearing in the same picture frame. As a preview, the preference for unbounded copying still holds.

2.5.3 Experiment 2b: Monosyllabic CVC copying in a different setting

2.5.3.1 Methods

57 self-reported English native speakers who did not participate in the previous experiment were recruited through Prolific. Five participants were excluded because they gave silent/inaudible recordings beyond 90% of the trials. One participants were excluded because of exposure to Hebrew, which contains productive reduplication (Bat-El, 2006). Other participants reported exposure to Spanish, French, Italian, and Swedish – none of these languages involve productive reduplication as a part of their morphophonological systems.[17] Data from 51 participants were transcribed and included into analyses (15 males, 31 females,

[16] Many thanks to Laurel Perkins for pointing out this issue.

[17] Some nicknames in Swedish are reduplicated (Riad, 2014, p158) and there is no other reported use.

5 others; mean age = 42, age range = 18 - 71). All familiarized and testing singulars were approximately the same as Expt. 2a,[18] except that the pictures to motivate the semantics of pluralization showed multiple items appearing in the same frame, instead of two repetitions of the same pictures. We ensured that each list had at least 10 participants. [19]

2.5.3.2 Analyses and results

Consistent with Expt. 2a, the repetition accuracy of the trained items was high. The repeated responses of the plural forms largely followed the intended familiarized rule. The average segmental similarity between the affix realization and the base all exceeded 0.98. All participants repeated the intended CVC copying at least twice. [20]

Are generalizations copying-based? We conducted the same Monte Carlo procedure and entropy analyses as described before, and found all participants had their faithfulness analyses fall beyond the 99% confidence interval of chance. An analysis of the entropy of their affix realizations showed convergent support (as in Figure 2.27): for all participants, the entropy values computed based on the segmental substances were greater than 2.3. In fact, upon a visual inspection of Figure 2.27, one can easily see that participants in Expt. 2b patterned with Expt. 2a in the same way.

Generalizing at what level of abstraction? Figure 2.16 shows the average response proportion of the number of phonological materials in the affix realizations conditioned on each testing type. Similar to Expt. 2a, Figure 2.24a establishes that the most frequent

[18] Due to the experimentor's scripting error, the Disyllabic CV testing type in list 3 mistakenly used four items of Disyllabic CV items in list 4. However, this should not affect the experiment, nor the conclusion, as these four items in list 3 did not introduce any potential confounds.

[19] 11 participants were recruited for list 3 and each other four lists had 10 participants.

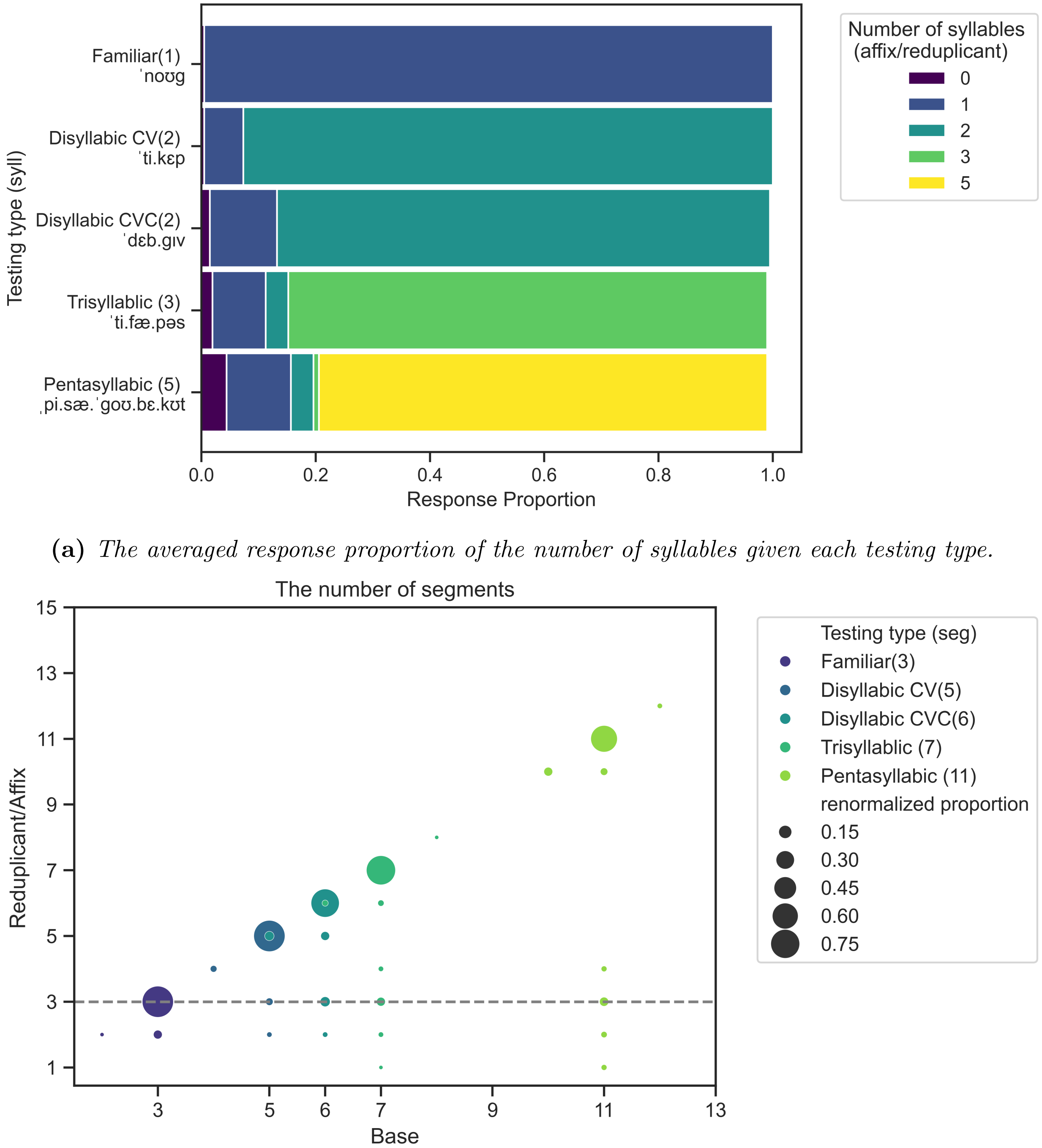
[20] Six participants produced CV partial reduplication for one familiarized trial (e.g., [ˈbæv-bæ]), one participant showed CV partial reduplication for two trials. One participant produced CV total reduplication for one trial (e.g., [ˈpu-pu]) and they repeated the wrong stems as well (e.g., [ˈpu]). Similar to the rationale in Expt. 2a, we did not exclude these participants for a rather conservative test. One participant provided silent recordings for the familiarized tokens but exhibited systematic copying-based generalizations in the testing phase. We did not exclude this participant.

number of syllables of the affix grew together with the stem/base. Figure 2.24b confirmed the same trend at a segmental level: based on the pairs of segment numbers condition on each testing type, the segment number of the base and that of the affix formed a linear relationship, suggesting that participants predominantly extrapolated a supra-finite-state computation. On the other hand, consistent with Expt. 2a, not all participants had formed a generalization of unbounded copying: there was a minority trend of having a fixed number of syllables (monosyllables and two syllables) in Figure 2.24a, mirrored by dots patterning along the horizontal line (i.e. always having three/four segments) in Figure 2.24b. We will discuss these minority patterns in Section 2.5.5.

We ran the same Bayesian logistic regression in `rStan` and conducted pairwise comparisons between testing types and within testing types. Our results are given in Table 2.16. The possible response type was unbounded copying or not. Our results confirmed that UNBOUNDED COPYING was indeed preferred within all testing types (all $p < 0.01$). There is no significant difference between testing types.

2.5.3.3 Interim summary

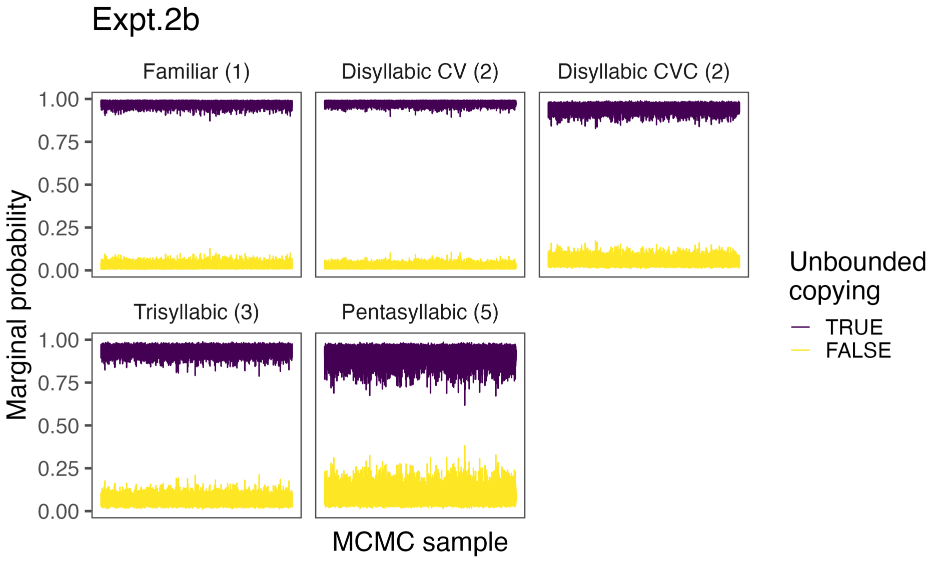
Together with Expt. 2a, this experiment shows strong support that participants preferred an unbounded copying response within all testing types, which suggests that strict finite-state might not be the right sense to characterize the hypotheses space for morphophonology. On the other hand, a true primitive copying operation should be incorporated, with its domain open to unboundedly many phonological constituents. The positive answer towards unbounded copying further indicated that participants generalized with a more abstract shape than a syllable. Specifically, they were able to copy a full prosodic word. Now, we turn to Experiment 2c, which differed minimally from the previous two experiments in familiarizing participants with CV copying (['pif] ↦ ['pi-pif]).



(a) *The averaged response proportion of the number of syllables given each testing type.*

(b) *The average proportion of the pair of base segment number and affix segment number in Experiment 2b ($p > 0.01$). The horizontal dashed line is the number of segments in the familiarized items, namely 3.*

Figure 2.24: *The averaged response proportion of the number of phonological materials in the affix realization conditioned on each testing type in Experiment 2b.*



Testing	type t	$\overline{p(\text{Unbounded}\|t)}$	$\overline{p(\text{Others}\|t)}$
Familiar	ˈnoʊg	$\boxed{\mathbf{0.97}}$	0.03
Disy CV	ˈti.kɛp	$\boxed{\mathbf{0.98}}$	0.02
Disy CVC	ˈdɛb.gɪv	$\boxed{\mathbf{0.95}}$	0.05
Trisyllabic	ˈti.fæ.pəs	$\boxed{\mathbf{0.94}}$	0.06
Pentasyllabic	ˌpi.sæ.ˈgoʊ.bɛ.kʊt	$\boxed{\mathbf{0.92}}$	0.08

Table 2.16: *The sampled posterior probabilities calculated based on the population-level parameter estimates, with the table showing the mean value. For all testing types, producing an unbounded copying response is significantly more probable than others ($p < 0.01$).*

2.5.4 Experiment 2c: Monosyllabic CV copying

2.5.4.1 Methods

105 self-reported English native speakers who did not participate in the previous experiment were recruited through Prolific. Seven participants were excluded because of the same scripting error as Experiment 2a. One participant was excluded because they gave silent responses for all trials. One participant was replaced due to failure to understand the task, giving total reduplication suffixed with English plural marker for all trials. Three participants were excluded because of exposure to Hebrew (Bat-El, 2006), American Sign Language (Abner, 2017), and Latin (Zukoff, 2017), which makes use of productive partial reduplication. Other participants reported exposure to Japanese, French, German, Spanish, which contained no

productive partial reduplication. Data from 93 participants were transcribed and included (30 males, 56 females, and 1 other; mean age = 39, age range = 17 - 68). All training and testing singulars were the same as Experiment 1a, except now participants were trained with CV copying (e.g. [ˈpif], plural [ˈpi-pif]; [ˈzæb], plural [ˈzæ-zæb]. We ensured each list contained at least 15 participants. [21]

2.5.4.2 Analyses and results

Participants in Expt. 2c also showed high repetition accuracy of the familiarized items. The repeated responses of the plural forms largely followed the intended familiarized rules. The average segmental similarity between the affix realization and the base all exceeded 0.95. We further excluded eight participants whose repetitions failed to comply with CV partial reduplication for at least 75% familiarized items (that is, three out of four items), so that we were confident to conclude the input indeed supported the intended familiarized pattern.

Are generalizations copying-based? For Experiment 2c, the Monte Carlo analysis indicates that twelve participants had their observed affix-base similarity fall into the 99% confidence interval of chance, as in Figure B.3 in Appendix B, which seemed to suggest that these participants had failed to adopt a copying-based generalization. The entropy analyses largely converged on the same conclusion: there were nine participants with low entropy of their responses (<2.0), suggesting that they may have adopted allomorphy-based generalizations. Upon further inspection, although the other three participants showed relatively high entropy, their variable responses were completely irregular and not based on copying at all. We excluded these twelve participants from the subsequent analyses on the affix shapes. We briefly discuss their responses in Section 2.5.5.

[21] List 2 ended up having 20 participants. List 3 and list 5 ended up having 19 participants; list 1 had 18 participants; list 4 had 17 participants. The imbalanced number at this stage was due to a goal of having a roughly balanced number in the shape analysis.

Generalizing at what level of abstraction? To further investigate the copying-based generalizations participants might have formed,[22] we coded the shape of their responses based on the syllabic properties. There were five categories. First, a response was coded as Light syllable (σ_{μ}) if the reduplicant was a monosyllabic light syllable (mostly CV, as in [ˈtu-ˈtu.kɑs]); [23] Heavy syllable ($\sigma_{\mu\mu}$) covered monosyllabic reduplicants with a coda as in [fut-ˈfu.tʊs]. Two syllables ($\sigma\sigma$) encoded disyllabic reduplicants as in [ˌpæ.kə-ˈpæ.kə.væf]. Total was coded based on the syllable numbers in the reduplicant as well as the base, specifically when a response had a trisyllabic reduplicant for a trisyllabic stem (both [ˈki.fæ.tə-ˈki.fæ.təs] and [ˈki.fæ.təs-ˈki.fæ.təs]) or a pentasyllabic reduplicant for a pentasyllabic stem. Note that Total was only construed in broad terms, and it did not apply to other testing types with shorter singulars. Lastly, we collapsed the rest of the responses into the fifth category Others, which included no realizations, affixing a consonant (usually English pluralization), and the longer forms such as three-syllable realization as in [ˌvoʊ.voʊ.ˈfi.dæ.kəs-ˈfi.dæ.kəs] and four-syllable realization as in [ˌgɛ.zʊ.ˈbɑ.ki-ˌgɛ.zʊ.ˈbɑ.ki.vɪd]. All provided examples here were actual participants' responses.

As a preliminary check, Figure 2.25 shows the averaged response proportion: Figure 2.25a is based on the annotated affix shapes with the raw values provided in Table 2.17, and Figure 2.25b shows the pairs of the segment number in the base and that in the reduplicant. From Figure 2.25a and the numerical values in Table 2.17, we could observe that the most frequent shape was a light syllable across all testing types, which suggests that most participants had extrapolated partial copying with a rather fixed syllable shape. This trend was corroborated in Figure 2.25b: more frequent segment number pairs patterned together along a horizontally constant line representing two segments in the affix realizations. Recall that the familiarized items also had two segments as the reduplicant. This is different from the linear trend that emerged in the previous two experiments and suggests a bound on how much to copy. The predominant pattern does satisfy the theorized finite-state kind of

[22] The discussion of the results in this section was based on data from 73 participants. List 1, list 4, and list 5 had 15 participants; list 2 and list 3 had 14 participants.

[23] Sporadically, participants gave V or CCV responses due to misperception of the stem. The onset was base-dependent.

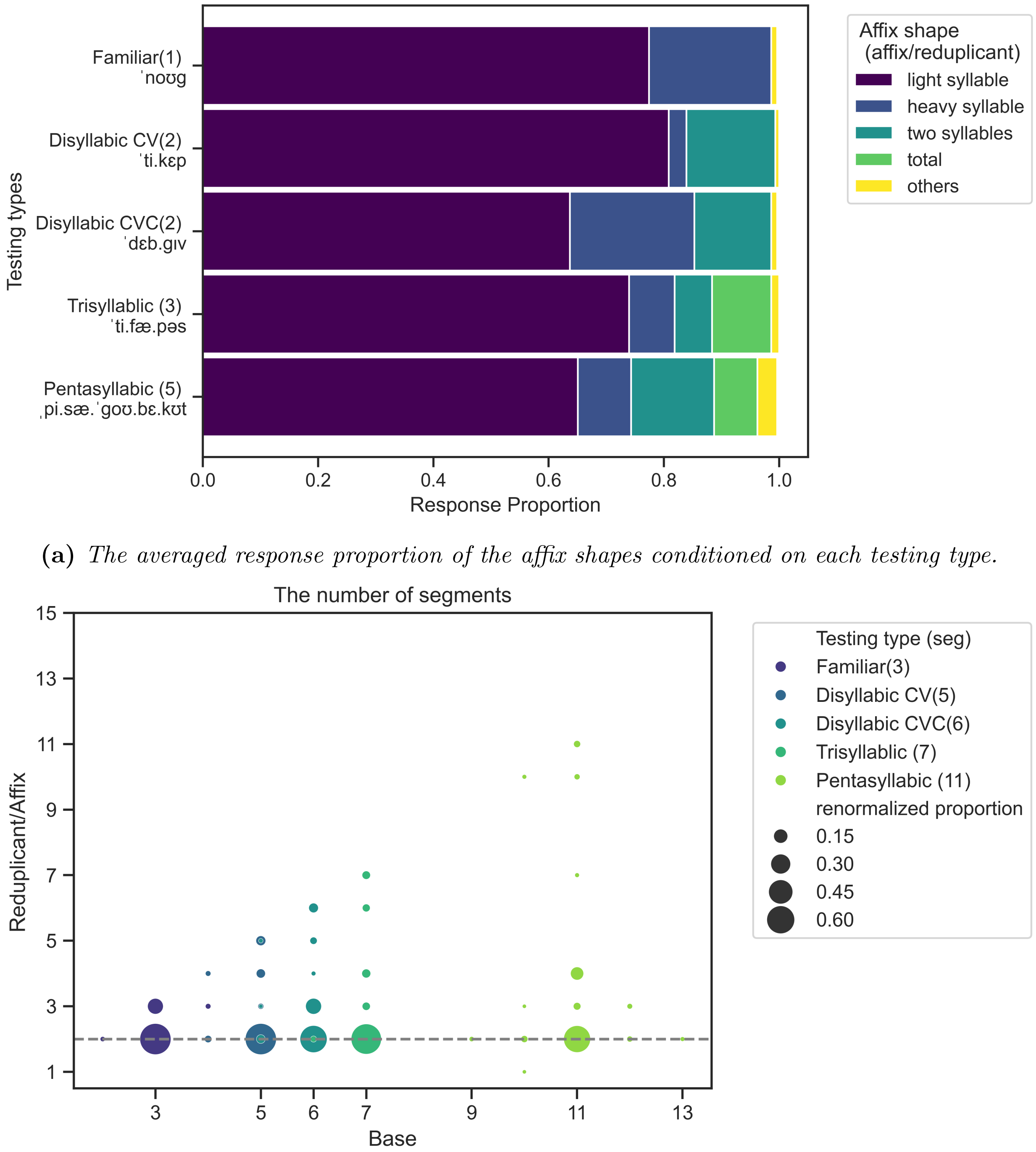
computation.

Testing type t		$\overline{p(\sigma_\mu\|t)}$	$\overline{p(\sigma_{\mu\mu}\|t)}$	$\overline{p(\sigma\sigma\|t)}$	$\overline{p(\text{total}\|t)}$	$\overline{p(\text{others}\|t)}$
Familiar	ˈnoʊg	$\boxed{\mathbf{0.77}}$	**0.21**	0.00	0.00	0.01
Disyllabic CV	ˈti.kɛp	$\boxed{\mathbf{0.81}}$	0.03	**0.15**	0.00	0.00
Disyllabic CVC	ˈdɛb.gɪv	$\boxed{\mathbf{0.64}}$	**0.22**	**0.13**	0.00	0.01
Trisyllabic	ˈti.fæ.pəs	$\boxed{\mathbf{0.74}}$	0.08	0.07	**0.10**	0.01
Pentasyllabic	ˌpi.sæ.ˈgoʊ.bɛ.kʊt	$\boxed{\mathbf{0.65}}$	0.09	**0.14**	0.08	0.03

Table 2.17: *The averaged response proportion of the affix shapes conditioned on each testing type.*

From the averaged response proportion as in Figure 2.25 as well as the cross-experimental comparisons based on the entropy analyses in Figure 2.27, we shall see that the results were, to a certain degree, more variable than in the previous two experiments and reflected greater between-subject variations. Generally speaking, beyond the fixed light syllable as the most frequent response shape, participants were still able to copy more than they had seen, with the notion of "seen" being built on the number of syllables or the number of segments. First, note that Total occurred at a non-negligible rate (11% for Trisyllabic forms and 8% for Pentasyllabic forms), which showed support for unbounded copying at the level of a prosodic word. Secondly, for Disyllabic CV forms and Disyllabic CVC, copying two syllables occurred at a considerable rate (∼ 14%), which was also observed in Trisyllabic items (∼ 7%; as in [ˌpi.su-ˈpi.su.fɑt]) and strikingly more in Pentasyllabic items (∼ 14%; as in [ˌgɑ.zu-ˌgɑ.zu.ˈfæ.bi.dɪb]), which indicates that some participants might have performed foot copying. The two findings provide further evidence for the possibility of extrapolating generalizations at the level of abstractions higher than a fixed syllable, though not the most preferred one. These observations were confirmed in Figure 2.25b: there exists a relatively minor linear trend, supporting the necessity of including unbounded copying.

Lastly, copying a heavy syllable seemed to occur at a higher rate in Disyllabic CVC (as in [kɪp-ˈkɪp.ʧʊf]) items and Familiar items (as in [ˈgɪs-gɪs]) than the other types. There were several plausible hypotheses on the motivation in these two conditions. One was the urge to copy a bimoraic foot immediately available at the left-word edge, especially given that



(a) *The averaged response proportion of the affix shapes conditioned on each testing type.*

(b) *The average proportion of the pair of base segment number and affix segment number in Experiment 2c* ($p > 0.01$)*. The horizontal dashed line is the number of segments in the familiarized items, namely 2.*

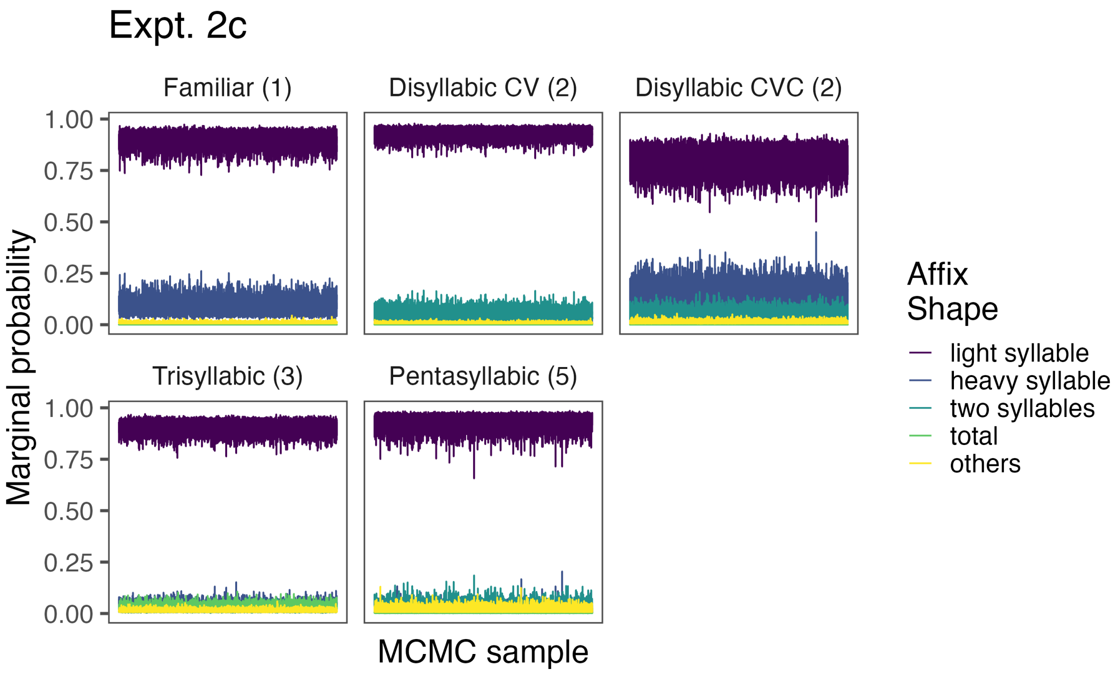
Figure 2.25: *The averaged response proportion of the number of phonological materials in the affix realization conditioned on each testing type in Experiment 2c.*

lax vowels were frequently produced as the target copied vowel in Disyllabic CVC cases, so participants needed to copy the coda. A rather wild possibility is syllable copying: all other testing types, namely Disyllabic CV, Trisyllabic and Pentasyllabic, all had a light syllable as the initial syllable. It could be that after a fair amount of light syllable copying for these testing types, the hypothesis of syllable copying became salient enough that prompted the participants to copy the first heavy syllable. Regardless of the underpinning motivation, the simple fact that there were more heavy syllable copying for Disyllabic CVC items and Familiar items seemed to point to a general preference to copy well-parsable constituents with respect to the phonological structures of the base.

Bayesian multinomial logistic regression model confirmed the trends described above. Table 2.25 provided the sampled posterior probabilities of each affix shape based on population-level parameter estimates with the mean of the sampled posterior probabilities organized in the corresponding table. First, within-testing type comparisons of possible shapes established that Light syllable was indeed the most predominant response (all $p < 0.01$). For Familiar forms, the second most preferred response was to copy a Heavy syllable (all $p <0.01$; [ˈgɪs-gɪs]). Within Disyllabic CV forms, the second most preferred response was Two syllables ([ˈzi.vɪb-ˈzi.vɪb]), which was significantly higher than the rest of the response shapes (all $p < 0.01$). As for Disyllabic CVC forms, the second most preferred response was to copy a Heavy syllable ([ˈkɪp-ˈkɪp.ʧʊf]): it was significantly higher than other types (all $p < 0.01$). The probability of producing a Two syllables ([ˈkɪp.ʧʊf]-[ˈkɪp.ʧʊf]) response was significantly higher than Total[24] and Others (all $p < 0.01$). Within Trisyllabic and Pentasyllabic forms, there is no significant difference among the other response types established.

As for between-testing type comparisons, Familiar, Disyllabic CV and Trisyllabic and Pentasyllabic forms were significantly more likely to have Light syllable responses than Disyllabic CVC (all $p < 0.05$). Secondly, same as described in the previous discussion, Disyllabic CVC and Familiar forms were more likely to yield Heavy syllable

[24] Note that Total only applied for Trisyllabic and Pentasyllabic forms.



Testing	type t	$\overline{p(\sigma_\mu\|t)}$	$\overline{p(\sigma_{\mu\mu}\|t)}$	$\overline{p(\sigma\sigma\|t)}$	$\overline{p(\text{total}\|t)}$	$\overline{p(\text{others}\|t)}$
Familiar	ˈnoʊɡ	$\boxed{\mathbf{0.90}}$	**0.09**	0.00	0.00	0.01
Disyllabic CV	ˈti.kɛp	$\boxed{\mathbf{0.93}}$	0.01	0.05	0.00	0.01
Disyllabic CVC	ˈdɛb.ɡɪv	$\boxed{\mathbf{0.80}}$	**0.14**	0.05	0.00	0.01
Trisyllabic	ˈti.fæ.pəs	$\boxed{\mathbf{0.91}}$	0.03	0.02	0.03	0.01
Pentasyllabic	ˌpi.sæ.ˈɡoʊ.bɛ.kʊt	$\boxed{\mathbf{0.93}}$	0.01	0.03	0.01	0.02

Table 2.18: *The sampled posterior probabilities calculated based on the population-level parameter estimates, with the table showing the mean value. For all testing types, the probability of producing a* Light syllable *response is significantly higher than other possible shapes (p < 0.01).*

reduplicant than Disyllabic CV, Trisyllabic, and Pentasyllabic (all $p < 0.001$). The difference between Disyllabic CVC and Familiar in producing Heavy syllable reduplicant was not significant ($p = 0.15$). Note that all syllables in Disyllabic CVC and Familiar were heavy, and the other three testing types had the word-initial CV syllables. We think these two results might provide evidence supporting a preference to copy a well-formed constituent with respect to the structure in the base.

The possible generalizations at a more abstract level than a syllable were supported. Disyllabic CV, Disyllabic CVC produced more Two syllables reduplicant than Familiar

Testing type t		$\overline{p(\text{Left}\|t)}$	$\overline{p(\text{Right}\,\|\,t)}$	$\overline{p(\text{Infixation}\,\|\,t)}$	$\overline{p(\text{Others}\,\|\,t)}$
FAMILIAR	ˈnoʊg	$\boxed{\mathbf{0.76}}$	0.03	0.00	**0.20**
DISYLLABIC CV	ˈti.kɛp	$\boxed{\mathbf{0.85}}$	0.02	0.03	**0.10**
DISYLLABIC CVC	ˈdɛb.gɪv	$\boxed{\mathbf{0.84}}$	0.03	0.03	**0.08**
TRISYLLABIC	ˈti.fæ.pəs	$\boxed{\mathbf{0.85}}$	0.04	0.05	**0.06**
PENTASYLLABIC	ˌpi.sæ.ˈgoʊ.bɛ.kʊt	$\boxed{\mathbf{0.80}}$	0.05	**0.08**	0.06

Table 2.19: *The averaged response proportion of the edge conditioned on each testing type.*

and TRISYLLABIC (all $p < 0.05$) – this is evidence for bisyllabic foot copying or prosodic word copying. Secondly, PENTASYLLABIC and TRISYLLABIC produced more TWO SYLLABLES reduplicants than FAMILIAR (all $p < 0.001$), suggesting that that the possibility of bisyllabic foot copying is non-trivially established. PENTASYLLABIC and TRISYLLABIC produced more TOTAL reduplicant than all other testing types (all $p < 0.001$), suggesting that copying at the prosodic word level received non-trivial support, albeit very small numerical values. Within the OTHERS response type, there was no significant difference between testing types established.

Edge-orientedness? Since most participants had extrapolated to a partial reduplication analysis, this led to the consideration of another level of ambiguity, namely, the relative position of the copy with respect to the stem. We coded the responses to be LEFT EDGE if the affix realization occurred at the left edge of the base (as in [deɪ-ˈdeɪz.gɪv]); RIGHT EDGE if the affix realization occurred at the right edge of the base [ˌtɑ.ki.ˈseɪ.və.sʌp-ˈseɪ.və.sʌp]); and INFIXATION if the affix realization occurred within the base ([ˈgɑ.vi-də-ˌdʌs]; [ˌgu.zi-ˌtoʊ.vɑ-ˈtoʊ.vɑ.fɛd]). We then collapsed the other possible answers to a fourth category OTHERS, which covered total reduplication ([ˈgɪb-ˈgɪb]) and no realization. Note that participants gave syllable-level adjacent copies, such as [ˈgɪb-gɪ], which would be regarded as non-adjacent copies at a segmental level. Such a response was coded as a RIGHT EDGE because the smaller copy occurred to the right. The non-adjacent syllable copies rarely occurred. Thus, the discussion here would roughly reflect which part of the stem is copied, concerning Dimension II of the discussed typological variation 2.2.

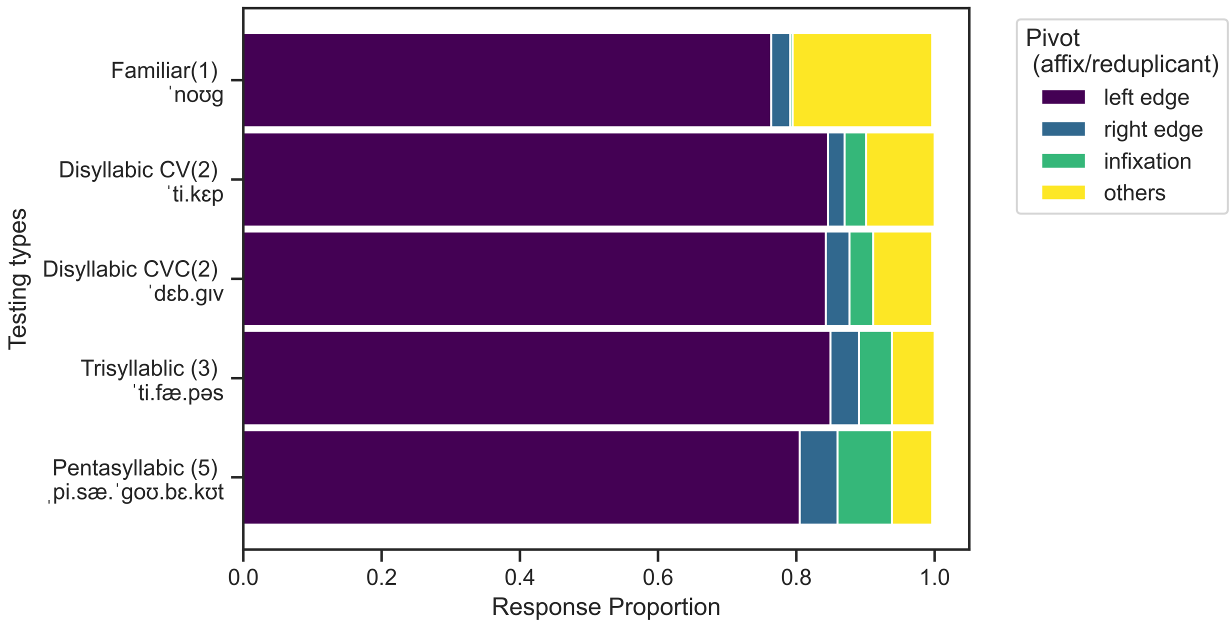


Figure 2.26: *The averaged response proportion of the relative position of the conditioned on each testing type.*

Figure 2.26 showed the averaged response proportion of the edge conditioned on each testing type, with raw numerical values provided in Table. 2.19. Participants predominantly copied the materials from the left word edge. Right edge reduplication and infixing reduplication occurred at some rate, especially infixation for Pentasyllabic ($\sim$ 8%) testing type. Bayesian multinomial logistic regression established the preference for the left edge: within all testing types, its occurrence was significantly more probable than other possible outcomes (all $p < 0.01$). Within all testing types except for Pentasyllabic forms, Others was significantly more probable than Right edge and Infixation (all $p < 0.01$). While for Pentasyllabic forms, Others was only significantly more probable than Right edge ($p < 0.01$) but only marginally higher than Infixation ($p = 0.025 < 0.05$).[25] We suppress the graph of the sampled posterior here. Readers may refer to Appendix B for more information.

[25] As for between testing type comparisons, Disyllabic CVC, Trisyllabic and Pentasyllabic led to more left edge responses than Familiar (all $p < 0.01$). Familiar led to more others responses than all other testing types (all $p < 0.01$). The preference for infixation in Pentasyllabic forms was not found in the population-level estimates.

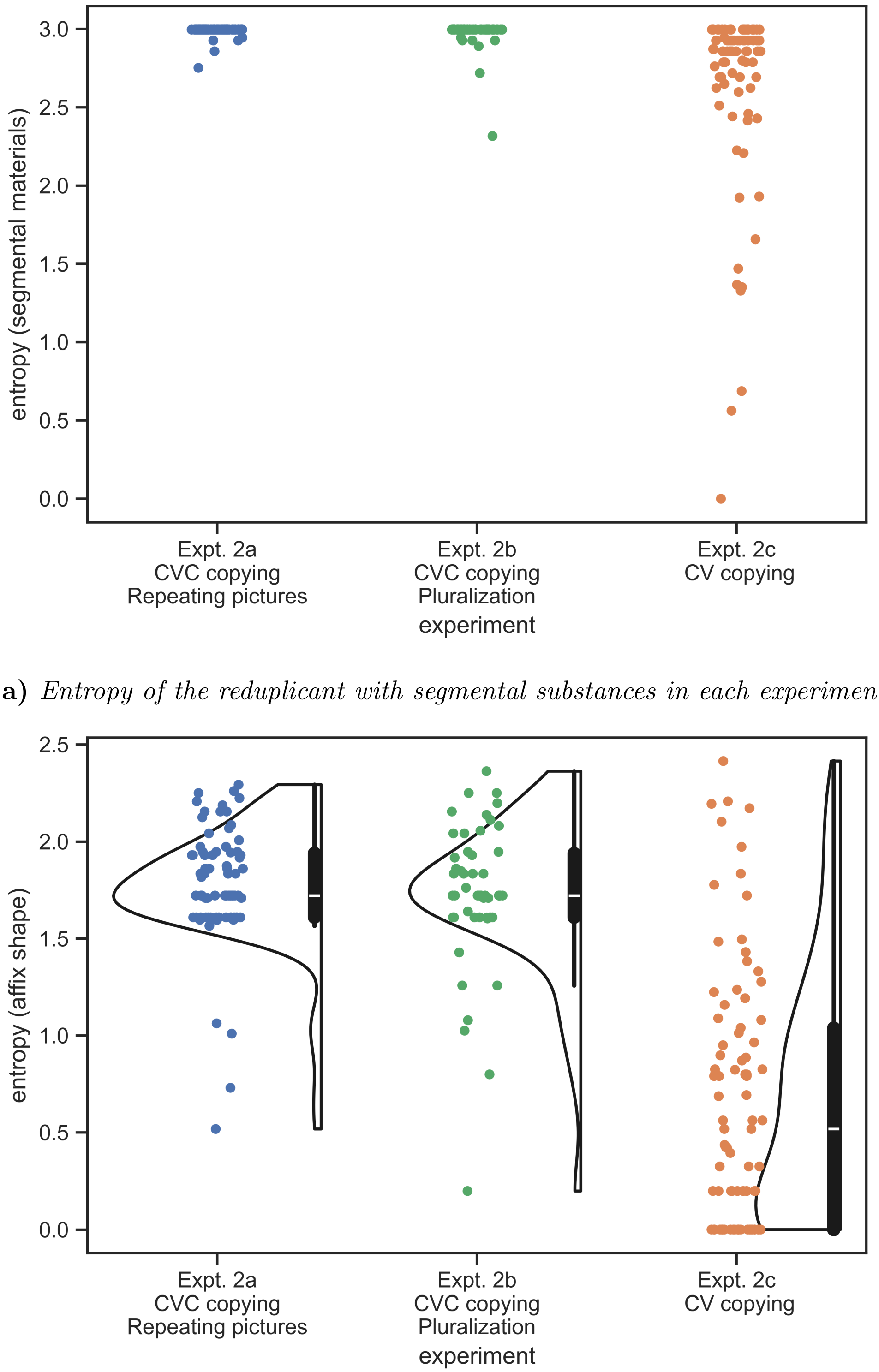
2.5.4.3 Interim summary

Different from Expt. 2a and Expt. 2b, the preferred generalization in Expt. 2c was a finite-state kind of computation, namely to copy a fixed light syllable. The emerging minority patterns supported generalizations characterizable at a higher level of phonological abstractions, here, mainly foot copying ([ˌpæ.kə-ˈpæ.kə.væf]). Apart from the reduplicant shape, we also investigated the relative position of the affix and the base and found a preference for copying pivoted to the left edge.

One might wonder whether in the averaged frequency-based analyses, Pentasyllabic seemed to associate with more Two syllable responses and more Infixation responses, which did not appear to be robustly attested based on the population-level parameter estimates. We think this is due to the idiosyncrasy of the individual grammars. In the next section, we study the spontaneous minority patterns and individual grammars of this experiment series in detail. We find that the spontaneous minority patterns were not random, nor arbitrary, but reflected some systematic principles of the phonological grammar. More broadly, they greatly reflected how reduplicative patterns vary across the world languages, which might provide a learning-based perspective on the typological variations of this morphophonological process.

2.5.5 The emergent variations in individual grammars

The previous sections discuss the preferred generalization of the reduplicant shape universally held at the population level. In Expt. 2a and 2b, with bounded and limited inputs, participants preferred unbounded copying over size-restricting hypotheses. However, participants in Expt. 2c showed a different learning outcome: they preferred a fixed light syllable as the reduplicant shape. Despite these apparent differences, these two observations are unifiable at a deeper level: participants generalized in a manner that is sensitive to prosodically defined templates, such as copying the full prosodic word in Expt. 2a and 2b, copying a light syllable in Expt. 2c. In this section, we study the individual grammars and find these variant spontaneous patterns mirror known regularities put forth in the literature of



(a) *Entropy of the reduplicant with segmental substances in each experiment*

(b) *Entropy of the reduplicant shape characterized based on syllabicity in each experiment*

Figure 2.27: *Entropy of participants' responses in the second series of experiments*

Characterization of the pattern	*# of participants*	*Singular*	*Reduplicated*
Word-final **Heavy syllable copying**	2	ˈtoʊk ˈzi.vɪb ˈtɛf.kʊp ˈgɑ.və.dus ˌpi.sæ.ˈgou.bæ.kʊt	ˈtoʊk-ˈtoʊk ˈzi.vɪb-vɪb ˈtɛf.kʊp-kʊp ˈgɑ.və.dus-dus ˌpi.sæ.ˈgou.bæ.kʊt-kʊt
Total copy up to disyllabic forms; then **word-final** **heavy syllable** copying	1	ˈtoʊk ˈzi.vɪb ˈtɛf.kʊp ˈgɑ.və.dus ˌpi.sæ.ˈgou.bæ.kʊt	ˈtoʊk-ˈtoʊk ˈzi.vɪb-ˈzi.vɪb ˈtɛf.kʊp-ˈtɛf.kʊp ˈgɑ.və.dus-dus ˌpi.sæ.ˈgou.bæ.kʊt-kʊt
Total copy up to Disyllabic forms; then **English suffixation**	1	ˈtoʊk ˈzi.vɪb ˈtɛf.kʊp ˈgɑ.və.dus ˌpi.sæ.ˈgou.bæ.kʊt	ˈtoʊk-ˈtoʊk ˈzi.vɪb-ˈzi.vɪb ˈtɛf.kʊp-ˈtɛf.kʊp ˈgɑ.və.dus-ɪs ˌpi.sæ.ˈgou.bæ.kʊt-s
Total copy up to Trisyllabic forms; then variable total copy and **English suffixation**	1	ˈtoʊk ˈzi.vɪb ˈtɛf.kʊp ˈgɑ.və.dus ˌpi.sæ.ˈgou.bæ.kʊt	ˈtoʊk-ˈtoʊk ˈzi.vɪb-ˈzi.vɪb ˈtɛf.kʊp-ˈtɛf.kʊp ˈgɑ.və.dus-ˈgɑ.və.dus ˌpi.sæ.ˈgou.bæ.kʊt-s (1) ˌpi.sæ.ˈgou.bæ.kʊt-ˌpi.sæ.ˈgou.bæ.kʊt (2)
Total copy up to trisyllabic forms then mainly **three-syllable** copying	1	ˈtoʊk ˈzi.vɪb ˈtɛf.kʊp ˈgɑ.və.dus ˌpi.sæ.ˈgou.bæ.kʊt	ˈtoʊk-ˈtoʊk ˈzi.vɪb-ˈzi.vɪb ˈtɛf.kʊp-ˈtɛf.kʊp ˈgɑ.və.dus-ˈgɑ.və.dus ˌpi.sæ.ˈgou.bæ.kʊt-ˈgou.bæ.kʊt (2) ˌpi.sæ.ˈgou.bæ.-kʊ-kʊt (1) ˌpi.sæ.ˈgou.bæ.kʊt-s (1)

Table 2.20: *The other patterns beyond total reduplication found in Expt. 2a. We keep stems as the same for all forms just for clarity. The number in the parathesis indicates the number of responses for this type of answer. No number specification means the rule had categorically applied.*

typological work on reduplication (Inkelas and Downing, 2015). The varying dimensions are (a). the reduplicant shape, (b). the copied materials, and (c). the base-reduplicant faithfulness (BR-faith). Here, we present these representative patterns and discuss their broader implications for the phonological theory.

Other patterns in Expt. 2a Eleven participants in Expt. 2a showed generalizations other than unbounded copying and we provide their responses in Table 2.20 and Table 2.21.

Based on these patterns, we make two remarks. First, the most frequent use of other possible generalizations was to copy a heavy syllable locally ([ˈzi.vɪb-vɪb]), indicating that some participants had indeed learned the partial reduplication grammar abstracted at a level

Characterization of the pattern	*# of participants*	*Singular*	*Reduplicated*
Total copy up to disyllabic forms; then variable responses along total copy and **word-final heavy syllable copying**	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp
		ˈgɑ.və.dus	ˈgɑ.və.dus-ˈgɑ.və.dus (1)
			ˈgɑ.və.dus-dus (3)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-ˌpi.sæ.ˈgou.bæ.kʊt (1)
			ˌpi.sæ.ˈgou.bæ.kʊt-kʊt (3)
Total copy up to monosyllabic forms; then **templatic backcopying**	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp (2)
			ˈtɛf.kʊp-kʊp (2)
		ˈgɑ.və.dus	ˈgɑ.və.dus-ˈgɑ.və.dus (3)
			ˈgɑv-ˈgɑv (1)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-ˌpi.sæ.ˈgou.bæ.kʊt (1)
			ˈpis-ˈpis (3)
Variable responses Interesting **base truncation**	1	ˈtoʊk	ˈtoʊk-ˈtoʊk (3)
			ˈtoʊk-ˈtoʊ (1)
		ˈzi.vɪb	ˈzi.və-ˈzi.vɪb (2)
			ˈzi.vɪb-ˈzi.vɪb (2)
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp (3)
		ˈtɛf.kʊp	ˈtɛf.kʊ-ˈtɛf.kʊ (1)
		ˈgɑ.və.dus	ˈgɑ.və.dus-eɪs (2)
			ˈgɑv-ˈgɑv (1)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-s (1)
			ˈpi.sæ-ˈpi.sæ (1)
			ˈpi.sæ.goʊ (2)
Variable responses **Foot/Word based copying**	1	ˈtoʊk	ˈtoʊk-ˈtoʊk (3)
			ˈtoʊ-ˈtoʊ (1)
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp (3)
			ˈtɛf.kʊp-kʊp (1)
		ˈgɑ.və.dus	ˈgɑ.və-ˈgɑ.və.dus (2)
			ˈgɑ.və.dus-dus (2)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.-ˌpi.sæ.ˈgou.bæ.kʊt (3)
			ˌpi.sæ.ˈgou.bæ.kʊt-kʊt (1)
Variable responses Total reduplication **English suffixation**	1	ˈtoʊk	ˈtoʊk-ˈtoʊk (3)
			ˈtoʊk-is (1)
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb (3)
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp (2)
			ˈtɛf.kʊp-s (2)
		ˈgɑ.və.dus	ˈgɑ.və.dus-ˈgɑ.və.dus (2)
			ˈgɑ.və.dus-ɪs (1)
		ˌpi.sæ.ˈgou.bæ.kʊt	pi.sæ.ˈgou.bæ.kʊt-ˌpi.sæ.ˈgou.bæ.kʊt (2)
			ˌpi.sæ.ˈgou.bæ.kʊt-s (1)

Table 2.21: *[Continued...]The other patterns beyond total reduplication found in Expt. 2a. We keep stems as the same for all forms just for clarity. The number in the parathesis indicates the number of trials. No number specification means the rule had categorically applied.*

that is smaller than the full word. In most of these responses, the heavy syllables were the word-final heavy syllables. That it is the word-final heavy syllable that was frequently copied is consistent with our previous discussion on a possible preference to copy a legal constituent in the base. Another possible explanation lies in the familiarized reduplicated forms (e.g., [ˈpif.pif]): stress was placed on the initial copy, and participants might have used this stress cue and parsed the final copy as the reduplicant (that is [ˈpif]$_{\text{base}}$.[pif]$_{\text{red}}$).

Secondly, despite copying the full words for shorter stems, one participant in Expt. 1a (the second participant in Table 2.21) copied the initial CVC, and truncated the base — this was sporadically found for trisyllabic stems (1/4; [ˈdi.zɪ.gɛb] ↦ [ˈdiz.ˈdiz]) but was frequently applied to pentasyllabic forms (3/4; [ˌpi.sæ.ˈgoʊ.bɛ.kʊt] ↦ [ˈpis.ˈpis]). This participant may well have tacitly implemented an instance of *templatic backcopying*. As discussed in Section 1.3, templatic backcopying had been taken by some scholars as a conundrum for the Base-Reduplicant correspondence theory (McCarthy and Prince, 1995) because this pattern is thought to not occur in natural languages yet BRCT predicts its existence (e.g., McCarthy and Prince, 1995; Spaelti, 1997; McCarthy et al., 2012). Our finding backs up the studies of Downing (2000), Caballero (2006) and Gouskova (2007), who offer real-language examples from Hausa (Chadic), Guarijio (Uto-Aztecan) and Tonkawa (Coahuiltecan, extinct) respectively, in supporting the conclusion that templatic back-copying is real. That participants spontaneously offered such a pattern indicates that it should be included in the grammatical space as a possible hypothesis.

Other patterns in Expt. 2b. Eleven participants in Expt. 2b provided patterns other than unbounded copying and one participant mainly adopted word-final heavy syllable copying as shown in Expt. 2a, hence we only provide data from ten participants in Table 2.22 and Table 2.23. Beyond heavy syllable copying as discussed above, we also observed frequent foot copying, suggesting that participants have extrapolated hypotheses that can copy more than one syllable even when they had only seen evidence showing monosyllabic CVC copying.

Characterization of pattern	*# of participants*	*Singular*	*Reduplicated*
Total copy up to Trisyllabic forms then **three-syllable copying**	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp
		ˈgɑ.və.dus	ˈgɑ.və.dus-ˈgɑ.və.dus
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-ˈgou.bæ.kʊt
Total copy up to monosyllabic forms; then variable responses of total copy and **suffixing word-final heavy syllable** copying	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb (2)
			ˈzi.vɪb-vɪb (2)
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp (1)
			ˈtɛf.kʊp-kʊp (3)
		ˈgɑ.və.dus	ˈgɑ.və.dus-ˈgɑ.və.dus (1)
			ˈgɑ.və.dus-us (1)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-kʊt
Total copy up to monosyllabic forms; then variable responses with **prefixing and suffixing** variable shapes	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb (2)
			zi-ˈzi.vɪb (2)
		ˈtɛf.kʊp	ˈtɛf.kʊp-kʊp
		ˈgɑ.və.dus	ˈgɑ.və.dus-dus (2)
			gɑv-ˈgɑ.və.dus (1)
			gɑvə-ˈgɑ.və.dus (1)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-kʊt (2)
			pi.sæ.-ˌpi.sæ.ˈgou.bæ.kʊt (1)
			pi.sæ-sæ (1)
Total copy up to monosyllabic forms; **word-final heavy syllable/foot copying** less copying when words grow longer	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb (1)
			ˈzi.vɪb-vɪb (2)
			ˈzi.vɪb (1)
		ˈtɛf.kʊp	ˈtɛf.kʊp-kʊp (3)
			ˈtɛf.kʊp (1)
		ˈgɑ.və.dus	ˈgɑ.və.dus-dus (2)
			ˈgɑ.və.dus-və.dus (1)
			ˈgɑ.və.dus (1)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-kʊt (1)
			ˌpi.sæ.ˈgou.bæ.kʊt-s (2)
			ˌpi.sæ.ˈgou.bæ.kʊt (1)
Total copy up to Disyllabic forms; then **English suffixation**	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp
		ˈgɑ.və.dus	ˈgɑ.vəd-ˈgɑ.vəd (1)
			ˈgɑ.və.dus-ɪs (3)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-s

Table 2.22: *The other patterns beyond total reduplication found in Expt. 2b. We keep stems as the same for all forms just for clarity. The number in the parathesis indicates the number of trials. No number specification means the rule had categorically applied.*

Characterization of pattern	*# of participants*	*Singular*	*Reduplicated*
Onsetless **rime copying**	1	ˈtoʊk	ˈtoʊk-oʊk
		ˈzi.vɪb	ˈzi.vɪb-vɪb (2)
			ˈzi.vɪb-ɪb (2)
		ˈtɛf.kʊp	ˈtɛf.kʊp-kʊp (1)
			ˈtɛf.kʊp-ʊp (3)
		ˈɡɑ.və.dus	ˈga.və.dus-ˈə.dus (1)
			ˈga.və.dus-us (2)
			ˈga.və-du.-dus (1)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-ʊt
Total copy up to disyllabic forms; then **word-final heavy syllable** copying	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp (3)
			ˈtɛf.kʊp-kʊp (1)
		ˈɡɑ.və.dus	ˈɡɑ.və.dus-ˈɡɑ.və.dus (3)
			ˈɡɑ.və.dus-dus (1)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-ˌpi.sæ.ˈgou.bæ.kʊt (1)
			ˌpi.sæ.ˈgou.bæ.kʊt-kʊt (3)
Total copy for shorter words fixed **word-final heavy syllable** copying	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp (2)
			ˈtɛf.kʊp-kʊp (2)
		ˈɡɑ.və.dus	ˈɡɑ.və.dus-ˈɡɑ.və.dus
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-kʊt (3)
			ˌpi.sæ.ˈgou.bæ.kʊt-bæ.kʊt (1)
Total copying up to Disyllabic forms then variable **foot/total copying**	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp
		ˈɡɑ.və.dus	ˈɡɑ.və.dus-ˈɡɑ.və.dus (1)
			ˈɡɑ.və-ˈɡɑ.və.dus (2)
			ˈɡɑ-ˈɡɑ.və.dus (1)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ.ˈgou.bæ.kʊt-pi.sæ.ˈgou.bæ.kʊt (1)
			ˌpi.sæ-pi.sæ.ˈgou.bæ.kʊt (3)
Total copying up to Disyllabic forms then variable **foot/word-final heavy syllable** copying	1	ˈtoʊk	ˈtoʊk-ˈtoʊk
		ˈzi.vɪb	ˈzi.vɪb-ˈzi.vɪb
		ˈtɛf.kʊp	ˈtɛf.kʊp-ˈtɛf.kʊp (3)
			ˈtɛf.kʊp-kʊp (1)
		ˈɡɑ.və.dus	ˈɡɑ.və.dus-ˈɡɑ.və.dus (2)
			ˈɡɑ.və-ˈɡɑ.və.dus (1)
			ˈɡɑ.və.dus-dus (1)
		ˌpi.sæ.ˈgou.bæ.kʊt	ˌpi.sæ-pi.sæ.ˈgou.bæ.kʊt (3)
			ˌpi.sæ.ˈgou.bæ.kʊt-kʊt (1)

Table 2.23: *[Continued...]The other patterns beyond total reduplication found in Expt. 2b. We keep stems as the same for all forms just for clarity. The number in the parathesis indicates the number of trials. No number specification means the rule had categorically applied.*

Non-copying behavior in Expt. 2c In Expt. 2c, we have excluded twelve people because they failed to apply copying at an acceptable level for novel singulars. We studied their responses and found four participants used a listed allomorphy account: they affixed one (or two) of the familiarized reduplicants [-zu/doʊ-/pi-/dɑ-/vi-]. Among these four participants, one participant systematically copied only monosyllabic words, suggesting that they recognized the effect of copying, but only applied it to monosyllabic forms. The other eight participants used English suffixation, or innovated affixation (dɑ-/gɑ-), that did not appear in their input. There was no copying effect for monosyllabic forms.

Bisyllabic trochaic foot copying in Expt. 2c We found Expt. 2c had ten participants showing clear evidence of copying a bisyllabic foot structure: they did so at least three times for pentasyllabic words (e.g. [ˌtɑ.ki-ˌtɑ.ki.ˈzeɪ.bə.zʌp]).[26] We also found one participant copied three syllables for all pentasyllabic items. As repeatedly discussed above, participants were able to perform phonological abstractions at a level higher than a syllable and these phonological abstractions can be described by true prosodically defined units (here, a trochaic foot), hence supporting the Prosodic Morphology. One of the participants offered one trial of templatic back-copying: [ˌgu.zi.ˈtoʊ.fɑ.vɛd] ↦ [ˈgu.zi-ˈgu.zi], adding further support to the psychological reality of templatic backcopying.

A closer look at the learned unbounded copying grammars for Expt. 2c We noticed that some participants may have extrapolated to unbounded copying grammar, and hence were curious to see what they had learned. Based on the criterion of producing at least one response annotated as Total, we have identified eleven participants who systematically learned an unbounded copying grammar. [27]

One participant had learned to apply total reduplication to three syllables and then copied the last three syllables for the Pentasyllabic forms [ˌtɑ.ki.ˈseɪ.və.zʌp-ˈseɪ.və.zʌp]. Likewise, two participants applied total reduplication to disyllabic forms ([ˈteɪ.pus-ˈteɪ.pus]), and adopted bisyllabic foot copying for words with three syllables ([ˌku.sə-ˈku.sɪ.tʊp]) and words with five syllables ([ˌti.pə-ˌti.pæ.ˈgɑ.fu.ˌsɪk]. They also entertained with longer copies for one or two trisyllabic stems ([ˈbæ.də.gɑ-ˈbæ.də.gɑs]),

[26] Eight of them learned opted for a rather categorical bisyllabic copying account.

[27] Two other participants entertained total reduplication for one or two trisyllabic stems.

Two participants systematically dropped the word-final coda for all trials ([ˌgu.zi.ˈtoʊ.fə.vɛ-ˌgu.zi.ˈtoʊ.fə.vɛd]), supporting the total reduplication plus NoFinalCoda grammar. Four participants had learned a total reduplication grammar and such a result might not be attributed to the input. We checked their repetitions of the familiarized items. Two participants repeated the CV partial reduplication as CV total reduplication once (1/4; [ˈzu-zu]). One participant mistakenly repeated one trial as CV total reduplication and one trial with CVC total reduplication (1/4; [ˈʧɑf-ʧɑf]). The other participant repeated all familiarization trials as expected. Thus, their input either showed no preference for the total reduplication analysis or simply defied the total reduplication analysis. The fact that they still extrapolated to total reduplication tends to support a general preference for total copying, which aligns with the preference for unbounded copying from Expt. 2a and Expt. 2b. On the whole, they converge with the typological generalization that total reduplication is more frequent than partial ones; as well as the hypothesis that most languages with partial reduplication also use total reduplication. Responses from these five participants indicate that we should include unbounded copying in the hypothesis space of a learner.

Infixing reduplication in Expt. 2c We found fourteen participants in Expt 2c provided infixing reduplication to a varying degree. Seven of them produced infixation for only one/two trials, which are provided as in Table 2.24.

Participant	*Stem*	*Reduplicated form*
1	[ˈsɑ.pi.ˌdɑf]	[sə-ˈpɑ-pi.ˌdɑf]
	[ˈbuz.dɛf]	[bu.ˈz uz.dɛf]
2	[ˌgɑ.zu.ˈfæ.bi.dɛp]	[ˌgɑ.zu.-ˌfæb.-ˈfæ.bi.dɛp]
	[ˌkɪ.pə.ˈzu.də.ˌtɛf]	[ˌkɪ.pə.-ˌzu.də-ˈzu.də.tɛf]
3	[ˈdɪf.gɑz]	[dɪv-ˈgʌv.gʌz][28]
	[ˈgɑ.bə.dus]	[ˌgɑ.bə.-ˈdu-dus]
4	[ˈgeɪ.bɛs]	[ˈgeɪ-gə-bɛs]
5	[ˈtæf.kʊp]	[ˈtæf-tə-ˌkʊp]
6	[ˈkeɪ.ˌbʌf]	[ˈkeɪ-kə-ˌbʌf]
7	[dɛ.ˈvɑ.gəs]	[dɛ.ˈvɑ-gæ-gəs]

Table 2.24: *Variable infixing reduplicative responses provided by participants*

[28] On a similar note, one of the trials by another participant is [ˈtoʊf.kʌs] ↦ [ˈtoʊf-kʌf-kʌs].

Seven participants consistently infixed reduplicants into the base. Responses from four participants were stress-driven, as they copied the primarily stressed syllable/foot locally for PENTASYLLABIC forms (e.g. [ˌtoʊ.fɑ-ˈveɪ.gə-ˈveɪ.gə.sɪk], [ˌdɛ.və.-gu-ˈgu.pə.zʌb]). Responses from the other three participants seemed to be driven by the right edge: they copied a light syllable at the word-final heavy syllable and incorporated it into the base (e.g., [ˈkeɪ-pɛ-pɛt]). These results are consistent with the predictions of the Pivot theory of infixation (Yu, 2003, 2007), which suggests that the infixes should attach to psychologically/phonologically salient positions, including word edges and stressed syllables.

The emergence of the unmarked in Expt. 2c Participants in Expt. 2c did not always copy faithfully but allowed unmarked structures to emerge. Twelve participants who copied a light syllable consistently reduced copied vowels to a [ə] over 50% of trials ([də-ˈduf]) and eight more participants showed reduction over 25% of the trials – such reduction is absent in the input. We hypothesize the motivation for such a reduction is the phonotactics of English, where stressless vowels are often reduced to schwa (Chomsky and Halle, 1968).

Regardless of the underlying motivation, we still need to explain how the independently motivated phonotactics could "override" the BR-faith enforced in the input. One can argue that this is due to frequency effects: four pairs of input were insufficient to establish a strong BR-faith-over-markedness grammar. We do not deny such a possibility, though it appears to fail to predict the differences between Expt. 2a+Expt. 2b (monosyllabic CVC reduplication; [pif], [pifpif]) and Expt. 2c (monosyllabic CV reduplication; [pif], [pipif]). This further fails to predict the difference between Expt. 1a ([dɔv.gə], [dɔv.dɔv.gə]) and Expt. 1b + Expt. 1c ([dɔv.gə], [div.dɔv.gə]/[dəv.dɔv.gə]).

Another conjecture that we think could also be possible is a "slippery slope" of faithfulness preservation in reduplication learning.[29] Specifically, if the BR-faith (here, MAX-BR violated by segment deletion) is ever observed to be violated, it could be violated to a further degree as long as the minimal structures are maintained. On the other hand, there is an

[29] We thank Kie Zuraw for suggesting this direction, and further to Sam Zukoff and Bruce Hayes for their helpful discussions.

independent learning bias that guides the learner to seek out an abstract analysis at a high level of abstraction once the BR-faith conditions are met, leading to the scenario that "the rich get richer, and the poor get poorer". This account could also explain the findings of experiment series 1. Recall that in Experiment Series 1, when we observed BR-faith violation (here, IDENT-F violations) due to vowel quality differences, especially when participants were familiarized with a fixed [ə], as in [ˈdɔv.gə]∼[dəv-ˈdɔv.gə], we further observed that the onset and the coda underwent some amount of reduction, such as cluster simplification as in [ˈstæb.gə]∼[səb-ˈstæb.gə] and coda non-incorporation as in [ˈstæb.gə]∼[stə-ˈstæb.gə].[30] To a certain extent, we think this may also explain why cross-linguistically, total reduplication does not exhibit active interactions with segmental phonology, albeit interactions with stress/tonal patterns, while segmental phonological patterns, such as reduction, are frequently found in partial copies (Zimmermann, 2021b).

If this hypothesis holds, we may have intriguing predictions for the evolution of reduplicative patterns. To see this, let us imagine ourselves as a learner in Expt. 2c, when we are expected to learn a partial reduplicative pattern with perfect identity reduplication ([ˈpi-pif]), but end up learning a reduplicative pattern with reduction ([pə-ˈpif]). Then, imagine the next generation of reduplication learners, when they get the reduced input, they may reduce it to a further degree, as participants in Expt. 1b who simplified the onset when trained with fixed [ə]. Hence, we will only expect partial reduplication to be more and more reduced over generations, which may come to an end when a generation of learners is unable (or, needs a lot of data) to recognize the copying effect and learn the intended reduplicative pattern as simple allomorphy or some other morphophonological regularities. We are not sure of to what degree this hypothesis holds, thus, noncommittal to this point of view. Future research should study the plausibility of such a hypothesis, and separate this hypothesis from other possible accounts.

[30] The coda non-incorporation might have some other possible explanations, such as perceptibility: whether the coda was perceived following a [ə].

2.5.6 Summary of findings in Experiment Series 2

Rapid generalization of unbounded copying from bounded input As a summary, recall the familiarized patterns of each experiment in this experiment series, as in (26).

(26) A summary of Experiment Series 2

Experiment	*Examples*	*Singular*	*Reduplicant*	*Base*	*Preferred generalization*
2a + 2b	[ˈpif] → [ˈpifpif]	$C_1V_2C_3$	$C_1V_2C_3$	$C_1V_2C_3$	total reduplication
2c	[ˈpif] → [ˈpipif]	$C_1V_2C_3$	C_1V_2	$C_1V_2C_3$	prefixing reduplication light syllable

In this experiment series, we studied how participants generalized copying-based patterns with limited, bounded input. Across the three experiments, participants were instructed to learn pluralization in an artificial language from a very small number (4) of auditorily presented singular-plural pairs. In the training, the singulars always had the shape CVC ([ˈpif]) while the reduplicated form copied CVC ([ˈpif-pif]). The training trials were highly ambiguous, compatible with both a generalization of total reduplication, i.e., copying the full word, and a generalization of imposing a size restriction by copying just CVC. To tease apart the many possible generalizations, the testing trials consisted of longer forms with multiple syllables ([ˈteɪ.pə.gæb]). Results suggest participants rapidly extracted reduplicative rules and predominately extended total copying, but not the size-restricting generalization, to longer words ([ˈteɪ.pə.gæb.ˈteɪ.pə.gæb]).

In Expt. 2c, during which participants were trained with a CV-shaped copy ([pif] ∼ [ˈpi-pif]), participants predominantly preferred to copy the leftmost light syllable ([teɪ.ˈteɪ.pə.gæb]). There was greater between-participant variation, largely reflecting typological variation. The dimensions of variation include which portion to copy ([teɪ.pə.gæ.gæb] versus [teɪ.teɪ.pə.gæb]), and how faithfully to copy ([tə.ˈteɪ.pə.gæb]). As for the size of the copy, the fact that some participants produced foot-based copying as [ˌteɪ.pə.ˈteɪ.pə.gæb] invalidates the bias favoring segment-based length restriction. These results suggest that unbounded copying, a supra-finite-state computation, should be included in an adequate hypothesis space of (morpho)phonology, an issue that will be directly addressed in Chapter 3.

We would like to bring up a qualitative observation and a follow-up speculation. Notice that between experiments, there might be some differences in the degree to which copying is learned. We did not exclude any participants from Expt. 2a + Expt. 2b (n = 129) for failing to copy but we identified 12 participants in Expt. 2c (n = 93) failing to copy for most trials (see discussion on non-copying behavior above). The familiarized patterns were minimally different: it is just a missing coda in Expt. 2c. This leads us to two questions. First, whether the difference in non-copying generalizations reflects a relative learning difficulty associated with [ˈpif] ↦ [ˈpi-pif] compared to [ˈpif] ↦ [ˈpif-pif], and if so, why. Could it be that partial reduplication in which the reduplicant fails to match any parsable constituent in the base (as in [ˈpif] ↦ [ˈpi-pif]) is harder to detect/learn compared to patterns that do match parsable constituent in the base (as in $\boxed{\text{pif}}$-$\boxed{\text{pif}}$)? Experiment series 1 seems to corroborate this hypothesis: [ˈdɔv.gə] ↦ [$\boxed{\text{dɔv}}$ -ˈ $\boxed{\text{dɔv}}$ .gə] did not lead to any exclusion of participants due to the failure to copy, but when familiarized with [ˈdɔv.gə] ↦ [div-ˈdɔv.gə]/[dəv-ˈdɔv.gə], we see some more exclusions based on the degree of copying in participants responses. This question may be informative in spelling out the exact algorithmic steps involved in how humans recognize and learn reduplication, which we leave as an open area for future research.

Support for Prosodic Morphology and further implications for the phonological theory We found the participants were guided by the theoretically grounded principles in creating their novel responses to the unseen stems. Our main results are in line with various theoretical proposals, such as **Prosodic Morphology** (McCarthy and Prince, 1986), **Base-Reduplicant correspondence theory** (McCarthy and Prince, 1995), **the Emergence of the Unmarked** (McCarthy and Prince, 1995), and the **Pivots account** for infixation (Yu, 2003, 2007). For more detailed discussions on how each theory is individually supported, see Section 2.5.5 above.

Let us focus on these results for Prosodic Morphology. In Experiment 2, we found participants generalize in a way that is sensitive to prosodically defined templates, which is strong evidence for prosodic morphology and consistent with results from Experiment Series

1 of generalizing at the level of syllables. Here, we expanded on the existing conclusion by observing that participants were capable of extrapolating to grammars anchored at all levels of the Prosodic Hierarchy. This includes (a). the preferred light syllable copying generalization in Expt. 2c ([teɪ-teɪ.pə.gæb]), (b). the word-final heavy syllable in Expt. 2a+Expt. 2b ([teɪ.pə.gæb.gæb]), (c). copying based on a trochaic foot ([ˌtɑ.ki-ˌtɑ.ki.ˈzeɪ.bə.zʌp]), and (e). copying at the level of a prosodic word, including the preferred total reduplication in Expt. 2a+Expt. 2b, as well as the NoFinalCoda grammar as in [ˌgu.zi.ˈtoʊ.fə.vɛ-ˌgu.zi.ˈtoʊ.fə.vɛd], supporting the psychological reality of each level. Each pattern shows varying degrees of preference depending on the different familiarization patterns.

From our results, we hope to demonstrate that experimental investigations are not merely a backup for our theories; they also provide informative empirical evidence for theory evaluation. For example, templatic back-copying ([ˌgu.zi.ˈtoʊ.fɑ.vɛd] ∼ [ˈgu.zi-ˈgu.zi]) was believed to never occur across the world's languages. Excluding such a pattern was considered one evaluation criterion for a better theory of reduplication (e.g., McCarthy and Prince, 1999; Kiparsky, 2010). However, this pattern did appear in participants' responses, though very infrequently provided. This suggests that theories such as Base-Reduplicant Correspondence Theory (McCarthy and Prince, 1995) and Morphological Doubling Theory (Inkelas and Zoll, 2005), among many others, should not be dismissed as inadequate because they generate templatic backcopying. Instead, the rarity of templatic backcopying might be attributed to learning biases. This further suggests that though we should continue to study the cross-linguistic unattestedness or rarity of certain linguistic structures and explore possible explanations for their potential absence, we might need to be more cautious in concluding the absolute non-existence of a pattern within the possible grammatical space.[31]

Universals and variations of learning biases reflect typological universals and variations. Our experiments asked for free-production responses. Most of the spontaneous responses offered by the participants were not arbitrary but appeared to reflect empirical

[31] In Chapter 3, I will argue that nesting dependencies should be excluded for morphophonology from typological evidence as well as experimental evidence from learning and recognition.

phenomena found in the typology of reduplication patterns. The systematicity in participants' responses, on the one hand, supports the ecological validity of our experimental design, and the possibility of the poverty of stimulus design as a sampler of the learner's hypothesis space conditioned on some training data, at least for reduplication learning.

Beyond this methodological remark, so far, we have identified two analytic biases for reduplication learning through this series of experiments. From Expt. 2a and Expt. 2b, we have identified a propensity to interpret patterns of total copy that are seen in relatively short strings as representative of unbounded copying of strings of *any* length. We will term this as an *unboundedness* bias. Such an unboundedness bias aligns well with the typological generalization that total reduplication is more frequent than partial reduplication. From Expt. 2c, we have identified a propensity to fix the edge and the reduplicant shape for patterns that are seen in relatively short strings as representative of a fixed-edge, fixed-shape reduplicative pattern for longer forms. This preference for reduplicant shape can be unified with results from Experiment Series 1, suggesting a general bias for a *fixed* prosodic template in partial reduplication. Regarding the question of which part of the stem is copied, we found that learners are biased towards making reference to the *word edge*, instead of the primarily stressed syllables, although the latter indeed occurred at a non-trivial rate. This aligns with the typological generalization that most reduplication patterns are found to be prefixing (as well as suffixing), but less frequently as infixes (Rubino, 2013). One explanation for these typological asymmetries is the presence of biases in learning. Here, we see that they predict the correct typological trends, possibly shaping typology.

Another piece of evidence for this typology-shaped-by-learning account is the emergent variations in individual grammars. We found that the significant between-subject variations in their analyses of the familiarized pattern reflect the typological variations observed in naturally occurring patterns, including foot-based templates (real language example: Diyari, Pama–Nyungan; Austin, 1981), imperfect copying with vowel reduction (real language example: Palauan, Austronesian; Zuraw, 2003), imperfect copying with consonant skipping in onset (real language example: Sanskrit, Indo-European; Steriade, 1988), infixing reduplication (real language example: Samoan, Austronesian; Broselow and McCarthy, 1983), templatic back-copying (real language example: Guarijio, Uto-Aztecan; Caballero, 2006), as well as attributing the missing coda as a separate phonological process instead of as a property of having a fixed template (real language example but with a vowel deletion process: Lardil, Tangkic; McCarthy et al., 2012). Some participants also learned to subcategorize based on the phonological shapes of the stems (see Paster, 2006). For example, they learned to actively copy when stems are relatively short but not copy for longer forms. To sum up, the point we hope to make here is an intuitive but often overlooked one: the great diversity of possible linguistic structures may also have its root in learning, reflected by the great variety of possible analyses towards which *individual* human learners are biased.

2.6 Discussion and future research

We conclude this chapter by first answering the research questions we have asked and then discussing possible directions for future research.

2.6.1 Main questions addressed in this chapter

Below in (27), we present our answers to four major questions discussed at the beginning of this chapter.

(27) a. *Can human learners rapidly learn copying-based generalizations?*

When prompted with reduplication as a morphophonological process, can learners recognize the effects of copying and extend copying-based generalizations to novel forms?

Answer Yes, the learner could *rapidly* recognize the effect of copying and extrapolate to a copying-based generalization when presented with a morphophonological operation. This is consistent with previous studies on the saliency of surface repetitions (*wofefe* versus *wowofe*), suggesting that identity-based structures are not hard to learn.

b. *What inductive biases of phonological abstractions do human learners exhibit?*

If multiple generalizations are equally compatible with the familiarized patterns, based on what levels of phonological abstraction do human learners form their generalizations?

Answer Human learners generalize in a manner that is sensitive to phonological abstractions characterizable by the vocabulary of prosody (e.g., syllables, feet, prosodic words). Under the condition that the identity requirements are satisfied in the input, they can extrapolate to a quite high level of abstraction, ignoring the fine-grained specifications of phonological materials.

c. *Are there learning differences among different attested patterns?*

Do different typologically attested patterns bear different learning results?

Answer The short answer is yes. Within each series of experiments, we aimed to study this question with minimally different familiarized patterns. In Experiment Series 1, we looked at how forcing a finer-grained specification of phonological materials in one location of the copy affects the learning results. We found that participants were more inclined to fix finer-grained phonological details in other positions, showing less higher-level abstraction. In Experiment Series 2, we investigated cases when two familiarized patterns had monosyllabic CVC stems, and differed only in copying a coda or not ([ˈpif-pif] versus [ˈpi-pif]). At the population level, participants extrapolated very different generalizations regarding how much to copy from these two patterns. If familiarized with [ˈpif-pif], they were more inclined to copy unbounded structures. On the contrary, if familiarized with [ˈpi-pif], they were more inclined to learn a fixed light syllable. We also found that CV partial reduplication led to more individual variations in terms of how much to copy, how faithful to copy, and what phonological materials to copy.

d. *Is a learner's hypothesis space of morphophonology limited to finite-state kinds of computation?*

When learners try to learn a morphophonological process and are only provided with bounded input, will they interpret the limited familiarized items as an instance of bounded copying and hence adopt a size-restricting partial reduplication generalization, or unbounded copying and hence adopt a non-size-restricting generalization?

Answer Participants in Experiment Series 2 were able to extrapolate unbounded copying grammars in which there is no restrictions on how much to copy. Unbounded copying is preferred when total reduplication analysis is plausible, as in Expt. 2a + Expt. 2b. An unbounded copying grammar is learnable even when total reduplication is not compatible with the input, as in Expt. 2c. This suggests that unbounded copying should be included in the hypothesis space of a possible human learner for reduplication as a morphophonological process. Thus, a finite-state kind of computation might not be the right sense to single out morphophonological regularities.

2.6.2 Future research

Our experimental results have several implications for the phonological structures beyond segments and morphophonological learning. First, the results of Experiment Series 1 suggest a tight bond between substructures within a copy.[32] We found when a kind of phonological computation targets a specific substructure (here, nucleus), it also affects other substructures (here, onset and coda) within the copy. This implies that the reduplicant functions as a complete sequence rather than as individual segments. One promising way to flesh this idea out is to formalize it within the framework of aggressive reduplication, as proposed by Zuraw (2002) – the theory of aggressive reduplication emphasizes the necessity to correspond phonological *substrings* within a word form, in addition to the segmental correspondences in

[32] One can think of the chemical bond associating chemical elements for a helpful analogy.

the classic base-reduplicant correspondence theory.[33]

Secondly, we hypothesized a plausible learning difficulty associated with the reduplicative patterns that do not match a constituent in the base (e.g., [ˈpi-pif]; [div-ˈdɔv.gə]), compared to the matching ones (e.g., [ˈ$\boxed{\text{pif}}$-$\boxed{\text{pif}}$]; [$\boxed{\text{dɔv}}$-ˈ$\boxed{\text{dɔv}}$.gə]). We think this may shed light on the algorithmic steps in recognizing and learning reduplicative patterns.

Lastly, from both experiments, we observe that participants allowed the unmarked structure to emerge. We propose a conjecture called the "slippery slope" of faithfulness preservation in reduplication learning. Specifically, if the two copies are similar enough but not perfectly identical, the faithfulness between them could be reduced further as long as the minimal structures are maintained. We think such an account can explain some typological generalizations on the difference between total reduplication and partial reduplication, which were often used as arguments to support a bipartite view of reduplication patterns. In this way, they may be unified from a learner's perspective. This account might also make testable predictions for the evolution and change of reduplicative patterns. Future research should study to what extent this hypothesis holds, its formalization, and its consequences.

Through our experiments, we hope to demonstrate how artificial grammar learning experiments with typological and theoretical grounding can contribute to morphophonological learning, and consequently supplement theoretical and typological work with converging (and sometimes conflicting) learning evidence. Future studies should consider more dimensions that potentially reveal other crucial aspects of the grammar and the learner. For example, one can ask whether the same trends hold for suffixing partial reduplication (i.e. when the partial copy is anchored at the right word edge), and/or manipulate the relative positions of two copies. We are working on factoring these other dimensions into future experiments.

Situating our work in a broader context, the bias favoring more abstract prosodic constituents in reduplication is also confirmed in learning another non-concatenative process, namely infixation, as in Wilson (2022). Together, these results suggest that an adequate learner readily for any morphophonological patterns ought to incorporate such a bias. In

[33] See the background (Section 3.2.1.1) of the next chapter for more discussion.

previous studies on morphophonological learning, reduplication and infixation did not receive much attention. Hence, we hope to call for more research on non-canonical (to spell out more, non-edge-oriented non-concatenative) morphophonological processes, and test computational models against empirical discoveries of these patterns.

One question that we hope to address in the future is *how much* the identified learning biases explain the typology of reduplicative patterns. We feel our current data is sufficient for qualitative comparisons but rather preliminary for a more precise understanding. One way to view the poverty of the stimulus paradigm adopted here is to regard it as a sampling of the learner's hypothesis space conditioned on some training data, which itself could be skewed. From the positive results of these two experiment series, we think one direction to pursue further is to collect a large-scale corpus as a benchmark, with carefully designed artificial grammar learning results investigating more varying linguistic dimensions. This might involve collecting evidence from more participants and engaging native speakers of more languages. Such a corpus would allow us to generate quantitative predictions for a better understanding of the relationship between the typology and the learning, and provide another source of evidence to evaluate the phonological theory and computational learning models.

CHAPTER 3

Finite-state buffered machines: A formal framework for recognizing surface repetitions

3.1 Introduction

Chapter 2 discusses how unbounded copying should be included in the grammatical space of morphophonology. Driven by this empirical result, this chapter[1] aims to introduce a formal model of possible natural language word forms which is restrictive enough to rule out many unattested patterns, but still expressive enough to allow for reduplication. Among the well-known existing classes of formal languages, there is a tension between these two goals. The overwhelming majority of attested phonological patterns fall within the finite-state class (Kaplan and Kay, 1994), and perhaps within even more restrictive subclasses (Heinz, 2007). Reduplication is the striking exception to this generalization. But at present, if we look for alternatives to the finite-state characterization which are powerful enough to express reduplication, we only find classes of formal languages which additionally allow a wide variety of unattested patterns — for example, nesting/mirror-image patterns, or arbitrary cross-serial dependency patterns significantly more general than reduplication itself. This gives us no way to retain the finite-state characterization's (apparently correct) prediction that mirror-image patterns and so on will be unattested, while avoiding the (apparently incorrect) prediction that reduplication will be unattested.

[1] A major portion of this chapter was published in *Journal of Language Modelling* in 2023, co-authored with Tim Hunter (Wang and Hunter, 2023). An earlier version was presented at SIGMORPHON 2021, ESSLLI 2021, and AMP 2021, and published as Wang (2021a). The new materials in this chapter include (1) how the formal proposal excludes some logically plausible but typologically unattested partial reduplication patterns (Example 2); (2) more discussion on the linguistic relevance of "mode-determinism" (Section 3.6.1); (3) how the formal model bears the results of prosodic units (Section 3.6.3).

Jäger and Rogers (2012) review other cases where natural language generalizations do not appear to correspond neatly to degrees of complexity as defined by the formalisms of the classical Chomsky Hierarchy, and the "refinements" of the hierarchy that these findings have prompted. In the case of natural language syntax, for example, it is widely accepted that context-free grammars are insufficiently expressive (Huybregts, 1984; Shieber, 1985; Culy, 1985); but the next level up on the classical hierarchy, context-sensitive grammars, are far too expressive to be a plausible characterization of possible natural languages. This situation prompted the development of many *mildly context-sensitive* formalisms (Joshi, 1985; Kallmeyer, 2010), whose generative capacity sits in between the context-free and context-sensitive levels. Another "mismatch" has been observed in phonology, where even the lowest level of the classical hierarchy, the finite-state languages, has been argued to be insufficiently restrictive. To address this, a number of researchers have developed *sub-regular* formalisms (e.g., Heinz et al., 2011; Chandlee, 2014; Heinz, 2018).

In this paper, the situation we are addressing is slightly less straightforward than the two mismatches just mentioned. The development of sub-regular formalisms was a response to a perception that *all* the levels of the classical hierarchy were too powerful. The mildly context-sensitive formalisms address the fact that, with regard to syntax, each of the classical levels is either too weak (finite-state, context-free) or too powerful (context-sensitive, recursively enumerable). The situation we address in this paper, in contrast, is one where the classical context-free class is both too powerful in some ways (since it allows mirror-image patterns) and too restrictive in other ways (since it disallows reduplication). We, therefore, seek a formalism that *cuts across* the levels of the classical hierarchy, rather than one which adds a level that sits within the existing hierarchical relationships.

We introduce *finite-state buffered machines (FSBMs)* as a step towards solving this problem. The idea is to preserve as much as possible of the restrictiveness of the finite-state class and add "just" what is necessary to generate copying patterns. FSBMs include unbounded memory in the form of a first-in-first-out buffer, but the use of this memory is restricted in two important ways. First, this memory buffer uses the alphabet of surface symbols, rather than a separate alphabet like the stack alphabet of a pushdown automaton (PDA). Second, the allowable ways of interacting with this memory buffer are closely tied to the surface string being generated: the only storage operation adds a copy of the current surface symbol to the memory buffer, and the only retrieval operation empties the entire memory buffer and adds its contents to the generated string. For example, in computing a string of the form $urrv$, an FSBM will proceed through three phases corresponding to the sub-strings u, r and v, much like a standard finite-state machine generating the string urv. But throughout the middle phase, a copy of each surface symbol of r will be stored in the FSBM's memory buffer, and at the transition from this middle phase to the third phase the buffer will be emptied and its contents appended to the computed string; thus ur has r appended to it, before the machine proceeds to compute the v portion in the third phase.

In Section 3.2 we discuss the computational challenge posed by reduplication in more detail, and outline the ways our approach differs from a number of other attempts to enrich otherwise restrictive formalisms with copying mechanisms. We present FSBMs in full in Section 3.3, give a pumping lemma in Section 3.4, and explore the mathematical properties of the generated class of languages in Section 3.5. Section 3.6 discusses some remaining issues, including various kinds of non-canonical reduplication, and a formal distinction between what we will call *symbol-oriented* generative mechanisms (such as string-copying) and the better-known mechanisms underlying the classical Chomsky Hierarchy. Section 3.8 concludes the paper.

3.2 Background

Section 3.2.1 outlines the important empirical properties of reduplication that make it a poor fit to the classical Chomsky Hierarchy; in particular, we aim to show that an appropriate characterization of possible natural language word forms should include the pattern ww, for unboundedly many strings w, but not ww^R, where w^R is the reverse of w. Section 3.2.2 reviews various modifications to classical automata, like our proposal, that incorporate some form of unbounded queue-like memory. In Section 3.2.3 we discuss other modifications to finite-state automata that were motivated by reduplication, but do not accommodate the

Total reduplication: Dyirbal plurals (Dixon, 1972, p. 242; Inkelas, 2008, p. 352)			
Singular	*Gloss*	*Plural*	*Gloss*
midi	‘little, small’	midi-midi	‘lots of little ones’
gulgir̮i	‘prettily painted men’	gulgir̮i-gulgir̮i	‘lots of prettily painted men’

Partial reduplication: Agta plurals (Healey, 1960, p.7)			
Singular	*Gloss*	*Plural*	*Gloss*
labáng	‘patch’	lab-labáng	‘patches’
takki	‘leg’	tak-takki	‘legs’

Table 3.1: *Total reduplication:Dyirbal plurals (top); partial reduplication:Agta plurals (bottom)*

crucial property of unboundedness.

3.2.1 The puzzle of reduplication

3.2.1.1 Reduplication in natural languages

As we have repeatedly discussed in the previous chapters, reduplication is common cross-linguistically. As we see in Table 3.1, Dyirbal exhibits *total reduplication*, with the plural form of a nominal comprised of two perfect copies of the full singular stem; whereas *partial reduplication* is exemplified in Agta, where plural forms only copy the first CVC sequence of the corresponding singular forms (Healey, 1960; Marantz, 1982).[2] In the sample reported by Rubino (2013), 313 out of 368 natural languages exhibit productive reduplication, of which 35 languages have total reduplication but not partial reduplication.

By comparison, context-free palindrome patterns are rare in phonology and morphology (Marantz, 1982) and appear to be confined to language games (Bagemihl, 1989; Gil, 1996), whose phonological status is unclear. Figure 3.1 illustrates the important difference between Dyirbal total reduplication (‘*midi-midi*’) and the logically-possible but unattested palindrome pattern (‘*midi-idim*’).

[2] For clarity, we adopt a simplistic analysis here. When the bases start with a vowel, Agta copies the first VC sequence, as in *uffu* ‘thigh’ and *uf-uffu* ‘thighs’. Thus, a more complete generalization is that Agta copies a (C)VC sequence.

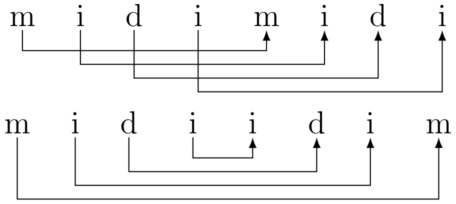


Figure 3.1: *Crossing dependencies in Dyirbal total reduplication 'midi-midi' (top) versus nesting dependencies in unattested string reversal 'midi-idim' (bottom)*

From the perspective of a computational analysis, it will be important to establish that (at least some) reduplication constructions are *unbounded*, in the sense that they are usefully modeled by string-sets of the form $\{ww \mid w \in S\}$ for some infinite set S. A partial reduplication construction, such as the Agta case above where an initial CVC sequence is copied, is obviously not unbounded in this sense, since — assuming a finite alphabet — there are only finitely-many CVC sequences (Chandlee and Heinz, 2012). But as observed by Clark and Yoshinaka (2014) and Chandlee (2017), even amongst total reduplication constructions we must take care to distinguish between unrestricted, productive total reduplication (which is unbounded in the relevant sense) and total reduplication on a *finite* set of bases. For example, it is important to establish that *midi-midi* is not simply part of a collection $\{ww \mid w \in S\}$ where S is some finite memorized set (e.g. the set of all lexemes of a particular category); in such a case, the resulting set of reduplicated forms would itself be finite, and therefore within most familiar language classes. Table 3.2 illustrates the relationship between productivity, the partial/total distinction, and unboundedness.

	Restricted to lexemes (not productive)	Not restricted to lexemes (productive)
Partial Reduplication	**Bounded**	**Bounded**
Total Reduplication	**Bounded**	**Unbounded**

Table 3.2: *Reduplication and bounded/unbounded copying*

A few cases complicate the picture in Table 3.2. If the definition of "partial" reduplication is just that one copy surfaces as a part of the relevant words, then, some attested partial reduplicative patterns could be unbounded. One such example is verb reduplication in Lardil (Tangkic). As we can see in (28), strictly following the definition, this pattern should be categorized as "partial". However, it is unbounded because the size of the partial copy grows together with the verb root. These kinds of partial reduplication are only illusionary if we look at the phonology of the whole language: copies are partial due to the interaction of total reduplication and other phonological phenomena that involve deletion. In the case of Lardil, enough evidence supports that there is a final vowel deletion process that applies to the first copy. [3]

(28) Lardil verb reduplication (McCarthy et al., 2012, p.189)

paɾel-paɾeli	‘to gather’
maɽbaɾ-maɽbaɾa	‘be cramped’
wuʈuwal-wuʈuwala	‘go around’

On a different note, in principle, a reduplicative pattern which copied, for example, *half* of the relevant stem, would be a case of unbounded copying in this sense that would likely nonetheless be described as partial reduplication. But the attested cases of true “partial reduplication” appear to all involve templates that do not directly depend on the length of the base (see the most frequent attested shapes in Moravcsik, 1978; Rubino, 2005; Dolatian and Heinz, 2020, and the emergent reduplicant shapes discussed in Chapter 2), like the Agta examples above.

A case of reduplication that is unbounded in the relevant sense is the Bambara ‘Noun *o* Noun’ construction (Culy, 1985). For example, the stem **wulu** *dog* can be copied to form **wulu o wulu** *whichever dog.* The important point about productivity comes from the interaction of this reduplication with the agentive *la* construction, illustrated in (29) (Culy, 1985, pp.346–347).

(29) a. wulu + nyini + la = wulunyinina
dog search for
“one who searches for dogs”, i.e.,-“dog searcher”

b. wulu + filè + la = wulufilèla
dog watch

[3] In fact, this apocope process applies when a prosodic word contains at least three moras. In Lardil verb reduplication, each copy is in its own prosodic word – hence apocope is expected to apply. The second copy retains its vowel because there is an underlying final /-t̪/ in verb stems, later deleted as non-apical codas. For more discussion, see McCarthy et al. (2012, pp. 189-190)

"one who watches dogs", i.e.,-"dog watcher"

This agentive construction itself is recursive, in the sense that it can build on its own outputs, as illustrated in (30); and the outputs of the agentive construction, including the recursively-formed ones, can be used in the 'Noun *o* Noun' reduplicative construction, as illustrated in (31).

(30) a. wulunyinina + nyini + la = wulunyininanyinina
dog searcher search for
"one who searches for dog searchers"

b. wulunyinina + filè + la = wulunyininafilèla
dog searcher watch
"one who watches dog searchers"

(31) a. wulunyinina **o** wulunyinina
dog searcher dog searcher
(29a) (29a)
"whichever dog searcher"

b. wulufilèla **o** wulufilèla
dog watcher dog watcher
(29b) (29b)
"whichever dog watcher"

c. wulunyininanyinina **o** wulunyininanyinina
(30a) (30a)
"whichever one who searches for dog searchers"

d. wulunyininafilèla **o** wulunyininafilèla
(30b) (30b)
"whichever one who watches dog searchers"

The set of all outputs of this reduplication process can therefore naturally be thought of as taking the form $\{ww \mid w \in S\}$, where S is the *infinite* set of nouns, including outputs of the agentive construction.

Further evidence that reduplication is productive in this sense comes from its applicability to borrowed words: Yuko (2001, p. 68) cites the totally-reduplicated plurals *teknik-teknik* 'techniques' and *teknologi-teknologi* 'technologies' attested in Malay, for example. Similarly, the code-switching data from Tagalog in (32) (Waksler, 1999), shows the English word 'swimming' being (partially) reduplicated.

(32)	Saan	si	Jason?	Nag-SWI-SWIMMING	siya.
	where	DET	Jason	PRESENT-REDUP-SWIMMING	he

'Where is Jason? He's swimming.'

In addition, in a few experiments that, either directly or indirectly, study the learnability of surface identity-based patterns, copying appears to be salient and easy to learn. Marcus et al. (1999) shows that infants can detect and habituate to different identity-based patterns: ABA vs. ABB and AAB vs. ABB, where A and B are CV syllables. Crucially, the particular syllables used at test time were distinct from any seen during training. That copying is salient is also supported by the learning results from the artificial language learning studies discussed in Chapter 2. Here, we briefly review our experiment design and results. In the first set of experiments, participants were instructed to learn pluralization in an artificial language from a small number (4) of auditorily presented singular-plural pairs. In the training, the singulars always had the shape CVC (['pif]) while the reduplicated form copied CVC (['pif-pif]). The training trials were highly ambiguous, compatible with both a generalization of unbounded copying, i.e., copying the full word, and a generalization of imposing a bound on the size of a copy, for example only copying CVC. To tease apart the two possible generalizations, the testing trials consisted of longer forms with multiple syllables (['teɪ.pə.gæb]). Results suggest participants rapidly extracted reduplicative rules and predominately extended total copying, but not the length-restricting generalization, to longer words (['teɪ.pə.gæb.'teɪ.pə.gæb). This supports that the learner is biased towards total reduplication. In the second experiment, during which participants were trained with a CV-shaped copy (['pif] ∼ ['pipif]), there was greater between-participant variation, largely reflecting typological variation. As for the size of the copy, there is a proportion of the responses longer than a CV shape, especially for longer forms, as in ['teɪ.pə.'teɪ.pə.gæb] and ['teɪ.pə.gæ.'teɪ.pə.gæb] – this invalidates the bias for size-based upper bounds. These results provide enough evidence that unbounded copying should be in the grammatical space of morphophonological patterns so that they can be readily learned, and in fact, preferred.

Evidence that reduplication/copying (ww) patterns have an importantly different status than reversal (ww^R) patterns — converging with the typological absence of reversal patterns noted above — comes from one recent artificial grammar learning study (Moreton et al., 2021). In this experiment, adult learners were trained to identify either a reduplication or a syllable reversal pattern. Participants were also asked to explicitly state the rule they had learned (if they could). Participants in the reduplication group showed final above-chance performance whether they could state the rule or not. However, in the syllable-reversal condition, only participants who could also correctly state the rule showed final above-chance performance; this suggests that learning the reversal pattern relied on some degree of explicit/conscious reasoning that the copying pattern did not. In further support of this distinction, correct syllable-reversal responses showed longer reaction times than correct copying responses. In a second variant of this experiment, the training phase was replaced with explicit instruction about the rule to be applied; participants in the reduplication group still showed shorter reaction times. These results suggest that, to the extent that reversal patterns can be learned or applied at all, this is achieved more by conscious application of a rule rather than unconscious linguistic knowledge, in contrast to reduplication.

A significant aspect of this AGL study is that the stimuli used were auditory, "*purely phonological*", "*meaningless*" strings (Moreton et al., 2021, p. 9), chunks of which are identical. We take this to indicate that cognitively representable reduplication or reduplication-like patterns need not be realizations of meaning-changing operations: identity between substrings can contribute to the phonotactic well-formedness of a surface form, in ways that can be separated from any morphological paradigms in which that surface form appears. This aligns with the general tendency that Zuraw (2002) called *aggressive reduplication*: human phonological grammar is sensitive to output forms with self-similar subparts, regardless of morphosyntactic or semantic cues. Such sensitivity is formalized by Zuraw as the constraint REDUP which requires string-to-string correspondence by coupling sub-strings together.

Direct evidence supporting aggressive reduplication comes from pseudo-reduplication. A pseudo-reduplicated word has one portion identical to another portion. But the decomposed form cannot stand alone and thus does not bear proper morphosyntactic or semantic information. Zuraw (2002) studied the transparency of phonological rule application within pseudo-reduplicated words in Tagalog loan words. For example, stem-final mid vowels in Tagalog usually raise to high vowels when suffixed, as in [ka:l**o**s] *'grain leveler'* but [kal**u**s-in] *'to use a grain leveler on'*. However, within English and Spanish loans, mid vowel raising is less frequently applied when a preceding mid vowel is present: /tod**o**+in/ *'to include all'* surfaces as [tod**o**-in] but not *[tod**u**-in]. The hypothesized motivation is that speakers preserve sub-string similarity between /to/ and /do/. A recent MEG study on visual inputs (Wray et al., 2022) further supports the reduplication-like representation for those pseudo-reduplicated words that fail to undergo a process due to similarity preservation.

3.2.1.2 Inadequacy of familiar language classes

Having established that the formal pattern ww, for unboundedly many strings w, is a reasonable model for reduplication, we can ask where this falls on the hierarchy of familiar language classes. The original Chomsky Hierarchy, shown in solid lines in Figure 3.2, classifies the ww pattern as properly context-sensitive; it is also included in the more recent *mildly context-sensitive* subclass (MCS; Joshi, 1985; Stabler, 2004), shown with a dashed line. This creates a puzzle with two parts.

The first part of the puzzle comes from the fact that reduplication is a counter-example to the otherwise overwhelming generalization that attested phonological and morphological patterns are regular. Aside from reduplication, it is very natural to hypothesize that the set of possible natural language word forms is regular (or even sub-regular). This is why the distinction above between bounded and unbounded copying is crucial: one way to save the regular hypothesis would be to demonstrate that reduplication is bounded, which would place it in the class of *finite* languages which is properly included in all of the classes shown in Figure 3.2. For example, Figure 3.3 shows a finite state automaton that successfully recognizes $\{ww \mid w \in S\}$ with a finite $S = \{aaa, aba, aab, abb, baa, bba, bab, bbb\}$. The finiteness makes it possible to essentially just memorize the desired list of surface forms.[4]

[4] Of course one might also dispute whether Figure 3.3, with its explosion in the number of states (Roark and Sproat, 2007; Dolatian and Heinz, 2020), represents a linguistically adequate model of even a bounded copying

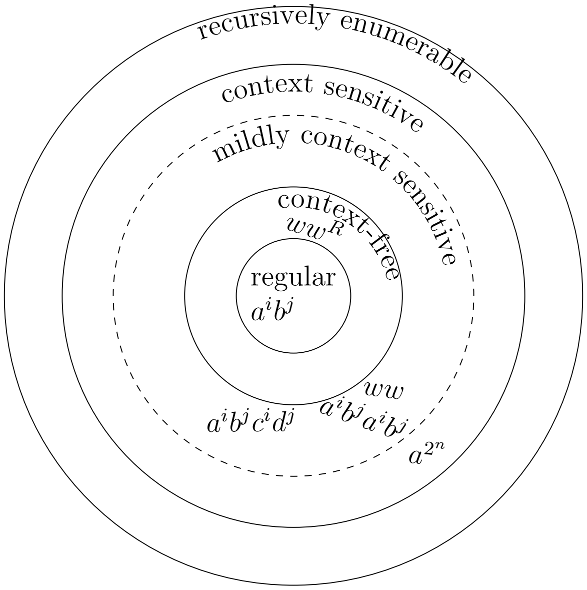


Figure 3.2: *Familiar language classes*

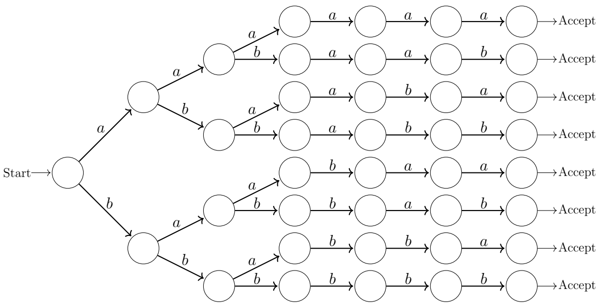


Figure 3.3: *A finite-state machine for whole-base copying with the set of bases =* {*aaa, aab, aba, abb, baa, bab, bba, bbb*}

The second part of the puzzle comes from considering the classes in Figure 3.2 that do include ww. The most restrictive of these is the mildly context-sensitive class.[5] This is not a good fit with natural language word forms because it also includes the ww^R pattern, which is unattested as discussed above; more generally, it includes *nesting* patterns as well

construction (cf. Cohen-Sygal and Wintner (2006) as in our discussion in Section 3.2.3.1). The distinction between arguing that Figure 3.3 is linguistically inadequate and arguing that copying is unbounded is subtle (Savitch, 1993).

[5] Joshi et al. (1990, p.13) provides a tree adjoining grammar for L_{ww}. A minimalist grammar can be found in Graf (2013, p.119). Multiple context-free grammars (MCF) are used to implement reduplication in Primitive Optimality Theory according to the base-reduplicant correspondence theory (Albro, 2000). Crysmann (2017) used head-driven phrase structure grammar to model partial and total reduplication in Hausa.

	linear/regular	nested	cross-serial
Morphology and Phonology	✓	✗	✓ restricted to symbol identity
Syntax	✓	✓	✓

Figure 3.4: *Attested types of dependencies in different language modules*

as *crossing* patterns (recall Figure 3.1). But the problem is slightly more subtle than the simple distinction between nesting and crossing suggests: the MCS class includes very general crossing patterns such as $a^ib^jc^id^j$, but reduplication represents a special case where the cross-serially dependent elements are identical symbols. MCS grammars are motivated by natural language syntax, where the more general kind of crossing patterns appear to be necessary[6] — the influential paper by Shieber (1985) on Swiss German appeals to exactly the aforementioned example $a^ib^jc^id^j$ — but for the purposes of morphophonology, there is reason to distinguish crossing patterns that involve surface symbol identity (e.g. ww and $a^ib^ja^ib^j$) from those that do not. This situation is summarized in Figure 3.4. We return to the distinction between formalisms where symbol identity plays a role and those where it does not in Section 3.6.2.

3.2.2 Language classes motivated by reduplication and queue automata

In response to essentially the puzzle introduced above, Gazdar and Pullum (1985, p.287) made the remark that

> We do not know whether there exists an independent characterization of the class of languages that includes the regular sets and languages derivable from them through reduplication, or what the time complexity of that class might be, but it currently looks as if this class might be relevant to the characterization of NL [natural language] word-sets.

One such proposal is offered by Manaster-Ramer (1986, p.87), who introduces the idea — closely related to that underlying our own proposal below — as follows:[7]

[6] And nesting patterns are at least as common as crossing patterns.

> Rather than grudgingly clambering up the Chomsky Hierarchy towards Context-sensitive Grammars, we should consider going back down to Regular Grammars and striking out in a different direction. The simplest alternative proposal is a class of grammars which intuitively have the same relation to queues that CFGs have to stacks.

The *Context-free Queue Grammars* (CFQGs) that Manaster-Ramer proposes adopt the format of right-linear rewrite rules for regular grammars (i.e. valid rule forms are 'A → a B' and 'A → a'), with an additional queue-based memory in which a string of terminal symbols can be accumulated. The queue-based memory is implemented by the additional capability to write terminal symbols at the right end of the output string — not only to the right of the current nonterminal, but also any terminals previously added to this queue.

There are significant similarities between CFQGs and the FSBM formalism that we introduce in this paper. Manaster-Ramer illustrates CFQGs via an example that generates $\{ww \mid w \in \{\mathrm{a}, \mathrm{b}\}^*\}$, and conjectures that they cannot generate the corresponding mirror-image (ww^R) language, but there is no careful exploration of the formalism's capacity or limitations. Also, it is clear that CFQGs can generate more general crossing patterns such as $a^ib^jc^id^j$ along with reduplication-like patterns, so FSBMs are more restricted in at least this (linguistically well-motivated) respect.

Along similar lines to Manaster-Ramer's proposal, Savitch (1989) introduced *Reduplication PDAs* (RPDAs), which are pushdown automata augmented with the ability to match reduplicated strings by using a portion of the stack as a queue. RPDAs are more powerful than CFQGs, since the language class they define properly includes context-free languages, so they do not exclude nesting/mirror-image patterns. This aligns with the fact that the motivations Savitch discusses mainly involve crossing patterns found in syntax rather than identity-based reduplication which is our focus here. But the technical formulation of RPDAs

[7] Taken literally, this quotation seems to lead in the direction of unrestricted queue automata which are known to be equivalent to Turing machines. What Manaster-Ramer actually proposes is significantly more restricted. Also, see Kutrib et al. (2018) for a more complete review of the history of queue automata and investigations on restricted versions that computer scientists have conducted.

has much in common with that of FSBMs below.

Finally, *Memory Automata* (MFAs; Schmid, 2016; Freydenberger and Schmid, 2019) introduce a kind of automata that is particularly similar to FSBMs. MFAs augment classical FSAs with a finite number of memory cells; each memory cell can store an unboundedly long sub-string of input, which can be matched against future input when it is recalled. The full class of MFAs can generate languages such as $\{a^i \,|\, i \text{ is not prime}\}$ (Câmpeanu et al., 2003, p.1013) and $\{a^{4^i} \,|\, i \geq 1\}$ (Freydenberger and Schmid, 2019, p.21), and is therefore much too powerful to be suitable as a model for natural languages.[8] But these unusually "complex" languages all rely on either interactions between distinct memory cells, or the ability to recall a particular string from a memory cell more than once. The FSBM formalism that we introduce corresponds closely to a restricted version of MFAs where there is only one memory cell, and its contents are erased when recalled.

To summarize: our goal is to identify a formalism whose class of languages aligns with Gazdar and Pullum's motivating quotation above; RPDAs do not match this description because they extend upwards from the context-free languages, rather than the regular languages; CFQGs and MFAs do adopt the regular languages as the starting point, but extend too far and therefore overshoot the mark in different ways.

This paper introduces FSBMs as a way of examining what minimal changes can be brought to regular languages to include string-sets with two copies of the same sub-strings, while excluding some typologically unattested context-free patterns, such as reversals, and crossing dependencies other than reduplication. We name the resulting class of languages *regular copying languages* (RCLs). The intended relation of this language class to other existing language classes is shown in Figure 3.5.

[8] MFAs were introduced to provide an automaton-based characterization of the languages generated by regular expressions extended with back-references (Câmpeanu et al., 2002; Câmpeanu et al., 2003; Carle and Narendran, 2009). There are some differences between the various definitions of these "extended regular expressions" in the literature; see Freydenberger and Schmid (2019, pp. 36–37) for discussion. We would like to thank an anonymous reviewer for pointing out the relevant research on extended regular expressions, which in turn led us to the literature on MFAs.

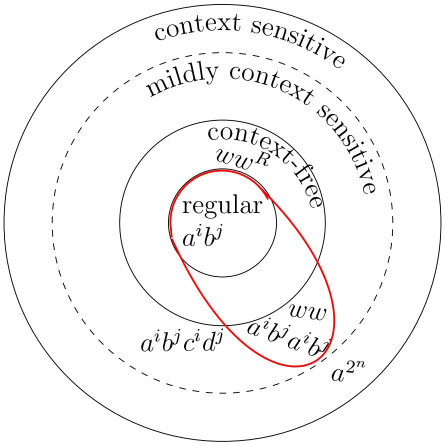


Figure 3.5: *The class of regular copying languages (oval shape) in the classical Chomsky Hierarchy*

3.2.3 Other computational models motivated by reduplication

Now we review other computational models motivated by reduplication, which can be categorized into two groups: those that limit attention to bounded copying, (Section 3.2.3.1), and those that consider transductions/mappings (Section 3.2.3.2).

3.2.3.1 Compact representations of bounded copying

moThe first line of work aims to improve upon the inelegant "memorization" strategy exemplified in Figure 3.3, while retaining the limitation to bounded copying. For example, Cohen-Sygal and Wintner (2006) introduce *finite-state registered automata* (FSRAs), which augment standard FSAs with finitely many memory registers. This allows for a more space-efficient representation of copying patterns, without the "duplicating paths" of Figure 3.3, by storing the symbols to be matched in registers rather than in the machine's central state. But because the registers themselves provide only a finite amount of additional memory, FSRAs do not extend upon the generative capacity of standard FSAs, and therefore do not accommodate productive total reduplication (i.e. unbounded copying).

An analogous proposal is the *compile-replace* algorithm (Beesley and Karttunen, 2000). This run-time technique first maps a lexical item to a regular expression representation for either morphological generation or analysis. Then the desired output is obtained by re-evaluating the output regular expression. Similarly, Walther (2000) added different types of transitions to represent the lexicon: *repeat* (for copying), *skip* (for truncation) and *self-loops* (for infixation). Then, intersecting these enriched lexical items with an FSA encoding language-specific reduplication rules would derive the surface strings. Last but not least, Hulden (2009) introduced an EQ function, a filter on a finite-state transduction which excludes input-output pairs where the output string does not meet a sub-string identity condition. In principle, this idea allows for an unbounded-copying output language such as $\{ww \mid w \in \{\mathrm{a}, \mathrm{b}\}^*\}$ to be specified, but in practice, Hulden's implementation restricts attention to cases where the equal sub-strings are bounded in length (p.125).

3.2.3.2 2-way Deterministic Finite-state Transducers

A finite-state method that computes unbounded copying elegantly and adequately is *2-way deterministic finite-state transducers* (2-way D-FSTs) (Dolatian and Heinz, 2018a,b, 2019, 2020), which differ from conventional (1-way) FSTs in being able to move back and forth on the input.[9] 2-way D-FSTs have been proven to describe string transductions that are MSO-definable (Monadic Second-Order logic; Engelfriet and Hoogeboom, 1999) and are equivalent to *streaming string transducers* (Alur and Černý, 2010). In these formalisms, reduplication is modeled as a string-to-string *mapping* ($w \mapsto ww$). To avoid the mirror image function ($w \mapsto ww^R$), Dolatian and Heinz (2020) further studied sub-classes of 2-way D-FSTs which cannot output anything during right-to-left passes over the input (cf. *rotating transducers*: Baschenis et al., 2017).

The issue addressed in Dolatian and Heinz (2020) is distinct from, but related to, the main concern of this paper: these transducers model reduplication as a function mapping underlying forms to surface forms ($w \mapsto ww$), while this paper aims to characterize only the identical-substrings requirement on the corresponding surface forms (ww). There are at least two reasons to address the string-set problem itself rather than considering only mappings

[9] 2-way FSTs are more restricted than Turing machines since they cannot move back and forth on the output tape, only the input tape.

between underlying and surface forms.

The first reason is a practical/strategic one, related to the problem of morphological *analysis* (rather than generation): the question of what kinds of transducers can implement the $ww \mapsto w$ mapping required for morphological analysis remains open, since 2-way D-FSTs (unlike standard 1-way FSTs) are not readily invertible as a class (Dolatian and Heinz, 2020, p.235). Although we do not directly address the morphological analysis problem here, recognizing the reduplicated ww strings is plausibly an important first step: applying the mapping $ww \mapsto w$ to some string x requires at least *recognizing* whether x belongs to the ww string set.

The second reason stems from a full consideration of the linguistic facts surrounding reduplication: there is evidence supporting meaning-free, non-morphologically-generated reduplication-like structures, as mentioned in the discussion of aggressive reduplication above. This suggests that the phonological grammar involves a *phonotactic* constraint requiring sub-string identity, and the natural formal model for such a constraint is an automaton that generates/accepts the strings satisfying it. A constraint of this sort could play a role in mappings relating underlying forms to surface forms, so we may be missing a generalization if we only model those mappings directly.

3.3 Finite-state Buffered Machines

The aim of proposing a new computing device is to add reduplication to FSAs and thereby gain a better understanding of the required computational operations. The new formalism is *finite-state buffered machines* (FSBMs), a summary of which is provided in Section 3.3.1. For ease of exposition, we introduce the new formalism by first presenting the general case of FSBMs in Section 3.3.2, along with illustrative examples. A clearer understanding of the formalisms' capacity for copying comes from identifying a subset of FSBMs that we call *complete-path FSBMs*, in Section 3.3.3; we show that the languages recognized by FSBMs are precisely the languages recognized by complete-path FSBMs in Section 3.3.4.

3.3.1 FSBM in a nutshell

FSBMs are two-taped automata with finite-state core control.[10] One tape stores the input, as in normal FSAs; the other serves as an unbounded memory buffer, storing reduplicants temporarily for future string matching. An FSBM can be thought of as an extension to the FSRAs discussed above (Cohen-Sygal and Wintner, 2006) but equipped with unbounded memory. FSBMs with a *bounded* buffer would be as expressive as FSRAs, and therefore also standard FSAs.

The interaction of the queue-like buffer with the input is restricted in two important ways. First, the buffer stores symbols from the same alphabet as the input, unlike the stack in a PDA, for example. Second, once one symbol is removed from the buffer, everything else must also be emptied from the buffer before symbols can next be added to it. These restrictions together ensure the machine will not generate string reversals or other non-reduplicative non-regular patterns.

Unlike a standard FSA, an FSBM works with two possible modes: in *normal* (N) mode, M reads symbols and transits between states, functioning as a normal FSA; and in *buffering* (B) mode, besides consuming symbols from the input and taking transitions among states, M adds a copy of just-read symbols to the queue-like buffer. At a specific point, M exits buffering (B) mode, matching the stored string in the buffer against (a portion of) the remaining input. Provided this match succeeds, it switches back to normal (N) mode for another round of computation. Figure 3.6 provides a schematic diagram showing how the mode of an FSBM alternates when it determines the equality of sub-strings and how the buffer interacts with the input. As presented here, FSBMs can only compute local reduplication with two adjacent, completely identical copies. They cannot handle non-local reduplication, multiple reduplication, or non-identical copies. We believe the current machinery can serve as the foundation for proposing different variants, and we discuss some potential modifications along these lines in Section 3.7.

Having introduced the important intuitions, we now turn to the formal definition of FSBMs.

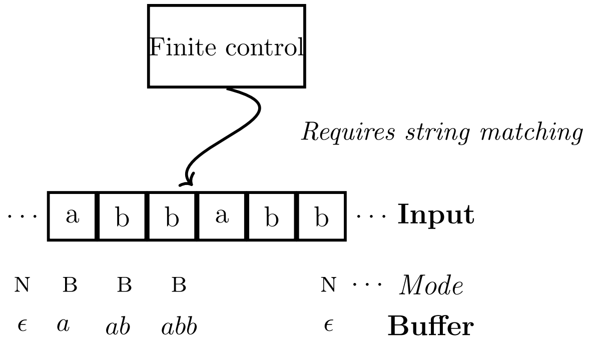

Finite control
Requires string matching
··· a b b a b b ··· Input
N B B B N ··· Mode
ϵ a ab abb ϵ Buffer

Figure 3.6: *Mode changes and input-buffer interaction of an FSBM M on "…abbabb…". The machine switches to* B *mode to temporarily store symbols in the queue-like buffer, and then at the point indicated by the arrow it compares the buffer contents against the remaining input. If the two strings match, the buffer is emptied, the matched input sub-string is consumed and the machine switches to* N *mode*

.

3.3.2 Preliminaries & Definitions

For any finite alphabet Σ of symbols, we use Σ^* to denote the set of all finite strings over Σ. For a string w, $|w|$ denotes its length. ϵ is the null string and thus $|\epsilon| = 0$. We denote string union by '+', and denote string concatenation by simple juxtaposition, assuming implicit conversion between symbols and length-one strings where necessary. If $u = vw$, then $v\backslash u = w$; otherwise, $v\backslash u$ is undefined. For example, $ab\backslash abb = b$.

Definition 1. *A* ***Finite-State Buffered Machine*** *is a 7-tuple* $\langle \Sigma, Q, I, F, G, H, \delta \rangle$ *where*

- Σ*: a finite set of symbols*
- Q*: a finite set of states*
- $I \subseteq Q$*: initial states*
- $F \subseteq Q$*: final states*
- $G \subseteq Q$*: states where the machine must enter buffering mode*
- $H \subseteq Q - G$*: states requiring string matching*
- δ*:* $Q \times (\Sigma \cup \{\epsilon\}) \times Q$*: transition relation*

The specification of the two sets of special states, G and H, serves to control what portions of a string are copied. To avoid intricacies, G and H are defined to be disjoint. The special case where $G = H = \emptyset$ corresponds to a standard FSA.

Definition 2. *A* ***configuration*** *of an FSBM is a four-tuple* $(u, q, v, t) \in \Sigma^* \times Q \times \Sigma^* \times \{\textsc{n}, \textsc{b}\}$, *where* u *is the input string;* q *is the current state;* v *is the string in the buffer; and* t *is the machine's current mode.*

Definition 3. *Given an FSBM* $M = (\Sigma, Q, I, F, G, H, \delta)$, *the relation* $\vdash_M$ *on configurations is the smallest relation such that, for any* $u, v, w \in \Sigma^*$*:*

- *For every transition* $(q_1, x, q_2) \in \delta$

  $(xu,\ q_1,\ \epsilon,\ \textsc{n}) \vdash_M (u,\ q_2,\ \epsilon,\ \textsc{n})$ *if* $q_1 \notin G$ *and* $q_2 \notin H$ $\qquad \vdash_{\textsc{n}}$

  $(xu,\ q_1,\ v,\ \textsc{b}) \vdash_M (u,\ q_2,\ vx,\ \textsc{b})$ *if* $q_1 \notin H$ *and* $q_2 \notin G$ $\qquad \vdash_{\textsc{b}}$

- *For every* $q \in G$

  $(u,\ q,\ \epsilon,\ \textsc{n}) \vdash_M (u,\ q,\ \epsilon,\ \textsc{b})$ $\qquad \vdash_{\textsc{n}\to\textsc{b}}$

- *For every* $q \in H$

  $(vw,\ q,\ v,\ \textsc{b}) \vdash_M (w,\ q,\ \epsilon,\ \textsc{n})$ $\qquad \vdash_{\textsc{b}\to\textsc{n}}$

Thus, $\vdash_M = \vdash_{\textsc{n}} \cup \vdash_{\textsc{b}} \cup \vdash_{\textsc{n}\to\textsc{b}} \cup \vdash_{\textsc{b}\to\textsc{n}}$. *When* $D_1 \vdash_M D_2$, *we say* D_1 ***yields*** D_2.

As is standard, $\vdash^*$ denotes the reflexive and transitive closure of $\vdash$, while $\vdash^+$ is the corresponding irreflexive closure.

Definition 4. *A* ***run*** *of* M *on* w *is a sequence of configurations* $D_0, D_1, D_2 \ldots D_m$ *such that*

- $\exists q_0 \in I,\ D_0 = (w, q_0, \epsilon, \textsc{n})$
- $\exists q_f \in F,\ D_m = (\epsilon, q_f, \epsilon, \textsc{n})$
- $\forall\, 0 \leq i < m,\ D_i \vdash_M D_{i+1}$

Definition 5. *The language recognized by* $M = \langle \Sigma, Q, I, F, G, H, \delta \rangle$, *denoted by* $L(M)$, *is the set of all strings* $w \in \Sigma^*$ *such that there is a run of* M *on* w. *That is,* $L(M) = \{w \in \Sigma^* \mid (w, q_0, \epsilon, \textsc{n}) \vdash_M^* (\epsilon, q_f, \epsilon, \textsc{n}), q_0 \in I, q_f \in F\}$.

Notice that we do not impose any notion of determinism on the transitions of an FSBM. We return to some discussion of this point in Section 3.6.1.

Now, we give examples of FSBMs. In all illustrations, G states are drawn with diamonds and H states are drawn with squares.

3.3.2.1 Examples: Total reduplication

Figure 3.7 offers an FSBM M_1 for L_{ww}, with arbitrary strings over the alphabet $\Sigma = \{a, b\}$ as potential bases. The initial state q_1 is also a G state, and the only H state is q_3. The machine stores a copy of string computed in between q_1 and q_3 in the buffer and requires string matching at q_3. Since the states where the machine enters ($q_1 \in G$) and leaves ($q_3 \in H$) buffering mode are also the initial and final states respectively, this machine will recognize simple total reduplication. Table 3.3 gives a complete run of M_1 on the string *abbabb*. As we can see in Step 8, the string *abb* in the remaining input is consumed in one step.

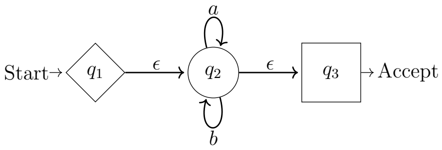
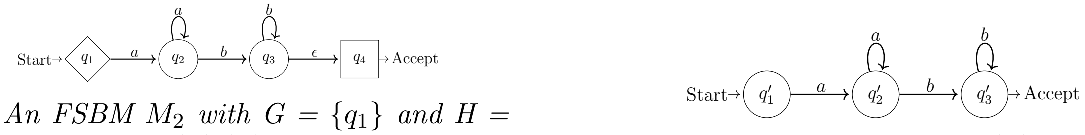


Figure 3.7: *M_1 with $G = \{q_1\}$ and $H = \{q_3\}$. $L(M_1) = \{ww \mid w \in \{a, b\}^*\}$*

	Used arc or *state*	⊢ types	*Configuration* (input, state, buffer, mode)
1.	*N/A*		(*abbabb*, q_1, ϵ, N)
2.	$q_1 \in G$	$\vdash_{\text{N}\rightarrow\text{B}}$	(*abbabb*, q_1, ϵ, B)
3.	(q_1, ϵ, q_2)	$\vdash_{\text{B}}$	(*abbabb*, q_2, ϵ, B)
4.	(q_2, a, q_2)	$\vdash_{\text{B}}$	(*bbabb*, q_2, *a*, B)
5.	(q_2, b, q_2)	$\vdash_{\text{B}}$	(*babb*, q_2, *ab*, B)
6.	(q_2, b, q_2)	$\vdash_{\text{B}}$	(*abb*, q_2, *abb*, B)
7.	(q_2, ϵ, q_3)	$\vdash_{\text{B}}$	(*abb*, q_3, *abb*, B)
8.	$q_3 \in H$	$\vdash_{\text{B}\rightarrow\text{N}}$	(ϵ, q_3, ϵ, N)
		Accept	

Table 3.3: *M_1 in Figure 3.7 accepts abbabb*

For the rest of the illustration, we focus on the FSBM M_2 in Figure 3.8a. M_2 in Figure 3.8a recognizes the non-context-free language $\{a^ib^ja^ib^j \mid i, j \geq 1\}$. This language can be viewed as total reduplication added to the regular language $\{a^i b^j \,|\, i, j \geq 1\}$ (recognized by the FSA M_0 in Figure 3.8b). q_1 is an initial state and more importantly a G state, forcing M_2 to enter B at the beginning of any run. Then M_2 in B mode always keeps a copy of consumed symbols until it proceeds to q_4, which is an H state and therefore requires M_2 to stop buffering and check for string identity to empty the buffer. Then, M_2 with a blank buffer can switch to N mode. It eventually ends at q_4, a legal final state. Table 3.4 shows one possible sequence of configurations of M_2 on *ababb*; this string is rejected because there is no way to reach a valid ending configuration.



(a) *An FSBM M_2 with $G = \{q_1\}$ and $H = \{q_4\}$; L(M_2) = $\{a^i b^j a^i b^j \,|\, i, j \geq 1\}$*

(b) *An FSA M_0; L(M_0)= $\{a^i b^j \,|\, i, j \geq 1\}$*

Figure 3.8: *One example FSBM and the corresponding FSA for the base language*

	Used Arc	$\vdash$ types	*Configuration*
	or *state*		(input, state, buffer, mode)
1.	*N/A*		(*ababb*, q_1, ϵ, N)
2.	$q_1 \in G$	$\vdash_{\text{N}\rightarrow\text{B}}$	(*ababb*, q_1, ϵ, B)
3.	(q_1, a, q_2)	$\vdash_{\text{B}}$	(*babb*, q_2, *a*, B)
4.	(q_2, b, q_3)	$\vdash_{\text{B}}$	(*abb*, q_3, *ab*, B)
5.	(q_3, ϵ, q_4)	$\vdash_{\text{B}}$	(*abb*, q_4, *ab*, B)
6.	$q_4 \in H$	$\vdash_{\text{B}\rightarrow\text{N}}$	(*b*, q_4, ϵ, N)
		Reject	

Table 3.4: *M_2 in Figure 3.8a rejects ababb*

	Used Arc	$\vdash$ types	*Configuration*
1.	*N/A*		(*taktakki*, q_1, ϵ, N)
2.	$q_1 \in G$	$\vdash_{\text{N}\to\text{B}}$	(*taktakki*, q_1, ϵ, B)
3.	(q_1, t, q_2)	$\vdash_{\text{B}}$	(*aktakki*, q_2, *t*, B)
4.	(q_2, a, q_3)	$\vdash_{\text{B}}$	(*ktakki*, q_3, *ta*, B)
5.	(q_3, k, q_4)	$\vdash_{\text{N}\to\text{B}}$	(*takki*, q_4, *tak*, B)
6.	$q_4 \in H$	$\vdash_{\text{B}\to\text{N}}$	(*ki*, q_4, ϵ, N)
7.	(q_4, *k*, q_5)	$\vdash_{\text{N}}$	(*i*, q_5, ϵ, N)
8.	(q_5, *i*, q_5)	$\vdash_{\text{N}}$	(ϵ, q_5, ϵ, N)
Accept			

Table 3.5: *M_3 in Figure 3.9 accepts taktakki*

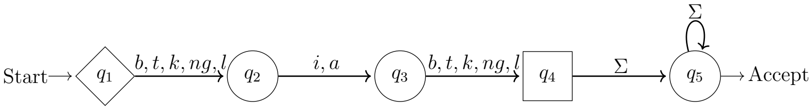


Figure 3.9: *An FSBM M_3 for Agta CVC-reduplicated plurals: $G = \{q_1\}$ and $H = \{q_4\}$*

3.3.2.2 Examples: Partial reduplication

Assuming $\Sigma = \{b, t, k, ng, l, i, a\}$, the FSBM M_3 in Figure 3.9 serves as a simple model of Agta CVC reduplicated plurals, as illustrated earlier in Table 3.1. Given the initial state q_1 is in G, M_3 has to enter B mode before it takes any transitions. In B mode, M_3 transits to a plain state q_2, consuming a consonant from the input and keeping it in the buffer. Similarly, M_3 transits to a plain state q_3 and then to q_4. When M_3 first reaches q_4, the buffer would contain a CVC sequence; q_4, an H state, requires M_3 to match this CVC sequence in the buffer with the remaining input. Then, M_3 with a blank buffer can switch to N mode at q_4. It transitions to q_5 to process the rest of the input via the normal loops on q_5. A successful run should end at q_5, the only final state. Table 3.5 gives a complete run of M_3 on the string "*taktakki*". Table 3.6 illustrates a case where the crucial step of returning from B mode to N mode is not possible, because of the non-matching sub-strings in "*tiktakki*"; this string is rejected by M_3.

	Used Arc	$\vdash$ types	*Configuration*
1.	*N/A*		$(\textit{tiktakki}, q_1, \epsilon, \textsc{n})$
2.	$q_1 \in G$	$\vdash_{\textsc{n}\to\textsc{b}}$	$(\textit{tiktakki}, q_1, \epsilon, \textsc{b})$
3.	(q_1, t, q_2)	$\vdash_{\textsc{b}}$	$(\textit{iktakki}, q_2, t, \textsc{b})$
4.	(q_2, i, q_3)	$\vdash_{\textsc{b}}$	$(\textit{ktakki}, q_3, ti, \textsc{b})$
5.	(q_3, k, q_4)	$\vdash_{\textsc{b}}$	$(\textit{takki}, q_4, tik, \textsc{b})$
$q_4 \in H$: *checks for string identity and rejects*			

Table 3.6: *M_3 in Figure 3.9 rejects tiktakki*

3.3.3 The copying mechanism and complete-path FSBMs

The copying mechanism is realized by four essential components: 1) the unbounded memory buffer, which has queue-like storage; 2) added modalities (i.e. the normal mode N and the buffering mode B); 3) added specifications of states requiring the machine to buffer symbols into memory, namely states in G; 4) added specifications of states requiring the machine to empty the buffer by matching sub-strings, namely states in H.

As shown in the definitions of configuration changes and the examples in Section 3.3.2, the machine must end in N mode to accept an input. There are two possible scenarios for a run to meet this requirement: either never entering B mode or undergoing full cycles of N → B → N mode changes. Correspondingly, the resulting languages reflect either no copying (functioning as plain FSAs) or full copying.

In any specific run, it is the states that inform a machine M of its modality. The first time M reaches a G state, it has to enter B mode and keeps buffering when it transits between plain states. The first time when it reaches an H state, M is supposed to match strings. Hence, it is clear that to go through full cycles of mode changes, once M reaches a G state and switches to B mode, it has to encounter some H state later. Then the buffer has to be emptied for N mode at the point when a H state transits to a plain state. A template for those machines performing full copying can be seen in Figure 3.10.

To allow us to reason about only the "useful" arrangements of G and H states, we impose an ordering requirement on G and H states in a machine. We define the *completeness restriction* on a path in Definition-7. We then identify those FSBMs in which all paths are

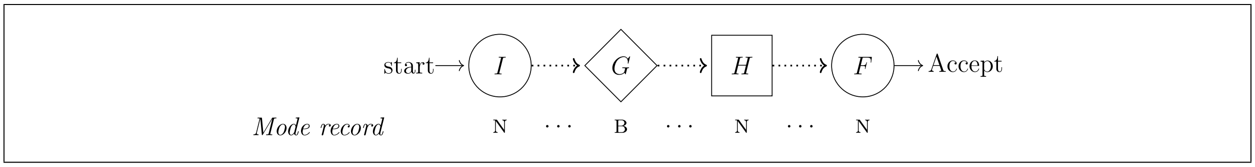


Figure 3.10: *The template for the implementation of the copying in FSBMs. Key components: G state, H states, and strict ordering between G and H. Dotted lines represent a sequence of transitions*

complete as *complete-path FSBMs*. The machine M_1 in Figure 3.7, M_2 Figure 3.8a and M_3 in Figure 3.9 are all complete-path FSBMs.

Definition 6. *A* ***path*** *from one state p_1 to another state p_n in an FSBM M is a sequence of states $p_1, p_2, p_3, \ldots p_n$ such that for each $i \in \{1, \ldots, n-1\}$, there is a transition $(p_i, x, p_{i+1}) \in \delta_M$.*

Definition 7. *A path in an FSBM M is* ***complete*** *if it is in the denotation of the regular expression $(P^*GP^*HP^* + P^*)^*$, where P represents any state in $Q - (G \cup H)$. A* ***complete-path FSBM*** *is an FSBM in which any path $p_1 \ldots p_n$ with $p_1 \in I$ and $p_n \in F$ is complete.*

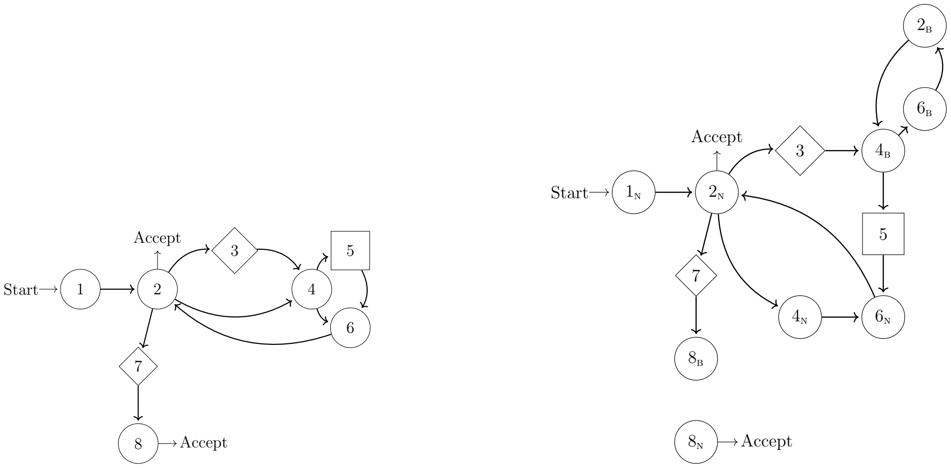
Definition 8. *A path is said to be* ***a copying path*** *if it is complete and there is at least one G state (or at least one H state).*

3.3.4 The sufficiency of complete-path FSBMs

Now, we show that the languages recognized by FSBMs are precisely the languages recognized by complete-path FSBMs; this will allow us to restrict attention to complete-path FSBMs when studying the formal properties of these machines below.

Proposition 1. *For any FSBM M, there exists a complete-path M' with $L(M) = L(M')$.*

Incomplete paths contribute nothing to the language generated by an FSBM, so showing this equivalence requires showing that, for any FSBM M_1, we can construct a new FSBM M_2 such that every path from an initial state to an accepting state in M_2 corresponds to some complete path from an initial state to an accepting state in M_1. The idea is that M_2 is a complete-path FSBM that keeps only those paths from M_1 that are indeed complete.



(a) M_1 **(b)** M_2

Figure 3.11: *Construction of a complete-path FSBM* M_2 *that is equivalent to* M_1*.*

The non-obvious cases of this construction involve scenarios where some plain state in M_1 might be reached either in normal (N) mode or in buffering (B) mode, depending on the path by which that plain state is reached. In Figure 3.11a, for example, this is the case for states 2, 4 and 6: intuitively, a path from state 2 back to itself might contain a G state (3) or an H state (5), or both or neither. To construct an equivalent complete-path FSBM M_2, we "split" each plain state q into two distinct states q_{N} and q_{B}. Transitions from a G state to q and transitions from q to an H state (i.e. transitions that only make sense in buffering mode) are carried over in M_2 for q_{B} but not for q_{N}. Similarly, transitions from an H state to q and transitions from q to a G state are carried over in M_2 for q_{N} but not for q_{B}. And the status of q as an initial and/or accepting state is carried over for q_{N} but not for q_{B}. Figure 3.11b shows the resulting complete-path FSBM for this example. In addition to keeping track of the mode in which states 2, 4 and 6 are visited, notice that this construction also prevents state 7 from occurring in any path from an initial state to an accepting state, since 8_{B} is not an accepting state and 8_{N} is unreachable.

3.4 Pumping Lemma

We define the *Regular Copying Languages* (RCLs) to be the set of all languages accepted by some (complete-path) FSBMs. To be able to prove that some languages are not RCLs, we present a pumping lemma in this section. The idea that is if an FSBM produces a string $urrv$ via a copying run, and r is sufficiently long, then some subpart of r will be pumpable in the manner of the familiar pumping lemma for regular languages; that is, r can be broken into $x_1x_2x_3$ such that $ux_1x_2^ix_3x_1x_2^ix_3w$ is also accepted.[11]

Theorem 1. *If $\mathcal{L}$ is a regular copying language, there is a positive integer k such that for every string $w \in \mathcal{L}$ with $|w| \geq 4k$, one of the following two conditions holds:*

1. *w can be rewritten as $w = xyz$ with*
   (a) $|y| \geq 1$
   (b) $|xy| \leq k$
   (c) $\forall i \geq 0, xy^iz \in \mathcal{L}$
2. *w can be rewritten as $w = ux_1x_2x_3x_1x_2x_3v$ such that*
   (a) $|x_2| \geq 1$
   (b) $|x_1x_2| \leq k$
   (c) $\forall i \geq 0, ux_1x_2^ix_3x_1x_2^ix_3v \in \mathcal{L}$

Proof. Since $\mathcal{L}$ is a regular copying language, there is a complete-path FSBM M that recognizes $\mathcal{L}$. Let k be the number of states in M. For an arbitrary string $w \in \mathcal{L}$ with $|w|$¿$4k$, there is at least one path through M that generates w. Let p be the shortest such path (or if there are ties, choose arbitrarily). Note that p does not contain any ϵ-*loops*; if it did, its length would not be minimal among all candidate paths.

Suppose first that p is not a copying path. The length of p is at least $|w|+1$, and so since $|w| \geq 4k > k$, some state must occur twice in p, in fact in the first $k+1$ elements of p. As in the standard pumping lemma for regular languages, this means that w can be rewritten as xyz, with $|xy| \leq k$, in such a way that M can also generate xy^iz by repeating the loop, and $y \neq \epsilon$ since p contains no ϵ-loops. So in this case, w satisfies Condition 1.

[11] This idea is largely inspired by Savitch (1989, p.256), who proposes a pumping lemma for context-free languages augmented with copying.

If $p = p_0p_1 \dots p_n$ is a copying path, then the run that generates $w = urrv$ must have the form $(urrv, p_0, \epsilon, \text{N}) \vdash^*_M (rrv, p_i, \epsilon, \text{N}) \vdash_M (rrv, p_i, \epsilon, \text{B}) \vdash^*_M (rv, p_j, r, \text{B}) \vdash_M (v, p_j, \epsilon, \text{N}) \vdash^*_M (\epsilon, p_n, \epsilon, \text{N})$ with $p_0 \in I$, $p_i \in G$, $p_j \in H$ and $p_n \in F$. Since $|w| \geq 4k$, at least one of $|u|, |r|, |v|$ is greater than or equal to k.

- If $|r| \geq k$, then $|p_i \dots p_j| \geq |r|+1 \geq k+1$, so at least one state must appear twice in the first $k+1$ elements of the sequence $p_i \dots p_j$, i.e there are ℓ and ℓ' such that $i \leq \ell < \ell' \leq j$ and $p_\ell = p_{\ell'}$, with $\ell' - i < k$. Then it must be possible to rewrite r as $x_1x_2x_3$, with $|x_1x_2| \leq k$, such that repeating the subpath $p_\ell \dots p_{\ell'}$ results in pumping x_2, and so any string of the form $x_1x_2^ix_3$ can be consumed from the input and stored in the buffer in the course of moving from $p_i \in G$ to $p_j \in H$, i.e. $(x_1x_2^ix_3x_1x_2^ix_3v, p_i, \epsilon, \text{B}) \vdash^*_M (x_1x_2^ix_3v, p_j, x_1x_2^ix_3, \text{B}) \vdash_M (v, p_j, \epsilon, \text{N})$. M will therefore generate all strings of the form $ux_1x_2^ix_3x_1x_2^ix_3v$, satisfying Condition 2.

- If $|u| \geq k$, then $|p_0 \dots p_i| \geq |u| + 1 \geq k + 1$, so at least one state must appear twice in the sequence $p_0 \dots p_i$, i.e. there are ℓ and ℓ' such that $0 \leq \ell < \ell' \leq i$ and $p_\ell = p_{\ell'}$, with $\ell' < k$. There are two cases to consider:

  - Suppose that M is in buffering mode throughout the part of the run from p_ℓ to $p_{\ell'}$. Therefore $p_\ell = p_{\ell'}$ is a plain state. Then it must be possible to rewrite u as $u'x_1x_2x_3x_1x_2x_3v'$, such that repeating the subpath $p_\ell \dots p_{\ell'}$ results in pumping x_2. And since the repeated state must occur in the first $k + 1$ elements of p, $|u'x_1x_2| \leq k$ and therefore $|x_1x_2| \leq k$. M will therefore generate all strings of the form $u'x_1x_2^ix_3x_1x_2^ix_3v'rrv$, satisfying Condition 2.

  - Otherwise, it must be possible to rewrite u as $x_1x_2x_3$ such that repeating this loop pumps x_2; since M is a complete-path FSBM, repeating the loop cannot create incomplete paths. And since the repeated state must occur in the first $k + 1$ elements of p, $|x_1x_2| \leq k$. M will therefore generate all strings of the form $x_1x_2^ix_3rrv$, satisfying Condition 1.

- If $|v| \geq k$, an analogous argument shows that either Condition 1 or Condition 2 is satisfied.

□

Having completed the proof of the pumping lemma, we turn to the question of how to use this theorem to mathematically check whether a language belongs to the regular copying class.

Theorem 2. $\mathcal{L}_{inv} = \{(a+b)^i c^j (a+b)^i c^j \mid i, j \geq 0\}$ *is not an RCL.*

Proof. Suppose $\mathcal{L}_{inv}$ is an RCL. Let $w = a^k c^{k+1} b^k c^{k+1} \in \mathcal{L}_{inv}$, where k is the pumping length from Theorem 1. Given $|w| > 4k$, one of the conditions from Theorem 1 must hold.

1. Assume condition 1 holds. That is $w = xyz$ such that (i) $|y| \geq 1$, (ii) $|xy| \leq k$ and (iii) $\forall i \geq 0, xy^i z \in L$. Given $|xy| \leq k$, y must only contain as. Therefore $xyyz$ must have the form $a^{k+|y|} c^{k+1} b^k c^{k+1}$, so $xyyz \notin \mathcal{L}_{inv}$, a contradiction.

2. Assume condition 2 holds. Then, $w = ux_1x_2x_3x_1x_2x_3v$ such that (i) $|x_2|$¿1, (ii) $|x_1x_2| \leq k$ and (iii) $\forall i \geq 0$, $ux_1x_2^ix_3x_1x_2^ix_3v \in \mathcal{L}_{inv}$. The string x_1x_2 cannot contain the sub-string ac, because x_1x_2 occurs twice in w but ac does not; similarly, x_1x_2 cannot contain cb or bc. There remain three possible ways of choosing x_1x_2 with $|x_1x_2| \leq k$, each incurring a contradiction.

   (a) If x_1x_2 contains only as, then x_3 must also contain only as because it occurs in between the two occurrences of x_1x_2 in w. Therefore $ux_1x_2^2x_3x_1x_2^2x_3v$ must have the form $a^\ell c^{k+1} b^k c^{k+1}$ with $\ell > k$, and is therefore not in $\mathcal{L}_{inv}$; a contradiction.

   (b) Similarly, if x_1x_2 contains only bs, then $ux_1x_2^2x_3x_1x_2^2x_3v$ must have the form $a^k c^{k+1} b^\ell c^{k+1}$ with $\ell > k$, and is therefore not in $\mathcal{L}_{inv}$; a contradiction.

   (c) Finally, suppose x_1x_2 contains only cs. If x_3 did not contain only cs, then it would need to cover the sub-string b^k since it appears in between the two occurrences of x_1x_2 in w; but if x_3 covered the sub-string b^k then this sub-string would occur twice in w, which it does not. So x_3 must also contain only cs. Therefore

$ux_1x_2^2x_3x_1x_2^2x_3v$ must have the form either $a^kc^\ell b^kc^{k+1}$ or $a^kc^{k+1}b^kc^\ell$, with $\ell > k+1$; a contradiction.

□

Example 1. *Some Non-RCL languages*

1. $\mathcal{L}_{SwissGerman} = \{a^ib^jc^id^j \mid i, j \geq 0\}$
2. $\mathcal{L} = \{a^nb^n \mid n \geq 0\}$
3. $\mathcal{L} = \{ww^R \mid w \in \Sigma^*\}$
4. $\mathcal{L} = \{www \mid w \in \Sigma^*\}$
5. $\mathcal{L} = \{w^{(2^n)} \mid n \geq 0\}$

To see that $\{w^{(2^n)} \mid n \geq 0\}$ is not an RCL, notice that the pumping lemma above requires that a constant-sized increase in the length of a string in the language can produce another string also in the language, but $w^{(2^n)}$ does not have this "constant growth" property (Joshi, 1985).

Example 2. *Unattested partial reduplication is non-RCL*

1. $L_{half\ copy} = \{\frac{w}{2}w \mid w \in \Sigma^*\}$ *where* $\frac{w}{2}$ *is the first half of* w*. For the sake of simplicity, let us assume* w *is even-lengthed.*
2. $\forall k \in \mathbb{N}$ *and* $k \geq 2$, $L_{kth\ copy} = \{\frac{w}{k}w \mid w \in \Sigma^*\}$*. For the sake of simplicity, let us assume the length of* w *is a multiple of* k*.*

To see that $L_{\text{half copy}}$ is not a regular copying language, let us first re-express it as $\{wwx \mid w, x \in \Sigma^*, |w| = |x|\}$. The main challenge for FSBM is that after checking the copied portion, the current machinery is impossible to ensure that the last portion x has the same length as w. The same rationale holds for $L_{\text{kth copy}}$.

3.5 Closure properties

The class of regular copying languages is closed under the following operations: intersection with a finite-state language (Section 3.5.1), some regular operations (union, concatenation, Kleene star; Section 3.5.2) and homomorphism (Section 3.5.3). But it is not closed under intersection, nor complementation (Section 3.5.4). More interestingly, it is not closed under inverse homomorphism (Section 3.5.5). In this section, we present proofs of these results.

3.5.1 Closure under intersection with regular languages

In this subsection, we write $\mathbf{0}$ for the zero matrix and $\mathbf{I}$ for the identity matrix, with the size of these matrices determined implicitly by context.

For any FSA $M = \langle Q, \Sigma, I, F, \delta \rangle$ and any symbol $x \in \Sigma$, $\mathbf{A}_x^M \in \{0,1\}^{|Q| \times |Q|}$ is the square matrix with rows and columns indexed by Q, whose (q_1, q_2) entry is 1 if $(q_1, x, q_2) \in \delta$ and is 0 otherwise. We will sometimes just write $\mathbf{A}_x$ where the FSA is clear from the context. We define $\mathbf{A}_\epsilon^M = \mathbf{I}$, and for any non-empty string $w = x_1 \dots x_n$ we define $\mathbf{A}_w^M = \mathbf{A}_{x_1}^M \dots \mathbf{A}_{x_n}^M$. Then it follows that the (q_1, q_2) entry of the matrix $\mathbf{A}_w^M$ is 1 if there is a path from q_1 to q_2 generating w, and is 0 otherwise.

We will assume, when we write any $\mathbf{A}_w^M$ in what follows, that the FSA M is supplemented with "sink states" as necessary to ensure that, for every $q_1 \in Q$ and every $x \in \Sigma$, there is at least one $q_2 \in Q$ such that $(q_1, x, q_2) \in \delta$. This ensures that, for any $w \in \Sigma^*$, there is at least one 1 on each row of $\mathbf{A}_w^M$, and therefore $\mathbf{A}_w^M \neq \mathbf{0}$.

We first define the relevant construction, then show below that it generates the desired intersection language. Without loss of generality, we assume that the FSA being intersected with the FSBM is ϵ-free.

Definition 9. *Given an FSBM $M_1 = \langle Q_1, \Sigma, I_1, F_1, G_1, H_1, \delta_1 \rangle$, and an FSA $M_2 = \langle Q_2, \Sigma, I_2, F_2, \delta_2 \rangle$, we define $M_1 \cap M_2$ to be the FSBM $\langle Q, \Sigma, I, F, G, H, \delta \rangle$, where*

- $Q = Q_1 \times Q_2 \times \{0,1\}^{|Q_2| \times |Q_2|}$

- $I = I_1 \times I_2 \times \{\boldsymbol{0}\}$

- $F = F_1 \times F_2 \times \{\boldsymbol{0}\}$

- $G = G_1 \times Q_2 \times \{\boldsymbol{A}^{M_2}_{\epsilon}\}$

- $H = H_1 \times Q_2 \times \{\boldsymbol{0}\}$

- $\delta = \delta_{\text{N}} \cup \delta_{\text{B}} \cup \delta_{\text{N}\to\text{B}} \cup \delta_{\text{B}\to\text{N}}$, *where*

  (a) $((q_1, q'_1, \boldsymbol{0}), x, (q_2, q'_2, \boldsymbol{0})) \in \delta_{\text{N}}$ *iff* $(q_1, x, q_2) \in \delta_1$ *with* $q_1 \notin G_1$ *and* $q_2 \notin H_1$, *and either*

  - $(q'_1, x, q'_2) \in \delta_2$, *or*
  - $x = \epsilon$ *and* $q'_1 = q'_2$.

  (b) $((q_1, q'_1, \boldsymbol{0}), \epsilon, (q_1, q'_1, \boldsymbol{A}^{M_2}_{\epsilon})) \in \delta_{\text{N}\to\text{B}}$ *iff* $q_1 \in G_1$

  (c) $((q_1, q'_1, \boldsymbol{A}), x, (q_2, q'_2, \boldsymbol{A}\boldsymbol{A}^{M_2}_{x})) \in \delta_{\text{B}}$ *iff* $\boldsymbol{A} \neq \boldsymbol{0}$ *and* $(q_1, x, q_2) \in \delta_1$ *with* $q_1 \notin H_1$ *and* $q_2 \notin G_1$, *and either*

  - $(q'_1, x, q'_2) \in \delta_2$, *or*
  - $x = \epsilon$ *and* $q'_1 = q'_2$.

  (d) $((q_1, q'_1, \boldsymbol{A}), \epsilon, (q_1, q'_2, \boldsymbol{0})) \in \delta_{\text{B}\to\text{N}}$ *iff* $q_1 \in H_1$ *and* $\boldsymbol{A} \neq \boldsymbol{0}$ *and the* (q'_1, q'_2) *entry of* $\boldsymbol{A}$ *is 1*

Notice that $|Q| = |Q_1| \times |Q_2| \times 2^{|Q_1| \times |Q_2|}$ is finite, since Q_1 and Q_2 are both finite.

The central challenge in setting up an FSBM to simulate the combination of an FSBM M_1 and an FSA M_2 is handling the effect on M_2 of $\vdash_{\text{B}\to\text{N}}$ transitions in M_1, where a string of arbitrary length is emptied from the buffer. Obviously the buffered string itself cannot be stored in the simulating FSBM's finite state. But, following an idea from Savitch (1989), any buffered string w determines a finite transition relation on the states of M_2, and it suffices to record this relation, which we encode in the form of the matrix $\mathbf{A}^{M_2}_{w}$.

The following lemma establishes the invariants that underpin the proof that this construction recognizes $L(M_1) \cap L(M_2)$.

Lemma 1. *Suppose a non-empty sequence of configurations $D_1, \dots D_m$ is the initial portion of a successful run (of any string) on an intersection FSBM $M = M_1 \cap M_2$, with each $D_i = (u_i, (q_i, q_i', \boldsymbol{A}_i), v_i, t_i)$. Then one of the following is true:*

(i) $t_i = \text{N}$ *and* $\boldsymbol{A}_i = \boldsymbol{0}$

(ii) $t_i = \text{N}$ *and* $(q_i, q_i', \boldsymbol{A}_i) \in (G_1 \times Q_2 \times \{\boldsymbol{A}_\epsilon^{M_2}\}) = G$

(iii) $t_i = \text{B}$ *and* $\boldsymbol{A}_i = \boldsymbol{A}_{v_i}^{M_2}$

(iv) $t_i = \text{B}$ *and* $(q_i, q_i', \boldsymbol{A}_i) \in (H_1 \times Q_2 \times \{\boldsymbol{0}\}) = H$

Proof. By induction on the length m of the sequence. If $m = 1$, then $t_m = \text{N}$ and $(q_m, q_m', \mathbf{A}_m) \in I = I_1 \times I_2 \times \{\mathbf{0}\}$, so $\mathbf{A}_m = \mathbf{0}$, satisfying (i). Now we consider a sequence $D_1 \dots D_m D_{m+1}$ where we assume that the requirement holds of D_m. Since $D_m \vdash_{M_1 \cap M_2} D_{m+1}$, there are four cases to consider.

- Suppose $D_m \vdash_{\text{N}} D_{m+1}$. Then $t_m = t_{m+1} = \text{N}$, $(q_m, q_m', \mathbf{A}_m) \notin G$, and $(q_{m+1}, q_{m+1}', \mathbf{A}_{m+1}) \notin H$. The inductive hypothesis therefore implies that $\mathbf{A}_m = \mathbf{0}$. Now there are four subcases, depending on the critical element of δ that licenses $D_m \vdash_{\text{N}} D_{m+1}$.
  - If the critical transition is in δ_{N}, then immediately $\mathbf{A}_{m+1} = \mathbf{0}$, satisfying (i).
  - If the critical transition is in $\delta_{\text{N}\to\text{B}}$, then $q_{m+1} \in G_1$ and $\mathbf{A}_{m+1} = \mathbf{A}_\epsilon^{M_2}$, satisfying (ii).
  - The critical transition cannot be in δ_{B}, since $\mathbf{A}_m = \mathbf{0}$.
  - The critical transition cannot be in $\delta_{\text{B}\to\text{N}}$, since $(q_{m+1}, q_{m+1}', \mathbf{A}_{m+1}) \notin H$ which implies that either $q_{m+1} \notin H_1$ or $\mathbf{A}_{m+1} \neq \mathbf{0}$.
- Suppose $D_m \vdash_{\text{N}\to\text{B}} D_{m+1}$. Then $t_m = \text{N}$, $t_{m+1} = \text{B}$, $v_m = v_{m+1} = \epsilon$, and $(q_m, q_m', \mathbf{A}_m) = (q_{m+1}, q_{m+1}', \mathbf{A}_{m+1}) \in G = G_1 \times Q_2 \times \{\mathbf{A}_\epsilon^{M_2}\}$. Therefore $\mathbf{A}_{m+1} = \mathbf{A}_\epsilon^{M_2} = \mathbf{A}_{v_{m+1}}^{M_2}$, satisfying (iii).

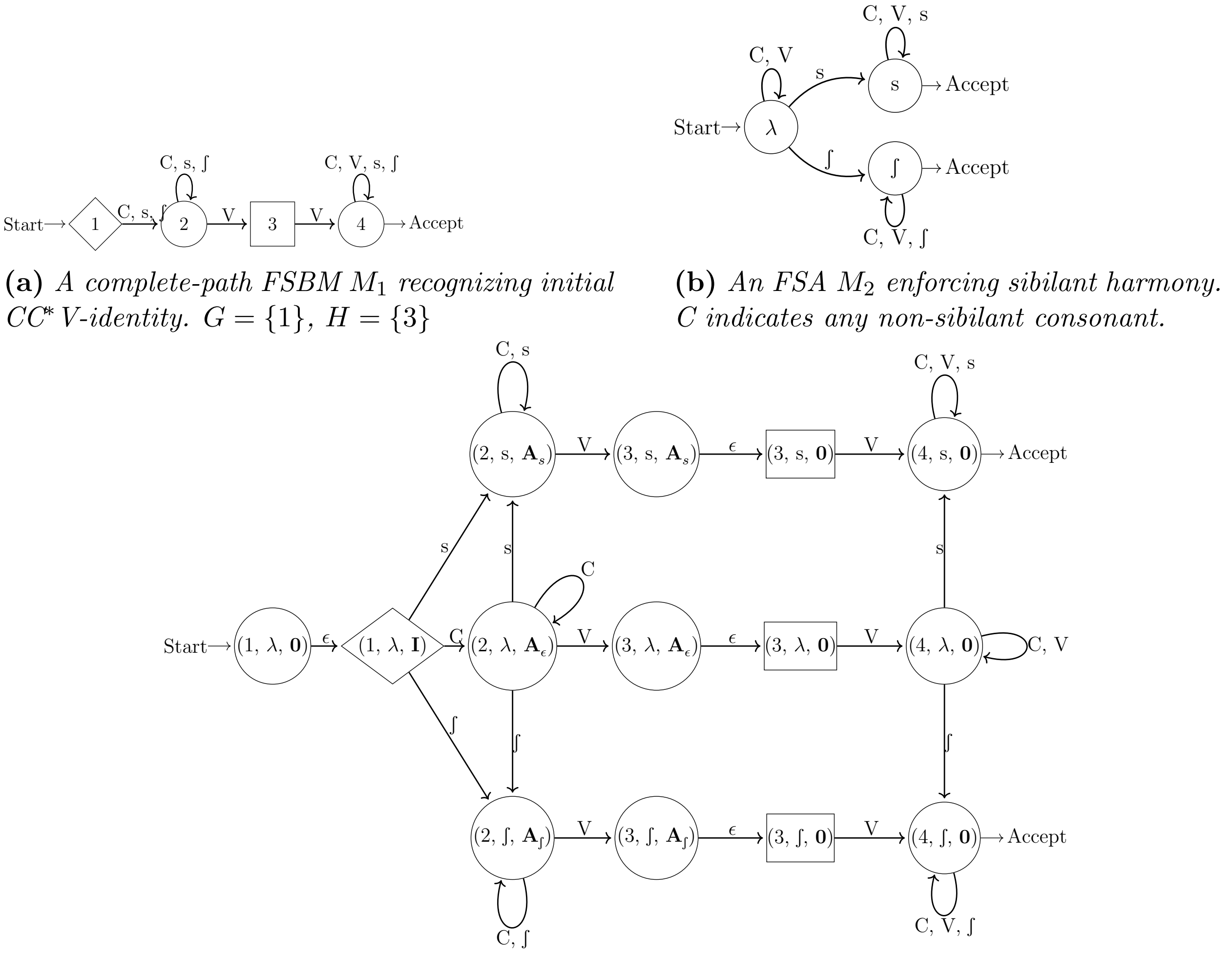
- Suppose $D_m \vdash_{\text{B}} D_{m+1}$. Then $t_m = t_{m+1} = \text{B}$, $(q_m, q'_m, \mathbf{A}_m) \notin H$, $(q_{m+1}, q'_{m+1}, \mathbf{A}_{m+1}) \notin G$, and $v_{m+1} = v_m x$ for some $x \in \Sigma \cup \{\epsilon\}$. The inductive hypothesis therefore implies that $\mathbf{A}_m = \mathbf{A}^{M_2}_{v_m}$. Now there are four subcases, depending on the critical element of δ that licenses $D_m \vdash_{\text{B}} D_{m+1}$.
  - The critical transition cannot be in δ_{N}, since $\mathbf{A}_m = \mathbf{A}^{M_2}_{v_m} \neq \mathbf{0}$.
  - The critical transition cannot be in $\delta_{\text{N}\to\text{B}}$, since $(q_{m+1}, q'_{m+1}, \mathbf{A}_{m+1}) \notin G$ which implies that either $q_{m+1} \notin G_1$ or $\mathbf{A}_{m+1} \neq \mathbf{A}^{M_2}_{\epsilon}$.
  - If the critical transition is in δ_{B}, then $\mathbf{A}_{m+1} = \mathbf{A}_m \mathbf{A}^{M_2}_x = \mathbf{A}^{M_2}_{v_m} \mathbf{A}^{M_2}_x = \mathbf{A}^{M_2}_{v_m x} = \mathbf{A}^{M_2}_{v_{m+1}}$, satisfying (iii).
  - If the critical transition is in $\delta_{\text{B}\to\text{N}}$, then $q_{m+1} \in H_1$ and $\mathbf{A}_{m+1} = \mathbf{0}$, satisfying (iv).
- Suppose $D_m \vdash_{\text{B}\to\text{N}} D_{m+1}$. Then $t_m = \text{B}$, $t_{m+1} = \text{N}$, $v_{m+1} = \epsilon$, and $(q_m, q'_m, \mathbf{A}_m) = (q_{m+1}, q'_{m+1}, \mathbf{A}_{m+1}) \in H = H_1 \times Q_2 \times \{\mathbf{0}\}$. Therefore $\mathbf{A}_{m+1} = \mathbf{0}$, satisfying (i).

□

This lemma establishes that the matrix component of the constructed machine's state tracks the information necessary to determine the appropriate "jump" to make through M_2 when a string is emptied from the buffer: in a $\delta_{\text{B}\to\text{N}}$ transition from $(q_1, q'_1, \mathbf{A})$ to $(q_1, q'_2, \mathbf{0})$, the base FSBM M_1 is in state $q_1 \in H_1$ and therefore leaves buffering mode, and the matrix $\mathbf{A}$ determines the appropriate states q'_2 for M_2 to jump to. The rest of the proof that $L(M_1 \cap M_2) = L(M_1) \cap L(M_2)$ is standard, but is provided in Appendix C.

An example demonstrating how the intersection works can be found in Figure 3.12. The FSBM in Figure 3.12a computes the language that shows initial CC*V-copying. The FSA in Figure 3.12b, adapted from Heinz (2007, p.38), encodes Navajo sibilant harmony (Sapir and Hoijer, 1967) on the feature [anterior], banning *s. . . ʃ and *ʃ. . . s sequences. The intersection FSBM is shown in Figure 3.12c, which recognizes the language of strings obeying both restrictions.

That FSBM-recognizable languages are closed under intersection with regular languages is an important step in clarifying the potential role of FSBMs for phonological theory. The



(a) *A complete-path FSBM M_1 recognizing initial CC^*V-identity. $G = \{1\}$, $H = \{3\}$*

(b) *An FSA M_2 enforcing sibilant harmony. C indicates any non-sibilant consonant.*

(c) *The intersection FSBM $M_1 \cap M_2$, ignoring states from which no accepting state is reachable. $\boldsymbol{A}_\epsilon$ is the M_2 transition matrix for any string without any s or ʃ (equal to $\boldsymbol{I}$); $\boldsymbol{A}_s$ is the transition matrix for all strings with at least one s and no ʃ; and $\boldsymbol{A}_ʃ$ is the transition matrix for all strings with at least one ʃ and no s*

Figure 3.12: *An example intersection construction*

overwhelming majority of phonotactic constraints that are not concerned with sub-string identity are regular (Heinz, 2018), and so any such constraint can be combined with an FSBM-enforcable identity constraint to yield another FSBM-recognizable language. In fact, since the regular languages are closed under intersection, FSBMs can also express the intersection of any *collection* of "normal" phonotactic constraints with any single FSBM-enforcable substring-identity constraint. In this way, FSBM provides a framework that can unify reduplication with other (morpho)phonological patterns.

An important issue that we leave open for future work is developing an algorithm for intersecting an FSA with an FSBM that assigns *weights* to strings expressing degrees of well-formedness. This kind of intersection algorithm has been used to implement the notion of competition between candidates from Optimality Theory (Smolensky and Prince, 1993), where violable constraints are expressed by weighted FSAs (Ellison, 1994; Eisner, 1997; Albro, 1998; Riggle, 2004b). Such an intersection algorithm for weighted FSBMs would allow for FSBM-defined reduplication constraints to be incorporated into implemented OT grammars. In other words, the point from the preceding paragraph might generalize beyond the special case of binary constraints which combine via simple intersection.

3.5.2 Closed under regular operations

Noticeably, given complete-path FSBMs are finite-state machines with a copying mechanism, most of the proof ideas in this subsection are similar to the standard proofs for FSAs, which can be found in Hopcroft and Ullman (1979) and **?**.

Theorem 3. *If L_1, L_2 are two FSBM-recognizable languages, then $L_1 \cup L_2$, $L_1 \circ L_2$ and L_1^* are also complete-path FSBM-recognizable languages.*

Proof. Assume there are complete-path FSBMs $M_1 = \langle \Sigma, Q_1, I_1, F_1, G_1, H_1, \delta_1 \rangle$ and $M_2 = \langle \Sigma, Q_2, I_2, F_2, G_2, H_2, \delta_2 \rangle$ such that $L(M_1) = L_1$ and $L(M_2) = L_2$, then ...

Union One can construct a new FSBM M that accepts an input w if either M_1 or M_2 accepts w. $M = \langle \Sigma, Q, I, F, G, H, \delta \rangle$ such that

- $Q = Q_1 \cup Q_2 \cup \{q_0\}$
- $I = \{q_0\}$
- $F = F_1 \cup F_2$
- $G = G_1 \cup G_2$
- $H = H_1 \cup H_2$
- $\delta = \delta_1 \cup \delta_2 \cup \{(q_0, \epsilon, q') \mid q' \in (I_1 \cup I_2)\}$

The construction of M keeps M_1 and M_2 unchanged, but adds a new state q_0. q_0 is the only initial state, branching into those previous initial states in M_1 and M_2 with ϵ-arcs. q_0 is a non-G, non-H plain state, so the constructed automaton is a complete-path FSBM.

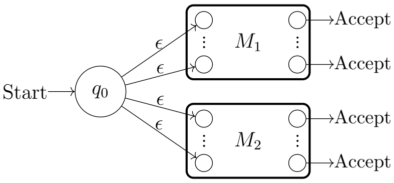


Figure 3.13: *The construction used in the union of two FSBMs*

Concatenation There is a complete-path FSBM M that can recognize $L_1 \circ L_2$ by the normal concatenation of two automata. The new machine $M = \langle \Sigma, Q, I, F, G, H, \delta \rangle$ satisfies $L(M) = L_1 \circ L_2$.

- Q $= Q_1 \cup Q_2 \cup \{q_0\}$
- I $= \{q_0\}$
- F $= F_2$
- $G = G_1 \cup G_2$
- $H = H_1 \cup H_2$
- $\delta = \delta_1 \cup \delta_2 \cup \{(p_f, \epsilon, q_i) \,|\, p_f \in F_1, q_i \in I_2\} \cup \{(q_0, \epsilon, p_i) \,|\, p_i \in I_1\}$

The new machine adds a new plain state q_0 and makes it the only initial state, branching into those previous initial states in M_1 ϵ-arcs. q_0 is not in H, nor in G. All final states in M_2 are the only final states in M. M also adds ϵ-arcs from all old final states in M_1 to all initial states in M_2.

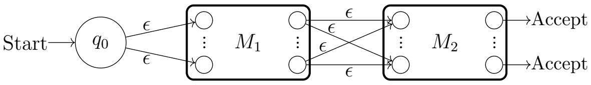
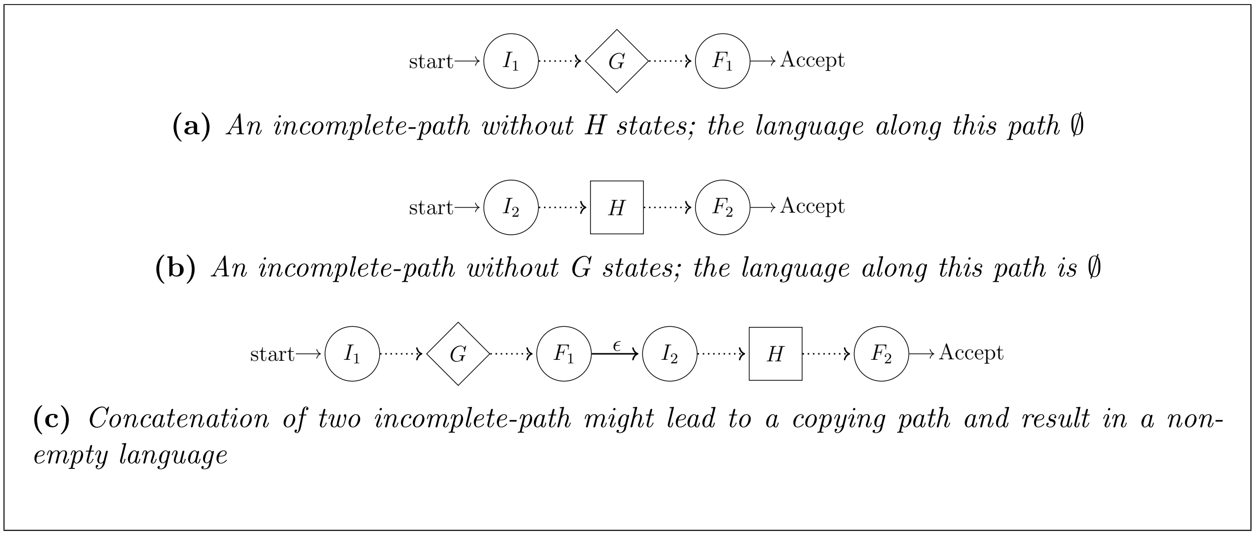


Figure 3.14: *The construction used in the concatenation of two FSBMs*

For this construction to work, it is important that we assume that M_1 and M_2 are complete-path FSBMs. Incomplete paths in two arbitrary machines might create a complete copying path, thus over-generating under the construction of concatenation mentioned here. For example, as illustrated in Figure 3.15, imagine one path in M_1 only has G states but no H states, and another path in M_2 contains only H states. They both recognize the empty language $L_{\emptyset} = \emptyset$. Therefore, the concatenation of these two languages should also be $L_{\emptyset}$. The assumption that M_1 and M_2 are complete-path FSBMs ensures that the construction has this result.



(a) *An incomplete-path without H states; the language along this path* $\emptyset$

(b) *An incomplete-path without G states; the language along this path is* $\emptyset$

(c) *Concatenation of two incomplete-path might lead to a copying path and result in a non-empty language*

Figure 3.15: *Problems arise in the concatenation of two incomplete paths. Dotted lines represent a sequence of normal transitions.*

Kleene Star $(L_1)^*$ is a complete-path FSBM-recognizable language. The new machine $M = \langle \Sigma, Q, I, F, G, H, \delta \rangle$ satisfies $L(M) = (L_1)^*$.

- $Q = Q_1 \cup \{q_0\}$
- $I = \{q_0\}$
- $F = F \cup \{q_0\}$
- $G = G_1$
- $H = H_1$
- $\delta = \delta_1 \cup \{(p_f, \epsilon, q_i) \,|\, p_f \in F_1, q_i \in I_1\} \cup \{(q_0, \epsilon, q_i) \,|\, q_i \in I_1\}$

M is similar to M_1 with a new initial state q_0. q_0 is also a final state, branching into old initial states in M_1. In this way, M accepts the empty string ϵ. q_0 is never a G state nor an H state. Moreover, to make sure M can jump back to an initial state after it hits a final state, ϵ transitions from any final state to any old initial states are added. Since all paths in M_1 are complete, concatenations of these paths do not overgenerate. □

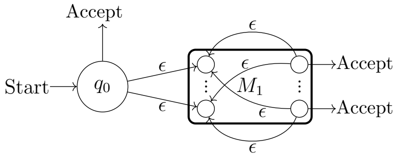


Figure 3.16: *The construction used in the star operation*

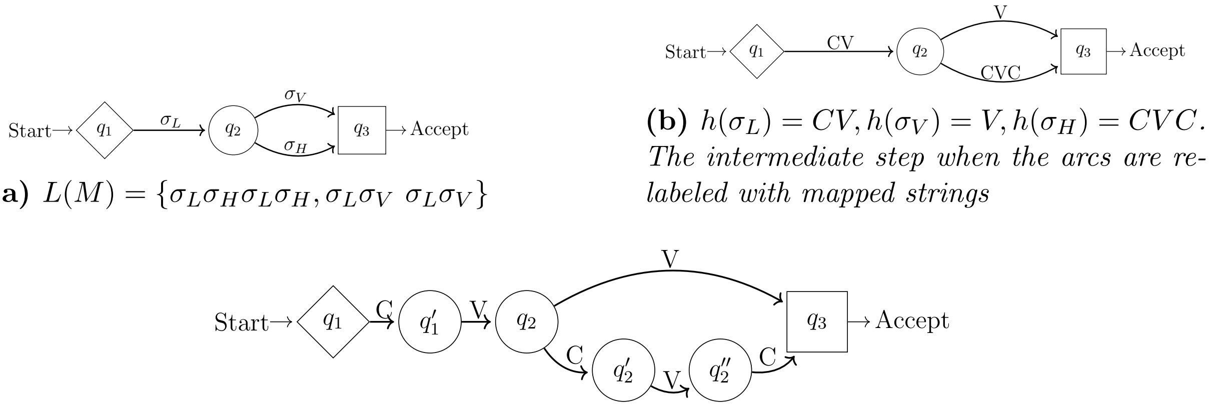
3.5.3 Closed under homomorphism

Theorem 4. *The class of languages recognized by complete-path FSBMs is closed under homomorphisms.*

Proof. We can construct a new machine M_h based on the base machine M, such that $L(M_h) = h(L(M))$. The construction goes as follows. Relabel each transition that emits x in M with the string $h(x)$, and add states to split the transitions so that there is only one symbol or ϵ on each arc in M_h. States added for this purpose are not included in G or H. all paths in M_h are complete since the construction does not affect the arrangements G and H states in paths. □

This construction is illustrated in Figure 3.17. The FSBM M uses the alphabet $\Sigma = \{\sigma_H, \sigma_L, \sigma_V\}$, and recognizes the finite language $\{\sigma_L\sigma_H\sigma_L\sigma_H, \sigma_L\sigma_V\sigma_L\sigma_V\}$. The constructed machine M_h recognizes the image of this finite language under the homomorphism $h : \Sigma^* \to \{C, V\}^*$ defined by $h(\sigma_L) = CV$, $h(\sigma_V) = V$, and $h(\sigma_H) = CVC$.

The fact that FSBMs are closed under homomorphism allows theorists to perform analyses at convenient levels of abstraction.



(a) $L(M) = \{\sigma_L\sigma_H\sigma_L\sigma_H, \sigma_L\sigma_V\ \sigma_L\sigma_V\}$

(b) $h(\sigma_L) = CV, h(\sigma_V) = V, h(\sigma_H) = CVC$. *The intermediate step when the arcs are relabeled with mapped strings*

(c) *States* q_1', q_2', q_2'' *are added to split the arcs.* $L(M_h) = \{CVVCVV, CVCVCCVCVC\}$

Figure 3.17: *Constructions used for the homomorphic language*

3.5.4 Not closed under intersection and complementation

Theorem 5. *The class of languages recognized by complete-path FSBMs is not closed under intersection, and thus not closed under complementation.*

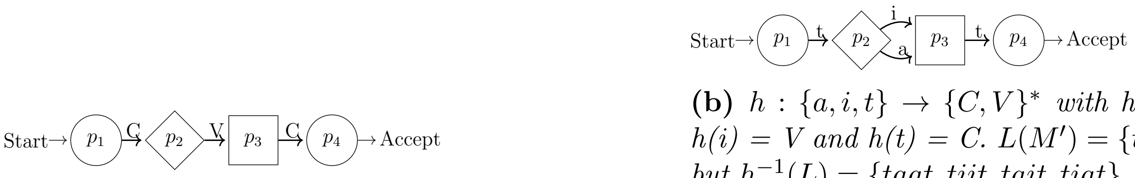
Proof. $L_1 = \{wwx \,|\, w, x \in a^*b\}$ and $L_2 = \{xww \,|\, w, x \in a^*b\}$ are FSBM-recognizable languages. However, $L_1 \cap L_2 = \{www \,|\, w \in a^*b\}$ is not an FSBM-recognizable language. Given FSBM is closed under union but is not closed under intersection, by De Morgan's law, FSBM is not closed under complementation. □

3.5.5 Not closed under inverse homomorphism

The class of languages recognized by complete-path FSBMs is closed under *one-to-one* alphabetic inverse homomorphism. One can directly relabel every mapped symbol in an FSBM to construct a new FSBM. But it is not closed under general inverse alphabetic homomorphisms and thus inverse homomorphism. Therefore, RCLs are not a trio.

Consider the complete-path FSBM-recognizable language $L = \{a^ib^ja^ib^j \ \,|\, i, j \geq 1\}$ (see Figure 3.8a), and an alphabetic homomorphism $h : \{0, 1, 2\}^* \to \{a, b\}^*$ such that $h(0) = a$, $h(1) = a$ and $h(2) = b$. Then, the inverse homomorphic image of L is $h^{-1}(L) = \{(0+1)^i2^j(0+1)^i2^j \,|\, i, j \geq 1\}$, which is not an RCL by Theorem 2.

Even though RCLs are not closed under inverse homomorphisms, analyzing exactly why this is not the case highlights something that distinguishes the languages of FSBMs from many other well-known language classes. The pivotal point comes from the one-to-many mapping. At first glance, one might try to apply the conventional construction for showing closure under inverse homomorphism of FSAs, i.e. build a new machine M', which reads any symbol x in the new alphabet and simulates M on $h(x)$, as shown in Figure 3.18.



(a) $L(M) = \{CVVC\}$

(b) $h : \{a, i, t\} \rightarrow \{C, V\}^*$ *with* $h(a) = V$, $h(i) = V$ *and* $h(t) = C$. $L(M') = \{taat, tiit\}$ *but* $h^{-1}(L) = \{taat, tiit, tait, tiat\}$

.

Figure 3.18: *The conventional construction of the inverse homomorphic image undergenerates*

But this construction fails to generate the full language $h^{-1}(L(M))$: the constructed machine M' still imposes an identity requirement, and therefore fails to accept strings such as 'tait' where the two occurrences of V are mapped by h^{-1} to distinct symbols. The application of an inverse homomorphism — unlike the application of a homomorphism — can "disrupt" sub-string identity relationships that the construction of a new FSBM will necessarily maintain.

3.5.6 An equivalent extension of regular expressions

The standard class of regular languages can be defined either via FSAs or via regular expressions. FSBMs constitute a minimal enrichment of FSAs that allow for copying. Here we present a corresponding way to enrich regular expressions that leads to the same class of languages as FSBMs. This provides an alternative characterization of the RCL class in terms of language-theoretic closure properties.

Definition 10. *Let Σ be an alphabet. The regular copying expressions (RCEs) over Σ and the languages they denote are defined as follows.*

- $\emptyset$ *is an RCE and* $\mathcal{L}(\emptyset) = \emptyset$

- ϵ *is an RCE and* $\mathcal{L}(\epsilon) = \{\epsilon\}$

- $\forall a \in \Sigma$, $\boldsymbol{a}$ *is an RCE and* $\mathcal{L}(\boldsymbol{a}) = \{a\}$

- *If* R_1 *and* R_2 *are RCEs, then* $R_1 + R_2$, R_1R_2, *and* R_1^* *are RCEs, and* $\mathcal{L}(R_1 + R_2) = \mathcal{L}(R_1) \cup \mathcal{L}(R_2)$, $\mathcal{L}(R_1R_2) = \{uv \mid u \in \mathcal{L}(R_1), v \in \mathcal{L}(R_2)\}$, *and* $\mathcal{L}(R_1^*) = (\mathcal{L}(R_1))^*$.

- *(new copying operator) If* R_1 *is a* ***regular*** *expression,* R_1^C *is an RCE and* $\mathcal{L}(R_1^C) = \{ww \,|\, w \in \mathcal{L}(R_1)\}$

RCEs introduce two modifications to regular expressions. First, a $\cdot^C$ expression operator for the copying-derived language is added. Then, the closure under other regular operations is extended to all RCEs. Therefore, languages denoted by regular copying expressions are closed under concatenation, union and Kleene star. Second, the copying operation is only granted access to regular expressions, namely to regular sets formed *without* the use of copying. In other words, the languages denoted by RCEs are not closed under copying, thus restricting the denoted languages by excluding w^{2^n}.

Given Σ^* is a regular language, an RCE for the simplest copying language $L_{ww} = \{ww \,|\, w \in \Sigma^*\}$ with $\Sigma = \{a, b\}$ would be $((a + b)^*)^C$. Assume $\Sigma = \{C, V\}$, a naive RCE describing Agta plurals after CVC-reduplication without considering the rest of the syllable structures could be $(CVC)^C(V + C)^*$. This denotes a regular language, unlike $((a + b)^*)^C$. Note, $((CVC)^C(V + C)^*)^C$ is not a regular copying expression, because recursive copying is prohibited, and the copying operator cannot apply to the expressions containing copying.

As noted in footnote 8, there are a number of definitions of "extended regular expressions" in the literature that incorporate some form of back-references (e.g. Câmpeanu et al., 2002; Câmpeanu et al., 2003; Carle and Narendran, 2009), and these motivated the development of Memory Automata (MFAs; Schmid, 2016; Freydenberger and Schmid, 2019). Just as FSBMs can be seen as a restricted special case of MFAs, RCEs correspond to a special case of extended regular expressions: essentially, an RCE of the form R^C is equivalent to $(R)\backslash 1$, where the back-reference necessarily immediately follows the captured group.

For further details of the equivalence of RCEs and FSBMs, see Appendix D.

3.6 Further formal issues

3.6.1 Determinism

A natural question to consider is whether the non-determinism that we have allowed in FSBMs is essential.[12] A proper treatment of this issue turns out to be more subtle than it might initially appear, but we offer some initial observations here.

The FSBM in Figure 3.19 is non-deterministic in the sense that the string *aa* might lead the machine either to q_2 or to q_3. This familiar kind of non-determinism brings no additional expressive power in the case of standard FSAs, where the subset construction can be used to determinize any FSA. But this method for determinization cannot be straightforwardly applied to FSBMs, because of the distinguished status of G and H states. Applying the construction to the FSBM in Figure 3.19 would yield a new state corresponding to $\{q_2, q_3\}$, and then the question arises of whether this new state should be an H state (like q_3) or not (like q_2). Neither answer is sufficient: in the new machine, the string *aa* will deterministically lead to the state $\{q_2, q_3\}$, but the prefix *aa* may or may not be the entire string that needs to be buffered and copied.

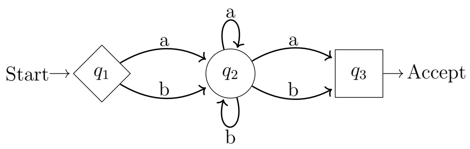


Figure 3.19: *An FSBM illustrating nondeterminism*

Stated slightly more generally, the subset construction can eliminate non-determinism *between states* ("state-nondeterminism"), but in FSBMs there is also the possibility of non-determinism *between modes* ("mode-nondeterminism"). The state-nondeterminism indicated in (33) could be eliminated, in a sense, by applying the subset construction to yield a new machine M' with transitions as in (34).

(33) a. $(aa\ldots, q_1, \text{N}, \epsilon) \vdash_M^* (\ldots, q_2, \text{B}, aa)$

[12] Thanks to two anonymous reviewers for drawing our attention to this issue.

b. $(aa\ldots, q_1, \textsc{n}, \epsilon) \vdash_M^* (\ldots, q_3, \textsc{b}, aa)$

(34) $(aa\ldots, \{q_1\}, \textsc{n}, \epsilon) \vdash_{M'}^* (\ldots, \{q_2, q_3\}, \textsc{b}, aa)$

But the two configurations reached in (33) differ in whether M will stop buffering after this prefix aa, and we suspect that there is no way to eliminate this kind of nondeterminism between modes. To bring out this important additional distinction, consider the transition sequences in (35) for the longer prefix $aaaa$.

(35) a. $(aaaa\ldots, q_1, \textsc{n}, \epsilon) \vdash_M^* (aa\ldots, q_2, \textsc{b}, aa) \vdash_M^* (\ldots, q_2, \textsc{b}, aaaa)$

b. $(aaaa\ldots, q_1, \textsc{n}, \epsilon) \vdash_M^* (aa\ldots, q_3, \textsc{b}, aa) \vdash_M (\ldots, q_3, \textsc{n}, \epsilon)$

So there is something distinctive about the kind of non-determinism in Figure 3.19, which lies not in the fact that the prefix aa might lead to either state q_2 or state q_3, but rather the fact that the prefix $aaaa$ might lead to either state q_2 in mode B, or state q_3 in mode N.

The following definition makes a first attempt at pinpointing the distinctive kind of non-determinism in Figure 3.19.

Definition 11. *An FSBM M is mode-deterministic if there do not exist three configurations $C = (w, q, m, v)$, $C_1 = (\epsilon, q_1, m_1, v_1)$ and $C_2 = (\epsilon, q_2, m_2, v_2)$, such that*

- $C \vdash_M^* C_1$ *and* $C \vdash_M^* C_2$,
- $C_1 \not\vdash_M^* C_2$ *and* $C_2 \not\vdash_M^* C_1$, *and*
- $m_1 \neq m_2$.

The FSBM in Figure 3.20, for example, is mode-deterministic in this sense, whereas (35) demonstrates that the FSBM in Figure 3.19 is not. We conjecture that the mode-deterministic FSBMs are properly less powerful than the full class of FSBMs, and in particular that there is no mode-deterministic FSBM that generates the same language as the FSBM in Figure 3.19.

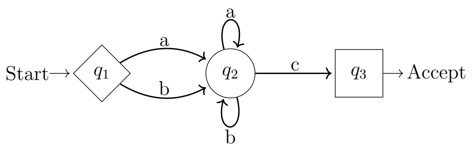


Figure 3.20: *An FSBM illustrating mode-determinism*

3.6.2 The role of symbol identity

A noteworthy trait of the RCL class is its non-closure under inverse homomorphisms. This distinguishes the RCL class from many of the familiar language classes that have played a role in the analysis of natural languages: the regular class and the context-free class are each closed under both homomorphisms and inverse homomorphisms, as are prominent classes in the "mildly context-sensitive" region, such as the tree-adjoining languages and multiple context-free languages (Joshi, 1985; Kallmeyer, 2010).

To illustrate, consider the relationship between the following two languages:

$$L_1 = (\mathrm{a} + \mathrm{b})^i \mathrm{c}^j (\mathrm{a} + \mathrm{b})^i \mathrm{c}^j$$
$$L_2 = \mathrm{a}^i \mathrm{c}^j \mathrm{a}^i \mathrm{c}^j$$

We showed above that L_1 is not an RCL, whereas L_2 obviously is. This sets the RCL class apart from the regular and context-free classes, which contain neither L_1 nor L_2, and from the tree-adjoining and multiple context-free classes, which contain both; recall Figure 3.5. For all these other formalisms, the surface differences between L_1 and L_2 are essentially irrelevant. For example, a multiple context-free grammar (MCFG; Seki et al., 1991; Kallmeyer, 2010) for L_1 is given in (36), and (37) shows an illustrative derivation for the string 'abcaac'. This grammar uses the nonterminals P and Q to control the assembly of (discontinuous) '$(\mathrm{a}+\mathrm{b})^i \ldots (\mathrm{a}+\mathrm{b})^i$' and '$\mathrm{c}^j \ldots \mathrm{c}^j$' portions respectively; P-portions can grow via the addition of X elements, and Q-portions can grow via the addition of Y elements.

(36) $S(u_1 v_1 u_2 v_2) \rightarrow P(u_1, u_2)\ Q(v_1, v_2)$

$P(\epsilon, \epsilon)$

$P(u_1 v, u_2 w) \rightarrow P(u_1, u_2)\ X(v)\ X(w)$

$Q(\epsilon, \epsilon)$

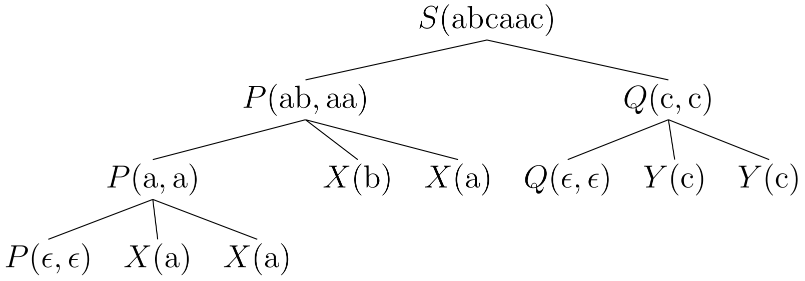
$Q(u_1 v, u_2 w) \rightarrow Q(u_1, u_2)\ Y(v)\ Y(w)$

$X(\mathrm{a})$

$X(\mathrm{b})$

$Y(\mathrm{c})$

(37)



Notice that to generate L_2 instead of L_1, we would simply omit the rule $X(\mathrm{b})$ from (36). What this highlights is that for either L_1 or L_2, the significant work is done by the rules that arrange the yields of the nonterminals X and Y appropriately, and this work can be dissociated from the rules that specify the terminal symbols that can appear as the yields of X and Y. The nonterminals provide a grammar-internal mechanism for doing the book-keeping necessary to enforce the abstract pattern shared by L_1 and L_2, and the relationship between these grammar-internal symbols and the terminal symbols that make up the generated strings is opaque.

In an FSBM, on the other hand, the machinery that extends the formalism beyond the regular languages has no analogous grammar-internal book-keeping mechanism that can be dissociated from surface symbols: the non-regular effects of an FSBM's string-buffering mechanism are inherently tied to the identity of certain surface symbols. This is what underlies the crucial difference between L_1 and L_2 for FSBMs, and the non-closure under inverse homomorphisms of RCLs.[13]

[13] Of course the states of an FSBM are grammar-internal symbols in the relevant sense, and this is in effect what allows FSAs to be closed under both homomorphisms and inverse homomorphisms. But the point of the discussion here is to look at the distinctive additional capacities of FSBMs, which are brought out by considering a non-regular language such as L_2.

A comparison with Savitch's RPDAs (discussed above; Savitch, 1989) is informative: RPDAs, while similar in some respects to FSBMs, generate a class of languages that *is* closed under both inverse homomorphism and homomorphism (in fact, under any finite-state transduction). This difference stems from the fact that

To put a label on this distinction, we might say that FSBMs are "symbol-oriented" (where by "symbol" we mean surface/terminal symbol), in contrast to the other formalisms mentioned above. Suppose, to make this precise, we say that a formalism (or a language class) is **symbol-oriented** if and only if it fails to be closed under both homomorphisms and inverse homomorphisms.

It is interesting to note that, while the symbol-oriented nature of FSBMs sets them apart from formalisms (such as MCFGs) motivated by the kinds of non-context-free cross-serial dependencies observed in syntax, this property of FSBMs is shared by other formalisms that have been argued to align well with observed phonological patterns. Many of the sub-regular language classes discussed by Heinz (2007), are also symbol-oriented in this sense. An easy example (Mayer and Major, 2018; De Santo and Graf, 2019) comes from the Strictly 2-Local (SL_2) languages: $(ab)^*$ is an SL_2 language, but applying the homomorphism h defined by $h(a) = c, h(b) = c$ yields $(cc)^*$, which is not an SL_2 language. So the SL_2 languages are not closed under homomorphisms.

The fact that the SL languages lack closure under homomorphisms, whereas the RCL class lacks closure under *inverse* homomorphisms, reflects the different role that symbol identity plays for the two formalisms. The move from $(ab)^*$ to $(cc)^*$ eliminates distinctions between surface symbols, which removes information that the SL_2 grammar for $(ab)^*$ was using to ensure that the length of each generated string was even. The move from L_2 to L_1, on the other hand, *introduces* distinctions between surface symbols which are incompatible with the string-buffering mechanism of an FSBM.[14]

But the broader point we wish to draw attention to here is the distinction between (i)-the context-free class and various mildly context-sensitive classes, which are closed under both homomorphisms and inverse homomorphisms, and (ii)-the RCL and SL classes, which are

an RPDA's queue-like memory arises from relaxing restrictions on a standard PDA's stack, and so the queue-like memory uses a distinct alphabet of "stack symbols" rather than surface symbols. These stack symbols are grammar-internal book-keeping devices whose relationship to surface symbols can be specified by the grammar-writer, as in the case of MCFGs such as (36) above.

[14] For similar reasons, the languages of regular expressions extended with back-references are also not closed under inverse homomorphism (Câmpeanu et al., 2003).

not and therefore exhibit a degree of sensitivity to surface symbol identity. It is intriguing that the insensitivity to surface symbol identity seems to be necessary for many important patterns found in natural language syntax — for example, the classic cross-serial dependencies in Swiss German (Shieber, 1985) correspond to $\mathrm{a}^i\mathrm{b}^j\mathrm{c}^i\mathrm{d}^j$, rather than $\mathrm{a}^i\mathrm{b}^j\mathrm{a}^i\mathrm{b}^j$ — whereas many phonological patterns that have been studied computationally are compatible with symbol-oriented formalisms. This includes both the sub-regular patterns that motivate formalisms such as SL grammars, and the non-regular reduplication patterns that motivate FSBMs.

A complication to this clear picture may come from copying patterns in syntax, for example the Yoruba constructions discussed by Kobele (2006), mentioned above in Section 2.3. The languages generated by parallel multiple context-free grammars (PMCFGs) are not closed under inverse homomorphisms (Nishida and Seki, 2000, p.145, Corollary 12), for reasons analogous to what we have seen for FSBMs, and so this is an example of a symbol-oriented formalism that has been argued to be appropriate for syntax. But it is clear that syntax requires at least some *non*-symbol-oriented mechanisms to generate the well-known cross-serial dependencies of the Swiss-German sort ($\mathrm{a}^i\mathrm{b}^j\mathrm{c}^i\mathrm{d}^j$), whereas those cross-serial dependencies that we do observe in phonology are compatible with the more restricted, symbol-oriented notion of cross-serial dependencies that appear in reduplication.

3.6.3 Representations matter: Segments or more abstract prosodic units?

At this point, readers might wonder why FSBMs work at a segmental level, but not at some higher levels, such as morphemic representations or more coarse-grained phonological representations, as we have kept emphasizing in Chapter 2.[15] Another related question is on how FSBMs could connect to these coarse-grained prosodic units. For these questions, we would like to offer the following opinions.

First, representations certainly matter. In fact, we believe there is substantial evidence

[15] Thanks to Jon Brennan, Aniello De Santo, Ben Eischens, and Jeff Heath for their questions along this line.

supporting the existence of a more abstract level of representation than segments for reduplicated strings – it helps leverage the *final* representational burden, making reduplication "easy," intuitive, and cognitively salient. Yet the choice of the representational level to start our investigations should not be the level based on which humans represent reduplicated string, but should also depend on the types of computation involved in reaching the final representation. In the case of proposing a *recognizer* for reduplicative structures, the computation always involves a matching step on whether some phonological objects are identical to other phonological objects. With this in mind, the representations to start our investigations with should be the lowest level X on which phonological identity is atomic and properly defined. The identity of any more abstract levels is built on the identity at X, and the identity of any more abstract levels naturally entails that everything at X is identical. In the case of (morpho)phonology, segments are usually assumed to be that level X (e.g., Chomsky and Halle, 1968, p.5; Base Reduplicant Correspondence Theory, McCarthy and Prince, 1995). In other words, to simply recognize the identity relations within reduplicative strings, there would always be a non-trivial step to check whether a bunch of segments are identical to another bunch of segments respectively. Indeed, we also find from our experimental results that participants extrapolated to different generalizations when the segmental level identity did not hold. Hence, the formal investigations in this chapter are necessary to suffice as a computational-level recognizer of the reduplicative strings.

How can FSBMs incorporate prosodic units? In our view, addressing this question will lead us one step further to propose an algorithmic explanation of computing reduplicative strings. For a satisfying answer, two aspects should be made more precise: first, the data structure of the "buffer" should go beyond the mere first-in-first-out, sequential organization of the surface segments; second, the exact procedures of the buffer-input interactions should be more spelled out. The prosodic units provide cutting points for possible solutions to the desiderata. These coarse-grained units could function as a way to organize sequential segments into more structured elements. Then, the "buffer" could not only store symbols, but also actively parse the segmental sequences into syllables, feet, and prosodic words. Moreover, the final emptying step could opt for a representational level that maximally leverages the encodings of the segmental identity relations. This works because the logical implications hold. That is, when a foot Ft_1 is identical to another foot Ft_2, all the syllables in Ft_1 must be identical to Ft_2. Similarly, that a syllable σ_1 is identical to another syllable σ_2 already entails that the segments within these syllables are identical. To make such an idea concreate, we think the string representation in the early works of formalizing finite-state optimality theory, such as Eisner (1997) and Albro (2005), and/or the tree representation as in Yu (2021), could be ultimately useful. It should be noted, again, that no matter what incarnations these detailed components are, the set of phonological strings they can compute is expected to be extensionally equivalent to the versions of FSBMs as proposed in this chapter. That said, the examined formal properties at the level of surface strings still hold, which is one of the main goals of this chapter.

3.7 Variants of FSBM for the typology of reduplication patterns

Reduplicative typology is much richer than the mere repetitions of the two copies. In this section, we briefly consider some more complicated kinds of reduplication that are beyond the capacity of the FSBMs as formulated here. We sketch some possible ways in which FSBMs might provide a starting point for future work that aims for a proper treatment of the full range of natural language reduplication phenomena. Through our discussions of these potential modifications, we would like to highlight the relevant dimensions of computation that reflect the dimensions along which the typology varies.

Non-local Reduplication Non-local reduplication is the case when the surface phonological strings have non-adjacent copies, incurring non-local correspondence among symbols.[16] A more comprehensive typology and linguistic analysis on non-local reduplication can be found in Riggle (2004a). Examples from Creek are shown in Table 3.7.

[16] Bambara 'Noun o Noun' illustrates a particularly simple kind of non-local reduplication where the intervening string is *always* the fixed string 'o'. This could be relatively easily handled by specifying a fixed string to each H state, to be inserted between the two copies when the buffer is emptied. The examples discussed in the main text are when the intervening elements are variable, different from the Bambara-like examples in important ways.

Non-local reduplication		
Creek plural		
Gloss	*Singular*	*plural*
‘precious’	a-cáːk-iː	a-**cáːca**k-íː
‘clean’	hasátk-iː	**ha**sat**ha**k-íː
‘soft’	lowáck-iː	**lo**wac**lo**k-iː

Table 3.7: *Creek plural; CV-copying placed before the final consonant of the root (Booker, 1979; Riggle, 2004a)*

Marantz (1982) described the adjacency between the reduplicant and the base as a general typological trend. There were proposals (e.g., Nelson, 2005) arguing the inviolability of Marantz's generalization, either classifying some patterns as non-reduplicative copying but due to drives to satisfy templates, or suggesting that copying is still local and deletion motivated by other phonological constraints complicates the situation. Riggle (2004a) used the Creek words in Table 3.7 to argue for true non-local correspondence relations.

FSBMs' current limitation to local reduplication comes from the requirement that B-mode computation has to be directly followed by the buffer-emptying process, and a filled buffer is not allowed in N mode. A possible modification to allow non-local reduplication would be to allow the buffer to be filled in N mode and encode such a possibility in another kind of special states, say J, which stops the machine from buffering, with the buffer only being matched against input and emptied when an H state is encountered. The transitions leading from a G state to a J state would consume symbols in the input tape and buffer symbols in the queue-like buffer. Then, if there is no adjacent H following the end of buffering, the machine can use plain transitions to plain states for only input symbols. The buffer with symbols in it should be kept unchanged. Ultimately, the machine has to encounter some H states to empty the buffer to accept the string, since no final configuration allows symbols on the buffer.

Such a modification might not affect much of the proof ideas of the theorems constructed so far. Regarding the pumping lemma, Condition 2 can be modified by including a sub-string of intervening segments in between two copies. That is, $w \in L$ with $w > 5k$ can be rewritten as $w = ux_1x_2x_3\mathbf{y}x_1x_2x_3v$ such that $\forall i \in \mathbb{N}, ux_1x_2^ix_3\mathbf{y}x_1x_2^ix_3v \in L$. It is worth pointing out that if the generalization in Creek is productive, the sub-string of intervening segments between copies could be unboundedly long.

Multiple Reduplication Here, multiple reduplication refers to the cases when two or more different reduplicative patterns appear in one word. One string can have multiple substrings identical to each other. Examples from Nlaka'pamux (previously known as Thompson; Salishan), are listed in Table 3.8. See Zimmermann (2019) for a complete typological survey and classification.

Multiple reduplication Nlaka'pamux (Broselow, 1983, p. 329)	
Gloss	Strings
calico	sil
DIM-calico	sí-sil
DIST-calico	sil-síl
DIST-DIM-calico	sil-sí-sil

Table 3.8: *Multiple reduplication in Nlaka'pamux*

While the computational nature of multiple reduplication in natural language phonology and morphology remains an open question,[17] FSBMs could be relatively easily modified to include multiple copies of the same base form ($\{w^n \mid w \in \Sigma^*, n \in \mathbb{N}\}$), where n might be tied to the number of copying operations in a language. Given a natural number n, an appropriate modification of FSBMs might allow for the buffered symbols to not be emptied until they have been matched n times against the input.

On the other hand, FSBMs cannot be easily modified to recognize the language $\{w^{2^n} \mid w \in \Sigma^*, n \in \mathbb{N}\}$, where ww strings are themselves copied (i.e. $\{w, ww, wwww, \dots\}$, excluding www).

It is worth carefully distinguishing between the sense of "copying" instantiated by ww and w^n on the one hand, and the sense instantiated by $w^{(2^n)}$ on the other. The former sense highlights the fact that certain portions of a string are identical to certain other portions,

[17] For recent phonological analyses, see Zimmermann (2021b) and Zimmermann (2021a). For a more detailed discussion on the string-to-string function version of this problem, see Rawski et al. (2023).

whereas the latter is a natural interpretation of the idea that there is a copying *operation* that can apply to *its own outputs*. The kind of recursive copying exhibited by $w^{(2^n)}$ means that this language does not have the constant growth property that Joshi (1985) identified as a criterion for mild context-sensitivity. Excluding this recursive copying from phonology seems relatively well-justified, on the grounds that triplication is attested (Zimmermann, 2019; Rawski et al., 2023). But the situation may be different for syntax, where Kobele (2006), for example, has argued for recursive copying of the $w^{(2^n)}$ sort on the basis of Yoruba relativized predicates. See also Clark and Yoshinaka (2014) on the relationship between parallel multiple context-free grammars (PMCFGs) and multiple context-free grammars (MCFGs); and Stabler (2004) on the comparison between what he calls *generating grammars* and *copying grammars*.

Reduplication with non-identical copies In natural languages, non-identical copies are prevalent. One type of non-identical copies involves a fixed, memorized segment/sub-string (Alderete et al., 1999). Examples are given in Mongolian, illustrated in Table 3.9, where whole stems are copied to create forms with the meaning 'X and such things'. However, the initial consonant is always rewritten as [m].[18]

Non-identical copies
Mongolian Noun Reduplication (Svantesson et al., 2005, pp. 60)

Gloss	*root*	*X and such things*
'gown'	teeɮ	teeɮ-meeɮ
'beard'	t^{h}aɮx	t^{h}aɮx-m^{h}aɮx
'eye'	nut	nut-mut

Table 3.9: *Non-identical copies in Mongolian*

The Mongolian case is a special instance of imperfect copying that involves fixed phonological materials. Other types of imperfect copying can be attributed to the interaction between reduplication and other (morpho)phonological processes, which usually involve vowel

[18] When the stem form starts with [m], it is always rewritten to [c]. For example, the reduplicated form of [maɮ] 'cattle' is [maɮ-caɮ]

reduction as in Table 3.10, and onset cluster simplification in Tagalog partial reduplication (Zuraw, 1996), e.g. *'X is working'* [nag-ta-tɾabahoh], mapped from [tɾabahoh].

Non-identical copies			
Paluan reduplication with vowel reduction (Zuraw, 2003, p. 8)			
unreduplicated	*Meaning*	*Reduplicated*	*Meaning*
tóɾð	'frustration'	bəkə-təɾ-tóɾð	'easily frustrated'
síkt^{h}	'cluster of fruit'	mə-sək-síkt^{h}	'covered with fruit'

Table 3.10: *Non-identical copies in Paluan*

One way to modify FSBMs to accommodate non-identical copies would be to allow the machine to either store or empty not exactly the same input symbols, but the image of the input symbols under some finite-state transduction, f. For example, to account for the fixed consonant in Mongolian, we can introduce a finite state transduction $f_{C_1\to m}$ that rewrites the first consonant to [m]. To empty the buffer, instead of checking the identity relation, it determines whether $f_{C_1\to m}(x) = y$ where x is in the buffer and y is a prefix of the remaining input. Regular vowel reduction patterns could also be encoded in a similar manner through a transduction $f_{reduce}(y) = x$ where x is the full sequences on the buffer and y is a prefix of the remaining input. For instance, we can enforce that the buffered material təɾ could only be emptied if $f_{reduce}(\text{tóɾ}) = \text{təɾ}$.

From results in Experiment 1b and 1c discussed in Chapter 2, we think the naturally occurring possible transductions exhibit a particular property that aligns with the motivation for proposing the current mechanics of FSBM with a one-fell-swoop matching step, as a direct comparison to the symbol-by-symbol emptying steps as in Wang (2021a) and Wang (2021b), though they are extensionally equivalent. That is, the domain of the transduction should be the whole copied sequence. In these two experiments, human learners were prompted with imperfect copying targeting a specific position within a copy, for example, [dɔv.gə] and [div-dɔv.gə], as well as [dɔv.gə] and [dəv-dɔv.gə]. Beyond responses that copied the onset and coda perfectly, we also observed that the encoding of copying in the other locations was affected. Using one of the novel stems [dɹap.moʊ] as an example, there was a fair amount of onset cluster simplification as in [dip-dɹap.mə], as well as the non-incorporation of the coda as in [dɹi-dɹap.moʊ]. If the finite-state transduction only targets segments at a particularly specified position, then, this kind of non-faithful copying observed at the other locations would be left unexplained. Hence, from this perspective, it appears that the option to match the buffered sequence with the rest of the sequence provides a more intuitive and naturalistic way for the actual implementation.

If no restrictions at all are imposed on the transduction, then the modified automata would recognize the context-free $\{a^n b^n \mid n \in \mathbb{N}\}$ with $f(a) = b$ in a manner that (unlike a context-free grammar) associates the first 'a' with the first 'b' and so on, though still excluding string reversals. Moreover, the resulting language set would also include $\{a^i b^j c^i d^j \mid i, j \geq 1\}$ with $f(a) = c, f(b) = d$. It could be fruitful for further studies to examine possible restrictions on the transduction.

Boundary-markers and the linguistic relevance of mode-determinism Determining the mode switches in finite-state buffered machines is equivalent to determining the boundary of two copies. Subsequently, what underpins the notion of mode-determinism is that specific surface properties of a prefix provide enough cues to signal boundaries between two copies. Mode-deterministic patterns have some intuitive natural language correlates. If a language requires a copy to be of a fixed shape, then such a pattern would belong to the mode-deterministic class. Additionally, any reduplicative patterns with fixed phonological material at the boundaries perfectly lie within the mode-deterministic sub-class. This directly corresponds to the Bambara 'Noun o Noun' case (as in (29)-(31)) and echo reduplication, a special kind of total reduplication with systematic phonological changes. These changes often target the boundaries within two copies, occupying the beginning of the second copy (McCarthy and Prince, 1986, p.67; Inkelas and Downing, 2015, pp. 509-510). The fixed [m] in Mongolian noun reduplication in Table 3.9 provides an illustrative example (e.g., [teeɮ-**m**eeɮ], [nut-**m**ut]). The kind of fixed phonological materials aligns nicely with the notion of mode-determinism, with the caveat that empirically the fixed phonological materials usually appear once.

The extent to which attested reduplication patterns are mode-deterministic remains an empirical question. However, the well-attestedness of fixed phonological materials at the boundary, particularly in total reduplication, does not seem to be a coincidence. The boundary-signaling property of these fixed phonological materials could, in principle, aid in parsing reduplicative strings, especially those of unbounded length. Hence, this property might make these echo-reduplicated words easier to recognize, learn, and transmit.

3.8 Conclusion and future research

This chapter has looked at formal computational properties of unbounded copying on regular languages, including the simplest copying language L_{ww} where w can be any arbitrary string over an alphabet. We have proposed a new computational device: finite-state buffered machines (FSBMs), which add copying to regular languages by adding an unbounded queue-structured memory buffer, with specified states restricting how this memory buffer is used. As a result, we introduce a new class of languages, which is incomparable to context-free languages, named regular copying languages (RCLs). This class of languages extends regular languages with *unbounded* copying but excludes non-reduplicative non-regular patterns. Context-free string reversals are excluded since the buffer is queue-like, and the mildly context-sensitive Swiss-German cross-serial dependency pattern, abstracted as $\{a^ib^jc^id^j \mid i, j \geq 1\}$, is also excluded, since the buffer works on the same alphabet as the input tape and only matches *identical* sub-strings.

We have also surveyed the class's closure properties and proved a pumping lemma. This language set is closed under union, concatenation, Kleene Star, homomorphism, and intersection with regular languages. It is not closed under copying, inhibiting the recursive application of copying and excluding non-semilinear $w^{(2^n)}$. This class is also not closed under intersection, nor complementation. Finally, it is not closed under inverse homomorphism, given it cannot recover the possibility of non-identity among corresponding segments when the mapping is many-to-one (and the inverse homomorphic image is one-to-many); we suggested that this might reflect an important difference between the string-generating mechanisms of phonology and syntax.

One potential direction for future research is to connect FSBMs with the 2-way D-FSTs studied by Dolatian and Heinz (2018a,b, 2019, 2020), which model unbounded copying as *function*s while excluding mirror image mappings. We briefly mention two possibilities along these lines. First, it will be interesting to compare the RCL class of languages with the image of the functions studied by Dolatian and Heinz (2020). A second possibility is to consider adding to FSBMs another tape for output strings, extending from acceptors (as presented here) to finite-state buffered transducers (FSBTs). The morphological analysis ($ww \mapsto w$) problem is claimed to be difficult for 2-way D-FSTs, since they are not invertible. Our intuition is that FSBTs might help solve this issue: after reading the first w in input and buffering this string in memory, the machine can write ϵ to the output tape when it matches the buffered string against the contents of the input tape. A more detailed and rigorous study is desirable in this direction.

The learning and learnability of FSBMs and copying in sub-regular phonology is also crucial. The RCL class itself cannot be identified in the limit, since it properly contains the regular class (Gold, 1967). However, we take positive learning results from Clark and Yoshinaka (2014) and Clark et al. (2016) on PMCFGs with copying, and from Dolatian and Heinz (2018b) on Concatenated Output Strictly Local functions for reduplication, as suggestions for future directions towards learning results for FSBMs. In particular, one of the most attractive properties of the sub-regular classes is their Gold-learnability (e.g. Garcia et al., 1990; Heinz, 2010; Chandlee et al., 2014; Jardine and Heinz, 2016). We hope to explore whether the learnability property remains once copying is added to these sub-regular classes.

The current class of languages excludes non-adjacent copies, multiple reduplication, and reduplication with non-identical copies. We briefly sketched possible modifications and formal effects. We hope that our proposal provides a useful framework for better understanding the formal issues raised by these more complex reduplication phenomena and guiding empirical research into their typology, processing, and learning. On a different note, one aspect we did not discuss regards efficiency and the pratical utility. We are optimistic that investigating their practical utility and developing more useful toolkits based on our proposal would be beneficial for computational modeling work, and we leave this for future research.

CHAPTER 4

Reduplication learning as hidden structure learning

4.1 Introduction

In this chapter,[1] we propose a morphophonological learner that handles reduplication, focusing on learning the unobserved prosodic template when the underlying representations of the stems and other affixes are likewise unknown. The proposed learner attributes the realization of copying to phonology and treats the candidate prosodic templates as phonological hidden structures, with the same status as phonological underlying representations. We show that under this view, the problem of learning reduplication can be integrated into the picture of learning concatenative morphophonological processes in intuitive ways. This is because they share the same learning question, namely how to learn phonological hidden structures together with an adequate phonological grammar. To address this question, we extend to the *expectation-driven maximum entropy lexicon learner* (hereafter, EM-MaxEnt; Wang and Hayes, resubmitted) with a component to handle the reduplicative morpheme.

We offer two lines of modeling results in this chapter. The first set of simulations reflects the ability of this learner to handle the typology of reduplication-phonology interactions. Using reduplication and word-final devoicing as an example, we have prompted the proposed learner with different types of interactions observed in the literature (e.g., McCarthy and Prince, 1995; see our discussion in Section 1.3), including normal application, overapplication, underapplication and templatic back-copying. We show that the learner can handle these patterns by learning the correct underlying representations, as well as the right grammar

[1] A part of this chapter is adapted from Wang and Hayes (resubmitted), which is co-authored with Bruce Hayes.

to derive the surface forms. Moreover, the learner recapitulates the predicted asymmetry between underapplication and overapplication (McCarthy and Prince, 1995, pp. 91-92) in a probabilistic way. Without extra treatment, underapplication leads to a phonological grammar with variation, exhibiting a frequency-matching behavior (Hayes et al., 2009).

The second set of simulations shows that our approach captures the population-level results of the experiments discussed in Chapter 2. Through these simulations, we find our proposal may provide a learning-based account for the debate of "templatic" and "a-templatic" approaches to reduplication (e.g., McCarthy and Prince, 1986; Spaelti, 1997; Gafos, 1998). Given the learning results, our tentative answer to this question is a learner must be equipped with the capacity of learning templates, but the final learned grammar does not necessarily need to use these representations. An "a-templatic" grammar might emerge as an outcome driven by principles of learning, under the assumption that the learner knows what constitutes a "base".

This chapter is organized as follows. We discuss the previous models relevant to reduplication learning and provide the desiderata we aim to achieve with our proposal (Section 4.2). Before introducing the reduplication learning component, we provide an overview of the workings of *expectation-driven maximum entropy lexicon learner* (EM-MaxEnt; Section 4.3). We introduce the details of the proposed reduplication learning component in Section 4.4, with an example of the normal application. Section 4.5 describes how the proposed learning system deals with a series of examples reflecting different types of reduplication-phonology interactions. Section 4.6 presents how the learner handles the results of the experiments and discusses its implications for the templatic-atemplatic debate. Section 4.7 concludes this chapter by discussing future directions for research.

4.2 Defining the problem

In this chapter, we focus on one specific task in morphophonological learning, the task of learning the underlying representations along with the phonological grammar such that the correct surface forms are derived. For example, if provided with {([kæt], 'cat'), ([kæts], ‘cats’), ([dɑg], ‘dog’), ([dɑgz], ‘dogs’)}, how can a learner successfully learn the UR for the plural morpheme is /-z/ and acquire a devoicing grammar as hypothesized by linguists and verified by experimental testing (Berko, 1958)? This task has proven to be a challenging one, and it has been undertaken with a wide variety of approaches, including error-driven learning with ranked constraints (e.g., Tesar et al., 2003; Apoussidou, 2006, 2007); distributional learning under the principle of Minimum Description Length with constraints (Rasin and Katzir, 2016) and with rules (Rasin and Katzir, 2018, 2020; Rasin et al., 2021); formal learning algorithms based on the subregular phonology hypothesis (Hua and Jardine, 2021), Bayesian Program Synthesis with ordered rules (Ellis et al., 2015; Barke et al., 2019; Ellis et al., 2022), algorithmic learning approaches guided by different evaluation criteria (Khalifa et al., 2023; Belth, 2023a,b), and maximum likelihood learning under different probabilistic frameworks (Jarosz, 2006; Pater et al., 2012; Cotterell et al., 2015; Johnson et al., 2015; O’Hara, 2017; Nelson, 2019; Tan, 2022; Wang and Hayes, resubmitted). For more complete literature surveys, see Jarosz (2013, 2015, 2019), Tesar (2014), Cotterell et al. (2015), and Rasin et al. (2021).

Despite being voluminous, the previous literature on morphophonological learning has only focused on concatenative processes. Very few have tried to capture reduplication. Previous attempts that investigated reduplication can be loosely divided into three categories.

First, computer scientists and linguists have worked on modeling low-resourced languages that often involve non-concatenative morphology, with the ultimate goal of modeling the morphological typology. Reduplication is usually included as a specific test case in these studies. Previous approaches along this line mainly focused on the morphological side of reduplication (i.e. whether a word contains copy structures), neglecting its phonological nature (i.e. how copying is realized). Vania and Lopez (2017) examined the types of representations that can capture the typology of morphology in broad strokes. Xu et al. (2020) proposes an unsupervised morphological learner, aiming to distinguish reduplicated words with distinct functions from unreduplicated words. Similarly, Todd et al. (2022) investigated the morphological segmentation task, finding support for prosodic templates in their experiments. They extended Morfessor (Creutz and Lagus, 2007), one of the widely used baseline models for unsupervised morphological segmentation, with the capability of dealing with reduplication. They added five templates for Māori reduplication. Models with reduplicative templates performed significantly better than those without. Despite this, the templates added were mainly moraic representations and were predefined according to Māori morphophonology. They acknowledged the potential difficulty if candidate templates were "too general or too numerous" (p. 17).

Secondly, as we discussed earlier in Chapter 2, the experimental work by Marcus et al. (1999) motivated a sequence of computational modeling work, including deep neural networks to answer whether variables are necessary for models of language learning (e.g., Gasser, 1994; Frank and Tenenbaum, 2011; Nelson et al., 2020; Beguš, 2021; Prickett et al., 2022; Beguš and Zhou, 2022; Ellis et al., 2022). These studies involve relatively simple rules, usually the Marcus et al.-style stimuli (ABB as in *wofefe*, AAB as in *wowofe*, ABA as in *wofewo*, where A, B are CV syllables). Among these efforts, Beguš (2021) treated reduplication as a morphophonological process. Beguš trained Generative Adversarial Networks (Goodfellow et al., 2014) on raw speech data, with CV reduplication for CVCV stems.[2] Ellis et al. (2022) attempted to add syllabic representation in their model for the Marcus et al.-style stimuli, and found that models enriched with syllabic representations captured the intended rule more rapidly, mirroring infants' rapid generalization.

Lastly, a few works (Dolatian and Heinz, 2018a; Wilson, 2019; Nelson et al., 2020) have looked at reduplication as a morphophonological process with more variable patterns. Dolatian and Heinz (2018a) had approached reduplication learning from a formal perspective. Driven by the typological discovery that most reduplicative patterns are *Concatenative Output Strictly k-local*, they proposed an algorithm that learns these output strictly local functions for each copy and concatenates these functions. Their algorithm operates under the assumption that the boundary between two copies is discovered beforehand. Wilson (2019) proposed an interpretable network using Tensor Product Representations (Smolensky, 1990) and tested it on various types of reduplication, including full reduplication, echo redupli-

[2] Actually, the studied reduplication pattern is [Cʌ-/Cə-] since the trained speech was based on a speaker of American English and the reduplicant was usually unstressed.

cation, CV reduplication, and Ilokano heavy syllable reduplication. Nelson et al. (2020) investigated the capacities of Encoder-Decoder models in handling CV copying and total reduplication.

Given the advantages and limitations of previous models, in (38), we provide the following desiderata for what we think an adequate morphophonological learner with reduplication should meet.

(38) Desiderata

a. *Empirical coverage*

The learner ought to be descriptively adequate for empirical findings, including typology and experimental results.

i. *Reduplication as phonological copying.*

Driven by the theoretical research on reduplication (Wilbur, 1973; Carrier, 1979; Marantz, 1982, *et seq.*), the learner should attribute the realization of the copying operation as a part of the phonology so that the need to satisfy a prosodic template may interact with other phonological and/or morphophonological processes.

ii. *Learn the right level of prosodic templates*

Agnostic to what prosodic template a reduplicative morpheme needs, the learner ought to learn the right template by itself.

b. *Transparency and Interpretability*

The learner should be transparent and interpretable so that we can easily examine when and why it succeeds, as well as when and why it fails.

We show that treating the prosodic templates as phonological hidden structures provides intuitive ways to meet these desiderata. To carry out learning, we follow the general framework of the *expectation-driven maximum entropy learner* (EM-MaxEnt) proposed in Wang and Hayes (resubmitted). We extend a reduplication learning component to their proposed learner and offer a preliminary investigation of its behavior. Before we describe our extension in Section 4.4, we start with a summary of their proposal for a concatenative morphophonological process in the next section.

4.3 Expectation-driven learning of phonological underlying representations

4.3.1 An overview of EM-MaxEnt

The notion of "expectation-driven" learning of phonological structures was first proposed by Jarosz (2006, 2015) for a probabilistic version of Optimality Theory (OT; Smolensky and Prince, 1993). This idea is rooted in the well-established machine learning literature, particularly, the extensive literature on *expectation-maximization* for parameter estimation (Dempster et al., 1977). Wang and Hayes (resubmitted) further applied expectation-maximization to Maximum Entropy Grammars (MaxEnt; Goldwater and Johnson, 2003; Hayes and Wilson, 2008) to tackle the joint learning of underlying representations and the phonological grammar.

Though it essentially uses the same GEN-cum-EVAL architecture as OT, MaxEnt differs by assigning weights to each constraint instead of ranking them. These weights reflect the strength of the constraints, which are grounded in broader principles based on typological and phonetic studies. Given the constraint weights and the pattern of constraint violations, a probability is computed for each candidate. With this probabilistic nature, MaxEnt is often adopted for its ability to capture gradient phenomena, such as free variation (Labov, 1969, *et seq.*), lexical frequency matching (Zuraw, 2000, *et seq.*), and soft-UG biases during learning (e.g., Wilson, 2006; Becker et al., 2012; White, 2017; Kuo, 2023a). Moreover, MaxEnt is computationally simple and tractable (e.g., yielding intuitive gradient for learning the constraint weights; see Della Pietra et al., 1997; Hayes and Wilson, 2008).

MaxEnt is used in several current systems for UR learning (e.g., Eisenstat, 2009; Pater et al., 2012; Johnson et al., 2015; O'Hara, 2017; Nelson, 2019; Tan, 2022). The approach in Wang and Hayes (resubmitted) differs from most previous applications in treating UR parameters as a categorical distribution over possible UR candidates for each morpheme, the same as O'Hara (2017). But different from O'Hara (2017), they apply Expectation-Maximization to carry out the parameter estimation, resulting in a more interpretable, transparent learning procedure. Another aspect of Wang and Hayes' (resubmitted) proposal is a principled method to generate underlying representation candidates. Specifically, they revisited the abstractness hierarchy set forth by Kenstowicz and Kisseberth (1977) (KK). For example, Seediq (Autronesian) stems for *hold* surface as [pemux], [pumex], and [pumux] at different slots in the paradigm (Kuo, 2023b). KK's level C ("pick and choose") uses this allomorph set as the possible UR candidates: {/pemux/, /pumex/, /pumux/} while KK's level D (segmentally-composite URs) producing a larger UR set {**/pemex/**, /pemux/, /pumex/, /pumux/} by freely combining alternating segments at each segment slot. They translated levels of the KK hierarchy into algorithms and ran the same training data with varying degrees of abstractness for URs. They found support for a relatively concrete position indicating that URs might not be too distant from their observed forms and concluded that KK-C might be the right level of abstractness. Inspired by Eisenstat (2009), they further employed segmental alternations for a principled surface candidate Gen, to include crucial surface candidates (SRs) that disambiguate different grammars.

The architecture of the EM-MaxEnt proposal is given in Figure 4.1. In the rest of this section, we will walk through the core components relevant to our enrichment for handling reduplication. Hence, we will skip their proposals for morpheme segmentation, the identification of the alternating segments driven by string alignment and phonetic similarity, and the different levels of UR Gen. Readers curious to know more about how these components work should refer to the original paper. We will assume KK's level C as the level for the UR Gen of concatenative processes and offer a toy example similar to the Pseudo German in the original paper (see also Pater et al., 2012).

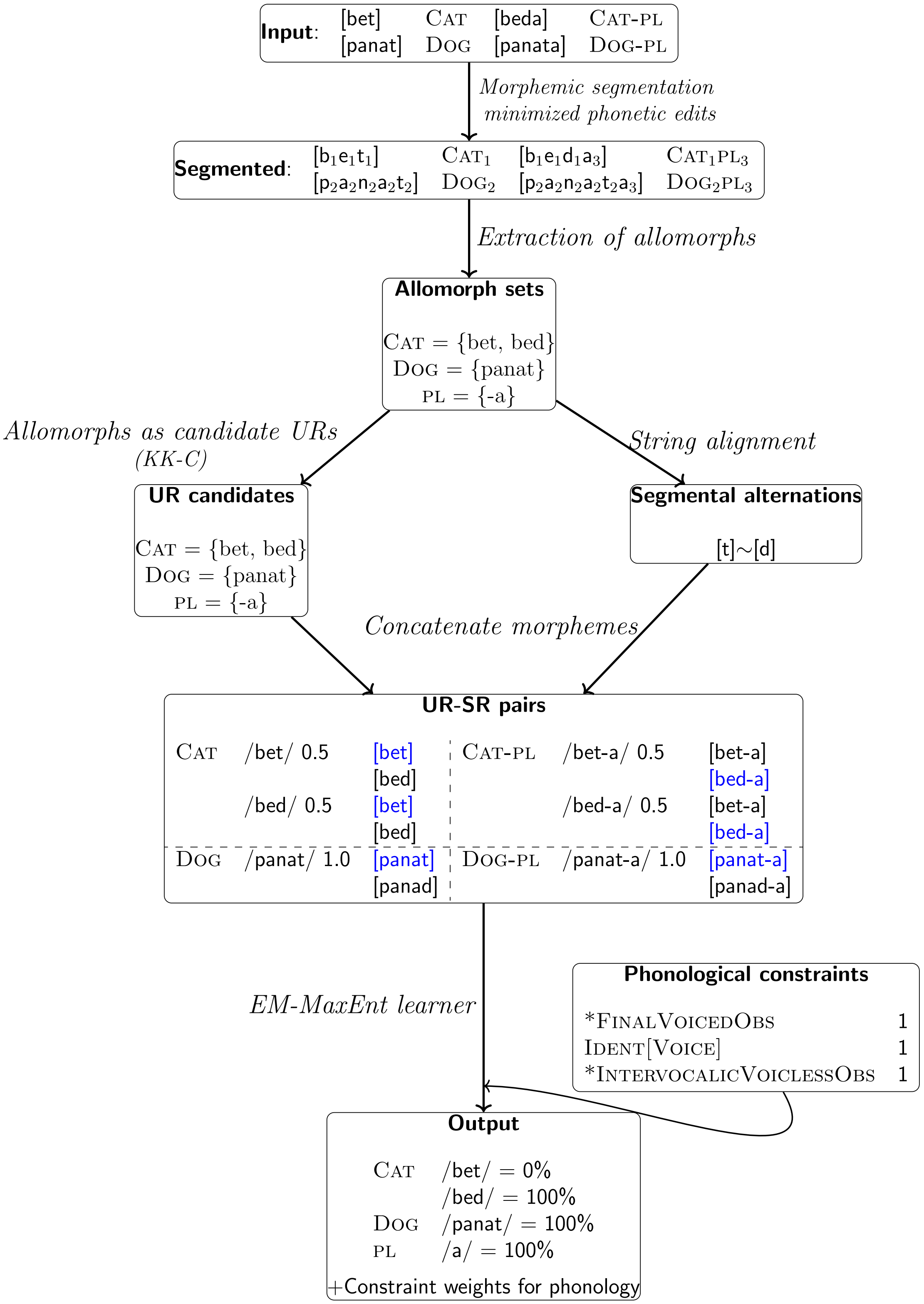


Figure 4.1: *The architecture of the proposed learning system for concatenative morphology in Wang and Hayes (resubmitted) with an illustration of a part of the data*

Meanings	*Forms*
Cat$_{\text{STEM}}$	bet
Cat-pl	beda
Dog$_{\text{STEM}}$	panat
Dog-pl	panata

Morphemes	*Intended URs*
Cat	/bed/
Dog	/panat/
pl	/-a/

Table 4.1: *Input data of the toy dataset (left) and the intended URs for each morpheme (right). The intended phonology: final devoicing.*

4.3.2 The toy dataset: Pseudo German

As shown in Table 4.1, the paradigm of Cat ([bet], sg; [beda], pl) shows that the segment [t] alternates with the segment [d]. The task for the learner is to decide the direction of the segmental alternation and construct a phonological grammar that correctly derives the right direction. In this case, the learner ought to choose between (a) a devoicing grammar such that /d/ ↦ [t] in the word-final position as seen in the isolation form, and (b) a voicing grammar such that /t/ ↦ [d] when surrounded by vowels as in the pl form. Both grammars work equally well in characterizing the paradigm Cat. As linguists, to decide between these two grammars, we find disambiguating data from the paradigm of Dog. Dog has no alternating allomorphs, and URs may be hypothesized as the non-alternating realization /panat/. As one can observe, the surface [t] in Dog-pl occurs in the same context as the surface [t] in Cat-pl, namely surrounded by vowels. A voicing grammar would predict Dog-pl to be *[panada], but in fact, [panata] surfaces. On the other hand, the devoicing account is consistent with both paradigms. Thus, the right grammar should be the devoicing account.

Like a linguist, an automated learner might carry out similar reasoning, needing the paradigm Dog to provide sufficient evidence. If so, there must be a way to include the crucial losing candidate *[panada] for the UR /panat-a/ as the disambiguating candidate so that the learner can see the incorrect predictions the voicing grammar makes. Then, the learner needs to attribute the different behaviors of the paradigm Dog and the paradigm Cat to a difference in their underlying representations, thus concluding that Cat must be underlying /bed/, instead of /bet/. The next section describes how the restrictive SR Gen proposed in the EM-MaxEnt learner includes [panada] in the set of surface candidates.

4.3.3 The alternation-substitution SR Gen

At this stage, the learner has already performed word segmentation by assigning each segment to its morpheme. It has also tallied up allomorph sets for each morpheme, and assigned each allomorph in the allomorph set as a possible underlying representation for its morpheme. Additionally, it has noticed that the paradigm CAT shows a [t] ∼ [d] alternation.

As just discussed, one of the goals of morphophonological learning is to figure out the direction of the alternation and learn a phonological grammar that accounts for such a direction. In the context of Pseudo German, once the learner notices the alternation [t] ∼ [d], the learning problem is to decide whether /t/ changes to [d] in certain contexts, or /d/ changes to [t] in other contexts. Inspired by Eisenstat (2009), Wang and Hayes (resubmitted) noted that the identified alternating segments could guide the learner in generating surface candidates for each underlying form: the learner should entertain /t/ → [d] and /d/ → [t] at all possible positions. Thinking of DOG-PL /panat-a/, this /t/ should have the possible alternation to [d]. Substituting /t/ with this possible alternant [d] creates the desired [panada]. A similar idea applies for /panad-a/: Substituting the underlying /d/ with the possible alternant [t] results in [panata]. Such a process should apply to the underlying segments with a possible alternant. Let us assume the learner also observes the alternation [p] ∼ [b]. The learner needs to determine whether /p/ changes to [b] in certain contexts or whether /b/ changes to [p] in others. Consequently, for the underlying form /panat-a/, the /p/ potentially alternates with [b], leading to further consideration of forms like [banat-a] and [banad-a].

The proposed SR GEN is given in (39) to define this process precisely.

(39) Alternation substitution SR GEN; denoted by $\textsc{Gen}(x_1x_2 \ldots x_n)$
Given the alternation relations ALT for a language, $\textsc{Gen}(x_1x_2 \ldots x_n) = \textsc{Alt}(x_1) \bullet \textsc{Alt}(x_2) \bullet \ldots \bullet \textsc{Alt}(x_n)$, where $S \bullet T = \{st \,|s \in S, t \in T\}$.

Note that the alternation relation of segmental pairs is reflexive to include the faithful candidate; namely for all possible segment x_1, $x_1 \in \textsc{Alt}(x_1)$. It is also symmetric: $y_1 \in \textsc{Alt}(x_1)$ implies $x_1 \in \textsc{Alt}(y_1)$. But it is not transitive. That $z_1 \in \textsc{Alt}(y_1)$ and $y_1 \in \textsc{Alt}(x_1)$ does not necessarily mean $z_1 \in \textsc{Alt}(x_1)$.[3] Hence, given $u = vx_1w$, $vz_1w \notin \textsc{Gen}(u)$. Excluding transitivity aims for conservative learning, in the spirit that the learner does not hypothesize unobserved alternations.[4] To see this, let us think about the possibility that in a language like English, /t,d/ undergo tapping and thus change to [ɾ]. That said, [t] ∼ [ɾ] and [d] ∼ [ɾ]. For the underlying representation /pæt/, the procedure discussed above includes [pæɾ] as a candidate SR, but [pæd] is not a competing candidate so that the learner does not need to account for the unobserved [t] ∼ [d] alternation. Yet for the underlying representation /pæɾ/, alternation substitution Gen generates both [pæt] and [pæd] as candidates.[5]

Under this procedure, given Alt(t) = {t, d}, then Gen(/panat-a/) = {[panata], [panada]}. If the alternation further includes Alt(p) = {p, b}, then Gen(/panat-a/) = {[panata], [panada], [banata], [banada]}. Note that the alternations do not, in principle, need to be segment-based. After seeing sufficient evidence, the learner might abstract away from these segmental alternations and form a feature-based alternation set, which we leave as future research.

4.3.4 The mixture of MaxEnt grammars

4.3.4.1 The mixture structure

Given an allomorph set for each morpheme, KK's level C uses these allomorphs as candidate URs. Hence, the possible URs for Cat are /bet/ and /bed/. At the beginning of learning, the learner does not know which plausible UR is correct, nor the phonological grammar.

[3] This is different from the proposal in Eisenstat (2009).

[4] That said, Alt can go beyond the observed alternating segments based on phonetic intermediateness. Assume k ∼ ɣ, leaping over the segment g and x, then k ∼ {k, x, g, ɣ}, as well as ɣ ∼ {ɣ, g, x, k}. See Wang and Hayes (resubmitted) for more details.

[5] Wang and Hayes assumes phonemicization before morphophonemic learning. The English tapping example here is just for an illustrative purpose. For previous computational models on allophonic learning, see Peperkamp et al. (2006); Calamaro and Jarosz (2015); Rasin et al. (2021); Richter (2021).

The EM-MaxEnt learner treats underlying representations as latent variables. The learner maintains a probability distribution over all plausible URs and starts by assuming that each plausible underlying representation is equally probable. The probability of each surface form (s) for a form (ω) sums across all plausible hidden structures (u), as illustrated in (40a). (40b) shows how this formula works with Cat-pl as a concrete example.

(40) a. A mixture of phonological grammars

$$P(s \mid \omega; \boldsymbol{W}, \boldsymbol{\theta}) = \sum_{u \in UR(\omega)} \overbrace{P(s \mid u; \boldsymbol{W})}^{\text{Grammar}} \overbrace{P(u \mid \omega;\, \boldsymbol{\theta})}^{\text{UR}}$$

where $\forall\omega$, $\sum_{u \in UR(\omega)} P(u \mid \omega) = 1$

b. An illustration with Cat-pl

$$P(\,[\text{bed-a}] \mid \textsc{Cat-pl}) = \overbrace{P(\,[\text{beda}] \mid /\text{bet-a}/)}^{\text{Grammar}}\ \overbrace{P(\,/\text{bet-a}/ \mid \textsc{Cat-pl})}^{\text{UR}}$$
$$+\, P(\,[\text{bed-a}] \mid /\text{bed-a}/)\ P(\,/\text{bed-a}/ \mid \textsc{Cat-pl})$$

The model as shown in (40a) is very similar to the classic (finite) mixture model in the statistics literature, where the component of the grammar is often referred to as the mixture components, and the UR portion identified above is referred to as the mixture weights.[6] Note that (40a) is a special version of the mixture models. In traditional mixture models, the mixture components could have different parameter weights in principle. However, the learner assumes the same set of weights for each component. In other words, the phonological model is held the same for all possible underlying representations. Assuming one phonological grammar is precisely the main reason why learning can take place. It is also the main reason why for most of the cases, the learner tries to learn a single underlying representation for each morpheme. Learning multiple underlying representations with the wrong grammar indicates failure, and the learner gets stuck in a local optimum, not the global solution. Yet sometimes, the learner can learn multiple underlying representations that work equally well

[6] See MacLahlan and Peel (2000).

under the correct phonological grammar, and we discuss the implications in later sections.

4.3.4.2 The MaxEnt grammar

We assume a MaxEnt grammar for the probabilistic phonological grammar. It assigns a non-negative real number as weight w_k to each constraint C_k, reflecting constraint strength. Based on the weights w_k and the pattern of constraint violations of input-output pairs $C_k(u, s)$, a probability distribution is computed with the softmax function, as in (41).

(41) The probability of a candidate s_i, given a UR u

$$P(s_i \,|\, u) = \frac{e^{-\sum_k w_k C_k(u,s_i)}}{\sum_j e^{-\sum_k w_k C_k(u,s_j)}}$$

The values $\sum_K w_k C_k(u, s)$ are often denoted as the *Harmony* scores, denoted as $\mathcal{H}$. For Pseudo German, the relevant constraints are *FinalVoicedObs, which penalizes word-final voiced obstruents, Ident[Voice], which penalizes voice mismatches between the input and the output, and *VTV, which penalizes intervocalic voiceless obstruents. A MaxEnt grammar that captures Pseudo German is given in Tableau (42). The weight of *FinalVoicedObs should be substantially higher than Ident[Voice] so that the grammar predicts the final voiced obstruent to devoice. *VTV must receive negligible weights so that the grammar does not make wrong predictions with intervocalic voicing.

(42)

	*FinalVoicedObs	Ident[Voice]	*VTV	$\mathcal{H}$	$P(s \mid u)$
	19.05	10.09	0.01		
Input: /bed/					
a. [bed]	1			19.05	0.0001
b. [bet]		1		10.09	0.9999
Input: /bed-a/					
a. [bed-a]				0	0.9999
b. [bet-a]		1	1	10.10	0.0001
Input: /panat/					
a. [panad]	1	1		29.14	0.0001
b. [panat]				0	0.9999
Input: /panat-a/					
a. [panad-a]		1		10.09	0.0001
b. [panat-a]			1	0.01	0.9999

4.3.4.3 The objective

The learning objective is based on the log conditional likelihood in order to carry out maximum likelihood/maximum a posteriori estimation. If there is no hidden structure, the MaxEnt learners are advantageous with a convex search space and intuitive gradients (Della Pietra et al., 1997; Goldwater and Johnson, 2003). This leads to the applicability of many optimization methods to obtain the optimal constraint weights.[7] MaxEnt learners are also flexible in incorporating soft UG biases (e.g., Wilson, 2006; White, 2017; Kuo, 2023a). This is often formulated as a Gaussian prior added to the log-likelihood (see 43b). Including a Gaussian prior can also prevent infinite weights and overfitting of the data.

However, as reflected in (43a) and (40a), hidden structures are indeed involved in our objective here. The unknown parameters are not only the set of weights $\boldsymbol{W}$, but also the hidden UR probabilities $\boldsymbol{\theta}$. The convexity of the search space is no longer guaranteed,

[7] Following many previous practices of MaxEnt, we opted for the L-BFGS-B, a bounded quasi-Newton method as the optimizer (Zhu et al., 1997). The constraint weights were enforced to be non-negative. All constraint weights presented in this chapter were fitted with Python's built-in implementation, in `scipy.optimize`. For R users, a useful package for fitting MaxEnt grammars and conducting model comparisons can be found in Mayer et al. (to appear).

requiring more complicated ways to search for the optimal parameter values. In the next section, we describe how expectation-maximization is applied to find the optimal constraint weights $\boldsymbol{W}$, as well as the parameters for the UR probabilities $\boldsymbol{\theta}$.

(43) a. The log conditional likelihood of the training data

$$\begin{aligned} ln(P(D \,|\, \boldsymbol{W},\, \boldsymbol{\theta})) &= ln(\prod_{(s,\,\omega)\in D} P(s \,|\, \omega \,; \boldsymbol{W} \,, \boldsymbol{\theta})^{f(s,\omega)}) \\ &= f(s, \omega) \sum_{(s,\omega)\in D} ln(P(s \,|\, \omega; \boldsymbol{W}, \boldsymbol{\theta})) \end{aligned}$$

where $P(s \,|\, \omega; \boldsymbol{W}, \boldsymbol{\theta})$ is defined in (40a).

b. The objective function

$$L(\boldsymbol{W}, \boldsymbol{\theta}) = ln(P(D \,|\, \boldsymbol{W}, \boldsymbol{\theta})) - \sum_i \frac{(w_i - \mu_i)}{2\sigma_i^2}$$

4.3.5 Finding the right combination of UR probabilities and constraint weights

The parameters whose values must be calculated are $\boldsymbol{\theta}$, the probability distributions over UR candidates, and $\boldsymbol{W}$, the weights of the phonological grammar. To carry out the search, the EM-MaxEnt learner uses expectation-maximization (Dempster et al., 1977; Do and Batzoglou, 2008). This algorithm is usually applied to find the parameter estimates to maximize certain probabilistic models with unobserved hidden structures (for its application in previous linguistic research, see e.g., Lari and Young, 1990; Eisner, 2001; Goldwater and Johnson, 2005; Jarosz, 2006; Dreyer and Eisner, 2011; Pater et al., 2012; Jarosz, 2013; Nelson, 2019; Cotterell et al., 2015; Tan, 2022). It iteratively breaks down an optimization task into a set of smaller steps and alternates between them. The two steps are the Expectation (E) step and the Maximization (M) step.

The E-step fills in the missing UR information in the input by calculating expected values. Intuitively, it calculates how much "responsibility" each UR u should take for any observed pairs of surface forms and the word, namely $(s,\, \omega)$. It first calculates the probability of u given an observed pair, as in (44a). To calculate the probability $P(u, s \,|\, \omega)$ in the denominator and the numerator, a posterior probability distribution over the hidden structures is calculated, as expanded in (44b).

(44) a. The probability of the UR u for an observed pair

$$\begin{aligned} P(u\,|s, \omega) &= \frac{P(u, s\,|\,\omega)}{P(s\,|\,\omega)} \\ &= \frac{P(u, s\,|\,\omega)}{\sum_{u' \in UR(\omega)} P(u', s\,|\,\omega)} \end{aligned}$$

b. E-step: Expected frequencies of the UR u in observation (s, ω)

For each $u \in UR(\omega)$,

$$E(u, s, \omega) = f(s, \omega) \frac{P(s\,|\,u\,; \boldsymbol{W})P(u\,|\,\omega\,; \boldsymbol{\theta})}{\sum_{u' \in UR(\omega)} P(s\,|\,u'\,; \boldsymbol{W})P(u'\,|\,\omega\,; \boldsymbol{\theta})}$$

where $P(s\,|\,u; \boldsymbol{W})$ and $P(s\,|\,u'; \boldsymbol{W})$ are as given in (41).

During the M-steps, the algorithm obtains a better estimate for the parameters $\boldsymbol{W}$ and $\boldsymbol{\theta}$ by maximizing the likelihood of "filled-in" data, based on the guess made at the Expectation step. The calculations for $\boldsymbol{W}$ are given in (45); those for $\boldsymbol{\theta}$ in (46).

(45) M-step: Using estimated frequencies to calculate constraint weights

$$W = argmax_{\boldsymbol{W}} \sum_{(s,\omega) \in D} \sum_{u \in UR(\omega)} E\left(u, s, \omega\right) \; ln\left(P\left(s \;|\; u; \boldsymbol{W}\right)\right) \; - \; \sum_{i} \frac{\left(w_i - \mu_i\right)^2}{2{\sigma_i}^2}$$

where $P(s\,|\,u; \boldsymbol{W})$ is defined in (41)

(46) M-step: Using estimated frequencies to re-estimate UR probabilities

$$\theta_{(\mu,\nu)} = \frac{\displaystyle\sum_{(s,\omega) \in M(\mu)} \sum_{\substack{u \in UR(\omega) \\ \text{s.t. } u \text{ contains } \nu}} E\left(u, \; s, \omega\right)}{\displaystyle\sum_{\nu' \in UR(\mu)} \sum_{(s,\omega) \in M(\mu)} \sum_{\substack{u' \in UR(\omega) \\ \text{s.t. } u' \text{ contains } \nu'}} E\left(u', \; s, \omega\right)}$$

where μ is an abstract morpheme, and ν is a possible UR

$M(\mu)$ is the set of word forms containing μ

E is as defined as in (44b).

With both the E-step and the M-steps complete, a single iteration is accomplished. The next iteration begins by inputting the new parameter values $\boldsymbol{\theta}^{\mathbf{t+1}}$ and $\boldsymbol{W}^{\mathbf{t+1}}$ to (44b), and the process continues.[8] The first iteration starts by assigning initial values $\boldsymbol{\theta}^{\mathbf{0}}$ and $\boldsymbol{W}^{\mathbf{0}}$ to $\boldsymbol{\theta}$ and $\boldsymbol{W}$. The process terminates when an iteration ceases to improve log-likelihood by more than some small threshold amount.[9] For one iteration with Pseudo-German, see Appendix E.1.

Figure 4.2 shows the learning trajectory of the grammar as well as the UR probability for the stem Cat. As we can see, over multiple iterations, the learner learned the linguist-hypothesized /bed/ as the correct UR ($\theta_{(\text{Cat, /bed/})} = 0.999$). It also learned a final devoicing grammar, reflected by the substantial constraint weight of *FinalVoicedObs ($w = 19.05$). It did not learn the intervocalic voicing grammar, as the Ident[Voice] ($w = 10.09$) received more substantial weight than *VTV ($w = 0.01$), preventing the intervocalic voiceless obstruent from voicing. In the rest of the chapter, for ease of presentation, we use $w_1 \succ w_2$ to indicate the substantial difference between constraint weights (or weights combinations) that would lead to near-1 probability of one particular candidate. Hence, here, $w_{*\text{FinalVociedObs}} \succ w_{\text{Ident[Voice]}} \succ w_{*\text{VTV}}$.

4.4 Learning prosodic templates as hidden structures

In this section, we introduce our extended reduplication learning component to the EM-MaxEnt learner. This component aims to learn (a). the correct reduplicative template(s), (b). the underlying representations for other stems and affixes, and (c). the realization of copying in tandem with other morphophonological alternations. We treat the realization

[8] Between (45) and (46), we also carried out another E-step (44b) based on the new constraint weights.

[9] Unless otherwise specified, the values we employed for the regularization term were μ_i at 0 and $2{\sigma_i}^2$ at 10^5 for all constraints. Learning was terminated on the first iteration at which log-likelihood increased by less than 10^{-3}.

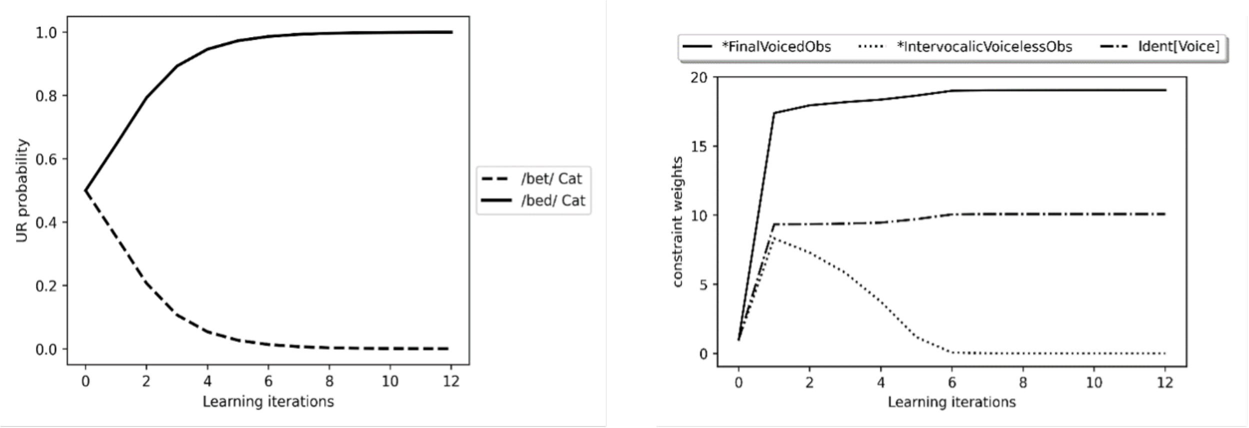


Figure 4.2: *The learning trajectory based on the Pseudo German training data*

of a reduplicative morpheme as phonological copying, allowing it to freely interact with other phonological processes. The extension includes the hidden prosodic templates for the reduplicative morpheme (heavy syllable $\sigma_{\mu\mu}$, light syllable σ_{μ}, foot Ft and prosodic word PrWd), the SR Gen that generates candidates with the reduplicative morpheme, and the constraints in the phonological grammar. We will discuss these aspects in the rest of this section. Other components of the model are the same as the concatenative version of the EM-MaxEnt, particularly the EM-based parameter estimation method.

4.4.1 The toy dataset: Pseudo German with word-final heavy syllable copying

To make each step concrete, we will also work with a toy example, which contains the previous word forms from Pseudo German and another paradigm slot dim. We assume that the learner has already performed the non-trivial step of segmenting words into their morphemes, as illustrated in Table 4.2. The intended hidden structure associated with each morpheme is given in Table 4.3. The intended phonology should reflect final devoicing and suffixing heavy syllable reduplication for the diminutive.

After segmentation, the learner needs to tally the allomorph sets of each morpheme. The original EM-MaxEnt learner for concatenative processes is *phonologically eager*. It always attempts to assign a single underlying representation to each morpheme and attribute observed surface differences within each allomorph set to (segmental) phonological regulari-

Meanings	*Forms*	*Meanings*	*Forms*
Cat$_{\text{STEM}}$	bet	Cat$_{\text{STEM}}$	bet
Cat-PL	beda	Cat-PL	bed-a
Cat-DIM	bedbet	Cat-DIM	bed-bet
Dog$_{\text{STEM}}$	panat	Dog$_{\text{STEM}}$	panat
Dog-PL	panata	Dog-PL	panat-a
Dog-DIM	panatnat	Dog-DIM	panat-nat

Table 4.2: *Input toy dataset (left) and the segmented data (right).*

Morphemes	*Intended hidden structures*
Cat$_{\text{STEM}}$	/bed/
Dog$_{\text{STEM}}$	/panat/
PL	/-a/
DIM	Red $= \sigma_{\mu\mu}$

Table 4.3: *The intended UR for each morpheme. The intended phonology should reflect final devoicing and suffixing heavy syllable reduplication.*

ties. Relaxing such phonological eagerness, we propose that at this stage, the learner should determine how likely the target stems/affixes are realized by different types of morphophonological structures, such as segmental alternation, listed allomorphy, and reduplication. The learner can generate probability distributions over these possible types based on various linguistic principles. For example, driven by paradigm uniformity (i.e. allomorphs of the same paradigm should be phonetically similar; Steriade, 2000, *et seq.*), in the paradigm Cat, the allomorphs [bet] and [bed] are so similar that a segmental alternation account is highly appealing compared to other types.

As for the suffix DIM, one observation the learner can make is that the allomorphs [nat] and [bet] are too dissimilar to be unified by a single underlying representation. That said, for DIM, the learner might assign a rather low probability to a segmental alternation account. Similarly, logically speaking, the learner could also attribute DIM to listed allomorphy. However, when multiple analyses can account for the data, an allomorphy account is often more "costly" due to the need to specify different allomorphs in the lexicon when it is not necessary. The third possibility is to hypothesize DIM as a case of reduplication, driven by the internal similarity within each relevant word form (e.g., [panat-nat], [bed-bet]). For the purpose of this chapter, we assume a categorical approach, namely that the learner adopts a

Morphemes	*Allomorph set*	*Morphophonological structures*	*Hidden structures*
Cat	{bet, bed}	alternation: [t] ∼ [d]	UR: {bet, bed}
Dog	{panat}		UR: {panat}
pl	{-a}	suffixation	UR: {-a}
dim	{-bet, -nat}	alternation: {[b] ∼ [n]; [e] ∼ [a]}; $p \approx 0$	
		allomorphy /-bet, -nat/; $p \approx 0$	
		Red; $p \approx 1$	Templates {PrWd, Ft, σ_{μ}, $\sigma_{\mu\mu}$}

Table 4.4: *Hypotheses on different types of moprhophonological structures for each morpheme.*

reduplicative account of dim and assigns zero probability to other possibilities, as illustrated in Table 4.4.

4.4.2 Reduplicative SR Gen closed under alternation substitution

Given that the experimental results in Chapter 2 support Prosodic Morphology (McCarthy and Prince, 1986, *et seq.*) and Base-Reduplicant Correspondence Theory (BRCT; McCarthy and Prince, 1995, *et seq.*), we base our learner on these theories to handle reduplication. The learner assumes the reduplicative morpheme to be abstract, the phonological shape of which adheres to the principles of Prosodic Morphology, as reiterated in (47).[10] Per BRCT, the segmental realization of the abstract morpheme is determined by BR faithfulness constraints. In the remainder of this section, we focus on discussing aspects of the prosodic shapes.

(47) Principles of Prosodic Morphology (adapted from McCarthy and Prince, 2017, p.283)

a. Prosodic Morphology Hypothesis

Templates are defined in terms of the authentic units of prosody: mora (μ), syllable (σ), foot (Ft) and prosodic word (PrWd)

b. Template Satisfaction Condition

Satisfaction of templatic constraints is obligatory and is determined by the principles of prosody, both universal and language-specific.

[10] We suppress the discussion of prosodic circumscription (McCarthy et al., 2012), the principle that governs the interaction between morphological operation and its prosodic criteria that need to refer to a prosodically delimited constituent within the base.

Following Prosodic Morphology (McCarthy and Prince, 1990, 1986, *et seq.*), we assume that the different levels of prosodic units are readily accessible to the learner through Universal Grammar (UG). The underlying representation of a reduplicative morpheme is always an abstract RED morpheme, referring to the prosodic units in (47a).[11]

During the process of learning, the learner needs to determine the most plausible prosodic template for the realization of the phonological shape. Just as the learner does not know whether the right direction of alternation is /t/ → [d] or /d/ → [t], the learner also does not know the right template for DIM. When looking at the paradigm of CAT, the learner cannot decide whether the reduplicative template is a heavy syllable ($\sigma_{\mu\mu}$), or a word (WD), as both are feasible. The paradigm of DOG helps disambiguate these two hypotheses. The hypothesis of copying a full word could not account for DOG-DIM, as [panat-panat] loses to [panat-nat]. On the other hand, copying a heavy syllable accurately captures DOG-DIM, letting [panat-nat] win. Thus, a heavy syllable provides a better analysis. One important goal, then, is to include the crucial disambiguating candidate [panat-panat] in the possible candidate set.

Recall that for concatenative processes, the alternation substitution GEN generates the crucial surface candidates that can disambiguate different grammars, as discussed above in Section 4.3.3. For reduplication, to select among different prosodic templates, the learner must consider the various possible effects of other templates. With this goal in mind, we describe the reduplicative SR GEN below by breaking it down into two steps.

First, the learner needs to form a set of possible shapes of the reduplicants. This is achieved by allowing each segment to surface or not. Naming such a function `Reduplicant`, it generates $2^3 = 8$ possible reduplicants for the base /bed/. These are {bed, be, ed, bd, b, e, d, ∅}. Similarly, `Reduplicant`(/panat/) produces 2^5 (= 32) reduplicants. Another way to conceptualize the possibility of each segment surfacing or not is to consider that each segment alternates with the null segment ∅ only in reduplication. Then, adopting a

[11] This idea resembles the templatic UR approach as discussed in Section 1.3, but not exactly the same because the learner also entertains the possibility of not using a specified prosodic template. See our discussion in Section 4.4.3.

gradual deletion account (e.g., McCarthy, 2008; McCarthy et al., 2018), one could form a more generalized version where these segments might also appear in their reduced forms, generating a larger set of possible candidates.[12] For the purpose of this chapter, we assume the straightforward segment $\sim$ null account.

Given the typology of reduplication and the participants' spontaneous responses discussed in Chapter 2, we know that the reduplicant can be placed to the left of the base (e.g., [be$_{\text{red}}$-bed$_{\text{base}}$]), the right of the base (e.g., [bed$_{\text{base}}$-be$_{\text{red}}$]), infixed into the base (e.g., [pa$_{\text{base}}$-na$_{\text{red}}$-nat$_{\text{base}}$]), or it can "back-copy" the properties of the reduplicant into the base (e.g., [be$_{\text{base}}$-be$_{\text{red}}$]). Gen must be capable of generating these possibilities. We define a function `Reduplicated` for these possible candidates, as shown in (48). The function definition for infixing reduplication is largely inspired by Wilson (2018).

(48) Generating the reduplicated forms for each input, $\texttt{Reduplicated}(u)$

$$\texttt{Reduplicated}(u) = \{u[0:k] + x + u[k:] \,|\, x \in \texttt{Reduplicant}(u), k \in \textsc{Pivot}\} \cup \{x + x \,|\, x \in \texttt{Reduplicant}(u)\}$$

where + indicates string concatenation.

Pivot could be defined according to Yu (2003)'s pivot theory of infixation, which argues that the potential placements of an infix are phonetic and/or psycholinguistic prominent positions. These possible pivots can be edge-oriented (e.g., the first syllable, the final syllable) or prominence-oriented (e.g., the stressed syllable).[13] The learner can further narrow down the possible set of Pivot points from the training data. In our case, [bed-bet] can be categorized as "after the first syllable", or "after the last syllable". [panat-nat] can only be categorized as "following the last syllable". Taking the *intersection* of these placements, the set of Pivot points for this dataset only includes the right word edge. An alternative, richer possibility is to consider the set of Pivot points to be at any position within the base, thus, allowing candidate forms like [b$_{\text{base}}$-bet$_{\text{red}}$-et$_{\text{base}}$]. The learner must learn the right phonology

[12] We think this may ultimately provide a learning-based perspective to the *Copying-Weakening-Implication* discussed in Zimmermann (2021b).

[13] See the typological survey in Yu (2003, §1) for more details.

to eliminate these funky candidates. For the sake of generality, for all simulations, we have included the left-edge candidates to demonstrate that the phonology can handle various possible placements.

One aspect we have not addressed is the ambiguity of the base and reduplicant in the surface form. For instance, at least with [bed-bet], there are two possible analyses: [bed$_{\text{base}}$-bet$_{\text{red}}$] and [bed$_{\text{red}}$-bet$_{\text{base}}$].[14] Consequently, the final output probability for [bed-bet] should be the summed probability of these two candidates. Implementation-wise, when generating candidates, in the function `Reduplicated`, we have assigned the `base` and `red` labels to each portion within a form. In the current simulations, we have not accounted for this ambiguity, but have assumed the output is strictly [bed$_{\text{base}}$-bet$_{\text{red}}$] and [panat$_{\text{base}}$-nat$_{\text{red}}$]. We plan to address this ambiguity in future work.

4.4.3 The family of constraints for reduplication

We follow the constraint set proposed by BRCT, the details of which are reviewed in Section 1.3. BRCT encompasses standard IO faithfulness for the realization of the base, and BR faithfulness for the realization of the reduplicant. The commonly adopted faithfulness constraints and their definitions can be found in Appendix A. (49) describes the full set of constraints we used to handle the toy dataset in Table 4.2.

(49) The adopted constraints and their definitions.

1. *FinalVoicedObs: assigns a violation for any word-final voiced obstruents
2. *VTV: assigns a violation for any intervocalic voiceless obstruents
3. Ident-IO(Voice): assigns a violation for any IO-corresponded segments without the same [Voice] specification
4. Ident-BR(Voice): assigns a violation for any BR-corresponded segments without the same [Voice] specification

[14] These two candidates will have different violation profiles for the base-reduplicant correspondence constraints and the input-base correspondence constraints. For the input /bed/, [bed$_{\text{base}}$-bet$_{\text{red}}$] violates Ident-BR(Voice) but not Ident-IO(Voice), but [bed$_{\text{red}}$-bet$_{\text{base}}$] violates both Ident-BR(Voice) and Ident-IO(Voice). We will make the definitions of these constraints explicit in the next Section (Section 4.4.3).

5. Max-IO: assigns a violation for any segment in the input without a correspondent segment in the output
6. Max-BR: assigns a violation for any segment in the base without a correspondent in the reduplicant
7. Contig-BR: assigns a violation for any non-contiguous copying (e.g., [b.t-bet])
8. *MarginCluster: a cover constraint which assigns a violation for any ill-formed consonant clusters in the margin (both onset and coda; e.g., [b..-bet])
9. AlignRedL: assigns a violation for any intervening segments between the left edge of the reduplicant and the left edge of the word
10. AlignRedR: assigns a violation for any intervening segments between the right edge of the reduplicant and the right edge of the word
11. L-Anchor-BR: assigns a violation if the leftmost segment of the base does not have its correspondent in the reduplicant
12. R-Anchor-BR: assigns a violation if the rightmost segment of the base does not have its correspondent in the reduplicant
13. RED = X: assigns a violation if the reduplicant does not have the shape of a particular hidden structure X (evaluated based on the (hidden prosodic unit, SR) pair)

Our proposal diverges from BRCT in that not *all* templatic constraints are included as a part of the constraint sets. Instead, our constraint set contains only one templatic constraint with Red = X, where X refers to the specific hidden specifications accordingly based on the pair of hidden templates and SR. This constraint assigns violations based on whether the surface realization of Red satisfies the hidden category X. For example, if the hidden prosodic category is σ_{μ}, with the reduplicant [be], the output form does not violate this templatic constraint. However, if the hidden prosodic category is $\sigma_{\mu\mu}$, then a reduplicant [be] in an output incurs a violation of this templatic constraint. In this way, the progression of learning can be thought of as a process that gradually uncovers the phonological specifications of this hidden templates. If the learner successfully learns that $\sigma_{\mu\mu}$ should be the prosodic template for the reduplicative morpheme, then the constraint is RED $= \sigma_{\mu\mu}$.

As a part of the phonological constraint set, this constraint receives some weight. If RED = X receives a negligible weight, the urge to obey the prosodic template for the reduplicant shape is also negligible. On the other hand, if RED = X receives a substantial weight, then, the adherence to the prosodic template for the reduplicant shape is substantial. In this way, the weight of this constraint acts like a gatekeeper, determining the extent to which the hidden prosodic templates influence the shape of the reduplicant.

4.5 The typology of reduplication-phonology interaction

We have outlined all the proposed components for reduplication learning, including the hidden structure of the reduplicative morpheme, the surface form GEN that accounts for reduplicated forms, and the set of constraints. In this section, we examine how the learner responds to different types of reduplication-phonology interactions, including normal application, overapplication, underapplication, and templatic backcopying.

4.5.1 Normal application

The toy dataset presented in Table 4.2 is an instance of the normal application of the reduplication-phonology interactions. Namely, when a phonological process interacts with reduplication, it applies to the right environment and only to the expected environment. Using the input in Table 4.2 and following the described procedures, the learner fed the input tableau with 1,992 pairs of hidden structures and surface forms into the EM learning procedure (Section 4.3). The learned results are presented in Table 4.5, which we will now unpack. For the ease of presentation, in all tableaux below, the reported harmony score $\mathcal{H}$ is based on the included constraints and their weights. The probability score is the actual probability with all considered candidates and constraints.

First, the learner successfully learned the correct underlying representations for the stem

Morphemes	*Learned hidden structures*	p
Cat$_{\text{STEM}}$	/bed/	0.999
Dog$_{\text{STEM}}$	/panat/	1.0
pl	/-a/	1.0
dim	Red = $\sigma_{\mu\mu}$	0.999

*FinalVoicedObs	17.32
*VTV	1.34
Ident-IO(Voice)	9.35
Ident-BR(Voice)	0.05
Max-IO	11.21
Max-BR	6.66
Contig-BR	10.45
*MarginCluster	6.28
AlignRedL	0.0
AlignRedR	5.1
L-Anchor-BR	0.39
R-Anchor-BR	9.02
Red = X	21.44

Table 4.5: *The learned results for the toy dataset*

Cat, and the right prosodic template for Red. It assigns a substantial weight to Red = X, or more precisely, Red $=\sigma_{\mu\mu}$ due to the learned heavy syllable template. In tableau (50), due to $w_{\text{Red}=\sigma_{\mu\mu}} \succ w_{\text{Max-BR}}$, the total reduplication candidate in (50a) loses to the winner (50b) with a heavy syllable template. Max-BR receives some weight to preserve a maximal heavy syllable structure (see 50c). The candidate with a light syllable template, as in (50d), violates many other constraints beyond the templatic constraint, such as R-Anchor-BR. Lastly, the templatic restriction for the reduplicant is not back-copied to the base, due to the substantial weight of Max-IO.

(50)

/panat-Red/	Red $=\sigma_{\mu\mu}$ 21.44	Max-IO 11.21	R-Anchor-BR 9.02	Max-BR 6.66	$\mathcal{H}$	$P(s\,\|\,u)$
a. $[p_1a_2n_3a_4t_{5\ \text{base}}\text{-}p_1a_2n_3a_4t_{5\text{red}}]$	1				21.44	0.000
b. $[p_1a_2n_3a_4t_{5\text{base}}\text{-}n_3a_4t_{5\ \text{red}}]$				2	13.32	0.999
c. $[p_1a_2n_3a_4t_{5\text{base}}\text{-}a_4t_{5\ \text{red}}]$				3	19.98	0.000
d. $[p_1a_2n_3a_4t_{5\ \text{base}}\text{-}n_3a_{4\ \text{red}}]$	1		1	3	50.44	0.000
e. $[n_3a_4t_{5\ \text{base}}\text{-}n_3a_4t_{5\ \text{red}}]$		2			22.42	0.000

Given $w_{\text{R-Anchor-BR}} \succ w_{\text{L-Anchor-BR}}$, and $w_{\text{AlignRedR}} \succ w_{\text{AlignRedL}}$, the grammar must copy phonological material at the right edge, and position the reduplicant at the right edge as well. This ensures a suffixation of the final heavy syllable, as demonstrated in Tableau

(51). First, non-adjacent copying is deemed ill-formed. The candidate that suffixes a copy of the first syllable, as in (51b) is less likely to be the output due to the substantial weight on R-Anchor-BR. Similarly, the candidate that prefixes a copy of the last syllable, as in (51c), loses out due to the substantial weight on AlignRedR. The candidate that prefixes a copy of the first syllable, as in (51d), violates both constraints and is therefore harmonically bounded by other candidates.

(51)

/panat-Red/	AlignRedL 0	AlignRedR 5.1	L-Anchor-BR 0.39	R-Anchor-BR 9.02	$\mathcal{H}$	$P(s\,|\,u)$
a. [$p_1a_2n_3a_4t_{5\text{base}}$-$n_3a_4t_{5\ \text{red}}$]	5		1		0.39	0.999
b. [$p_1a_2n_3a_4t_{5\text{base}}$-$p_1a_2n_{3\ \text{red}}$]	5			1	9.02	0.000
c. [$n_3a_4t_{5\text{red}}$-$p_1a_2n_3a_4t_{5\text{base}}$]		5	1		25.89	0.000
d. [$p_1a_2n_{3\text{red}}$-$p_1a_2n_3a_4t_{5\text{base}}$]		5		1	34.52	0.000

What about those possible forms with suffixing reduplication that satisfy the heavy syllable template but involve copying from both edges? They are ruled out by the substantially weighted Contig-BR and *MarginCluster, as in Tableau (52).

(52)

/panat-Red/	Contig-BR 10.45	*MarginCluster 6.28	Max-BR 6.66	L-Anchor-BR 0.39	$\mathcal{H}$	$P(s\,|\,u)$
a. [$p_1a_2n_3a_4t_{5\text{base}}$-$n_3a_4t_{5\ \text{red}}$]			2	1	13.71	0.999
b. [$p_1a_2n_3a_4t_{5\text{base}}$-$p_1a_4t_{5\ \text{red}}$]	1		2		23.77	0.00
c. [$p_1a_2n_3a_4t_{5\text{base}}$-$p_1n_3a_4t_{5\ \text{red}}$]	1	1	1		23.39	0.000

We have established that the learned grammar correctly enforces suffixation of the word-final heavy syllable, realized by copying a contiguous substring. Now, we turn to the interaction between reduplication and final devoicing. The necessary constraint weighting for non-reduplicated words is very similar to the plain Pseudo-German example discussed in Section 4.3, with $w_{*\text{FinalVociedObs}} \succ w_{\text{Ident-IO(Voice)}} \succ w_{*\text{VTV}}$. Therefore, we will skip further details here and focus on the relevant aspects of the reduplicated forms.

For the reduplicated forms, BR faithfulness constraints are involved in determining which copy undergoes devoicing, as demonstrated in Tableau (53). The stem with an underlying word-final /t/ (/panat/) is straightforward: the winner [panat$_{\text{base}}$-nat$_{\text{red}}$] preserves the voicing in both copies without violating any relevant constraints. As for the stem with an underlying word-final /d/ (/bed/), the critical relative weighting requires $w_{*\textsc{FinalVoicedObs}} \succ w_{\textsc{Ident-BR(Voice)}}$ – this is to ensure the underapplied [bed$_{\text{base}}$-bed$_{\text{red}}$] as in (53b) loses to [bed$_{\text{base}}$-bet$_{\text{red}}$] as in (53a), which undergoes devoicing in the reduplicant as expected. Additionally, the relative weighting $w_{\textsc{Ident-IO(Voice)}} \succ w_{\textsc{Ident-BR(Voice)}}$ guarantees that the overapplied [bet$_{\text{base}}$-bet$_{\text{red}}$] as in (53c) loses to [bed$_{\text{base}}$-bet$_{\text{red}}$] as in (53a).

(53)

	*FinalVoicedObs 17.32	Ident-IO(Voice) 9.35	Ident-BR(Voice) 0.05	$\mathcal{H}$	$P(s \mid u)$
Input: /bed-Red/					
a. [bed$_{\text{base}}$-bet$_{\text{red}}$]			1	0.05	0.999
b. [bed$_{\text{base}}$-bed$_{\text{red}}$]	1			17.32	0.000
c. [bet$_{\text{base}}$-bet$_{\text{red}}$]		1		9.35	0.000
d. [bet$_{\text{base}}$-bed$_{\text{red}}$]	1	1	1	26.72	0.000
Input: /panat-Red/					
a. [panad$_{\text{base}}$-nat$_{\text{red}}$]		1	1	9.38	0.000
b. [panad$_{\text{base}}$-nad$_{\text{red}}$]	1	1		26.67	0.000
c. [panat$_{\text{base}}$-nat$_{\text{red}}$]				0.00	0.999
d. [panat$_{\text{base}}$-nad$_{\text{red}}$]	1		1	17.37	0.000

In sum, all of these constraints are meaningful for the specific range of (hidden structure, SR) pairs we have considered. The learned grammar derives all the observed forms with a probability close to 1. The final learned result accurately reflects the intended grammar and correctly selects the appropriate hidden structures.

4.5.2 Overapplication

4.5.2.1 Normal overapplication

Overapplication describes the kind of reduplication-phonology interactions where the phonological process applies not only in the expected environment but also in an unmotivated environment due to base-reduplicant faithfulness. As illustrated in Table 4.8, devoicing applies at the word-final position, leading the obstruent in the base to surface as a [t] in the second copy. Such a process is not expected to apply to the first copy, because the voiced obstruent in the first copy is not word-final, thus not violating the markedness requirement. However, the first copy undergoes devoicing as well. Different from the previous toy example, to test the generality of this learner, we have made the reduplication process an instance of prefixing reduplication. This example is also a case of normal overapplication: word-final devoicing applies to the base, and then the devoicing requirement is copied into the reduplicant. This is different from back-copying overapplication, which occurs when a phonological process applies to the reduplicant and then gets copied to the base. We will discuss backcopying in Section 4.5.2.2 and in Section 4.5.2.3.

Meanings	*Forms*
CAT$_{\text{STEM}}$	bet
CAT-PL	bed-a
CAT-DIM	be**t**-be**t**
DOG$_{\text{STEM}}$	panat
DOG-PL	panat-a
DOG-DIM	pan-panat

Morphemes	*Intended hidden structures*
CAT$_{\text{STEM}}$	/bed/
DOG$_{\text{STEM}}$	/panat/
PL	/-a/
DIM	RED $= \sigma_{\mu\mu}$

Table 4.6: *Input data of the overapplication toy dataset (left) and the intended URs for each morpheme (right). Final devoicing + heavy syllable prefixation as the phonology. Final devoicing overapplies to the unmotivated environments.*

Feeding the input in Table 4.8 to our learner yields the results shown in Table 4.7. First, as before, the learner successfully learned the correct underlying representations for the stem CAT, and the right prosodic template for RED. $w_{\text{RED}=\sigma_{\mu\mu}} \succ w_{\text{MAX-BR}}$ indicates that the reduplicant must be a heavy syllable. Both CONTIG-BR and *MARGINCLUSTER received non-trivial weight for the well-formed contiguous copy. All these aspects regarding

Morphemes	*Learned hidden structures*	p
Cat$_{\text{STEM}}$	/bed/	0.999
Dog$_{\text{STEM}}$	/panat/	1.0
PL	/-a/	1.0
DIM	Red = $\sigma_{\mu\mu}$	0.999

*FinalVoicedObs	21.35
*VTV	0.30
Ident-IO(Voice)	10.45
Ident-BR(Voice)	10.87
Max-IO	10.74
Max-BR	0.00
Contig-BR	11.53
*MarginCluster	2.32
AlignRedL	4.48
AlignRedR	0.00
L-Anchor-BR	12.15
R-Anchor-BR	0.00
Red = X	12.10

Table 4.7: *The learned results for the normal overapplication toy dataset*

the realization of the copy follow the discussion in the previous section. As for prefixation instead of suffixation, given $w_{\text{L-Anchor-BR}} \succ w_{\text{R-Anchor-BR}}$, and $w_{\text{AlignRedL}} \succ w_{\text{AlignRedR}}$, the copying operation should copy materials at the left edge, and place the reduplicant at the left edge, as illustrated in (54).

(54)

/Red-panat/	AlignRedL 4.48	AlignRedR 0.00	L-Anchor-BR 12.15	R-Anchor-BR 0.00	$\mathcal{H}$	$P(s\,\|\,u)$
a. [$p_1a_2n_3a_4t_{5\text{base}}$-$n_3a_4t_{5\text{ red}}$]	5		1		34.55	0.00
b. [$p_1a_2n_3a_4t_{5\text{base}}$-$p_1a_2n_{3\text{ red}}$]	5			1	22.4	0.000
c. [$n_3a_4t_{5\text{red}}$-$p_1a_2n_3a_4t_{5\text{base}}$]		5	1		12.15	0.000
d. [$p_1a_2n_{3\text{red}}$-$p_1a_2n_3a_4t_{5\text{base}}$]		5		1	0.00	0.999

The derivation of overapplication is as illustrated in Tableau (55). $w_{\text{*FinalVoicedObs}} \succ w_{\text{Ident-IO(Voice)}}$ ensures the devoicing of the word-final obstruent in the base. On the other hand, that $w_{\text{Ident-BR(Voice)}}$ receives a substantial weight ensures that the normal application candidate [bed$_{\text{red}}$-bet$_{\text{base}}$] as in (55a) loses to the overapplication candidate [bet$_{\text{red}}$-bet$_{\text{base}}$] as in (55c).

(55)

	*FinalVoicedObs 21.35	Ident-IO(Voice) 10.45	Ident-BR(Voice) 10.87	$\mathcal{H}$	$P(s\,\vert\,u)$
Input: /Red-bed/					
a. [bed$_{\text{red}}$-bet$_{\text{base}}$]		1	1	21.32	0.000
b. [bed$_{\text{red}}$-bed$_{\text{base}}$]	1			21.35	0.000
c. [bet$_{\text{red}}$-bet$_{\text{base}}$]		1		10.45	0.999
d. [bet$_{\text{red}}$-bed$_{\text{base}}$]	1		1	32.22	0.000
Input: /Red-panat/					
a. [pan$_{\text{red}}$-panad$_{\text{base}}$]	1	1		31.8	0.000
b. [pan$_{\text{red}}$-panat$_{\text{base}}$]				0.000	0.999

4.5.2.2 Backcopying

Table 4.8 shows backcopying overapplication: devoicing applies at the word-final position, leading the word-final obstruent to surface as a [t] in the reduplicant. Such a process is not expected to apply to the base because the voiced obstruent in the base is not word-final. Yet we observe devoicing of the voiced obstruent in the base as well.

Meanings	*Forms*
Cat$_{\text{stem}}$	bet
Cat-pl	bed-a
Cat-dim	be**t**-be**t**
Dog$_{\text{stem}}$	panat
Dog-pl	panat-a
Dog-dim	panat-nat

Morphemes	*Intended hidden structures*
Cat$_{\text{stem}}$	/bed/
Dog$_{\text{stem}}$	/panat/
pl	/-a/
dim	Red = $\sigma_{\mu\mu}$

Table 4.8: *Input data of the backcopying overapplication toy dataset (left) and the intended URs for each morpheme (right). Final devoicing + heavy syllable suffixation as the phonology. Final devoicing overapplies to the unmotivated environments.*

Feeding the input Table 4.8 to our learner yields the results shown in Table 4.9. First, as

[15] This non-trivial weight on *VTV penalizes [bet$_{\text{base}}$-et$_{\text{red}}$]. This weight is needed since Max-BR and L-Anchor-BR receive zero weights. However, the weight is not substantially higher than the other constraints for the voicing process guaranteeing that intervocalic voicing is never wrongly predicted.

Morphemes	*Learned hidden structures*	p
CAT$_{\text{STEM}}$	/bed/	0.999
DOG$_{\text{STEM}}$	/panat/	1.0
PL	/-a/	1.0
DIM	RED = $\sigma_{\mu\mu}$	0.999

*FINALVOICEDOBS	22.19
*VTV	6.99[15]
IDENT-IO(VOICE)	13.38
IDENT-BR(VOICE)	21.8
MAX-IO	7.7
MAX-BR	0
CONTIG-BR	9.32
BADCLUSTER	3.53
ALIGNREDL	0
ALIGNREDR	7.69
L-ANCHOR-BR	0
R-ANCHOR-BR	21.64
RED = X	7.34

Table 4.9: *The learned results for the backcopying overapplication toy dataset*

before, the learner successfully learned the correct underlying representations for the stem CAT, and the right prosodic template for RED. $w_{\text{RED}=\sigma_{\mu\mu}} \succ w_{\text{MAX-BR}}$ indicates that the reduplicant must be a heavy syllable. Given $w_{\text{R-ANCHOR-BR}} \succ w_{\text{L-ANCHOR-BR}}$, and $w_{\text{ALIGNREDR}} \succ w_{\text{ALIGNREDL}}$, the copying operation should copy materials at the right edge, and place the reduplicant at the right edge. CONTIG-BR and BADCLUSTER received non-trivial weight for the well-formed contiguous copy. All these aspects regard the realization of the copying follow the discussion in the case of normal application, and therefore we suppress further discussions here.

The derivation of backcopying overapplication is as illustrated in Tableau (56). Beyond getting $w_{\text{*FINALVOICEDOBS}} \succ w_{\text{IDENT-IO(VOICE)}}$, the critical relative weighting $w_{\text{IDENT-BR(VOICE)}} \succ w_{\text{IDENT-IO(VOICE)}}$ ensures the normal application candidate [bed$_{\text{base}}$-bet$_{\text{red}}$] as in as in (56a) loses to the overapplied candidate [bet$_{\text{base}}$-bet$_{\text{red}}$] as in (56c). This reflects the constraint ranking scheme, with more strength on the markedness constraint and the BR-faithfulness constraint to outweigh the IO-faithfulness constraint. We found that the grammar derives all the observed forms with a probability close to 1, indicating successful learning.

(56)

	*FinalVoicedObs 22.19	Ident-IO(Voice) 13.38	Ident-BR(Voice) 21.8	$\mathcal{H}$	$P(s \mid u)$
Input: /bed-Red/					
a. [bed$_{\text{base}}$-bet$_{\text{red}}$]			1	21.8	0.000
b. [bed$_{\text{base}}$-bed$_{\text{red}}$]	1			22.19	0.000
c. [bet$_{\text{base}}$-bet$_{\text{red}}$]		1		13.38	0.999
d. [bet$_{\text{base}}$-bed$_{\text{red}}$]	1	1	1	57.47	0.000
Input: /panat-Red/					
a. [panad$_{\text{base}}$-nat$_{\text{red}}$]		1	1	35.18	0.000
b. [panad$_{\text{base}}$-nad$_{\text{red}}$]	1	1		35.57	0.000
c. [panat$_{\text{base}}$-nat$_{\text{red}}$]				0.00	0.999
d. [panat$_{\text{base}}$-nad$_{\text{red}}$]	1		1	43.99	0.000

Let us compare the learned grammar from this simulation with that in the previous section (Table 4.5). Beyond the obvious differences related to reduplication-phonology interactions, the sub-component of the grammar responsible for the realization of reduplication also differs from that obtained in normal application simulation. In this case, Max-BR receives zero weight, whereas in the previously obtained grammar, Max-BR was assigned a non-negligible weight. On the other hand, in this simulation, *VTV receives a non-negligible weight, while in the previous simulation, the effect of *VTV was negligible. The substantial weight on *VTV subsequently enhances the relative strength of constraints related to voicing alternations, namely *FinalVoicedObs, Ident-IO(Voice) and Ident-BR(Voice). Note that both learned grammars are capable of deriving the right reduplicative pattern (i.e., suffixing the word-final syllable), indicating that the solution space, excluding the reduplication-devoicing interactions, contains multiple optimal weight vectors. The reason why multiple possible solutions exist is due to ambiguity in determining the culprit. Compared to the winner, the losing candidate of the form [bet$_{\text{base}}$-et$_{\text{red}}$], which satisfies the heavy syllable template, can be ruled out by either Max-BR or *VTV. This ambiguity leads to the bifurcation in the learned grammars, reflecting the possible variations in the learned outcomes.

4.5.2.3 Templatic back-copying

Templatic back-copying refers to the phenomenon where the base is truncated to match the shape of the fixed template. In other words, the templatic effect enforced on the reduplicant is back-copied into the base, making this a special case of overapplication. As shown in Table. 4.10, the light syllable template affects the realization of the base as well, truncating it to a light syllable. In this simulation, our primary interest was whether the templatic effect of the reduplicant, along with its transfer to the base, could be learned. Therefore, for ease of presentation, we have omitted the effects of final devoicing, as they are relatively straightforward.

Meanings	*Forms*
Cat$_{\text{STEM}}$	bed
Cat-PL	bed-a
Cat-DIM	be-be
Dog$_{\text{STEM}}$	panat
Dog-PL	panat-a
Dog-DIM	pa-pa

Morphemes	*Intended hidden structures*
Cat$_{\text{STEM}}$	/bed/
Dog$_{\text{STEM}}$	/panat/
PL	/-a/
DIM	Red = σ_μ

Table 4.10: *Input data of the toy dataset (left) and the intended URs for each morpheme (right). Light syllable copying is realized in the phonology and back copied to the base.*

Given the base truncation, it is reasonable to consider that the learner may not know whether the deletion of the base is due to the underlying representation of the stem, or the results of a process applied to the base. Thus, the learner could entertain the possibility of having /be/ and /pa/ as the underlying representations, leading to the hidden structures outlined in Table 4.11. We slightly modified our SR-Gen so that the candidates [bed-bed] also appear for the UR /be/ for a more complete comparison, and also included Dep-IO and Dep-BR constraints.

As in Table 4.12, the learner successfully learned the intended hidden structures, which include the UR for the stem Dog (/bed/) and the stem Cat (/panat/), as well as the light syllable template for the reduplicative morpheme DIM. The grammar assigns substantial

Morphemes	*Hidden structures*
Cat	UR: {bet, bed, be}
Dog	UR: {panat, pa}
pl	UR: {-a}
dim	Templates: {PrWd, Ft, σ_{μ}, $\sigma_{\mu\mu}$ }

Table 4.11: *Hidden structures considered in this simulation.*

weight to the templatic constraint Red = σ_{μ}.

Morphemes	*Learned hidden structure*	p
Dog$_{\text{STEM}}$	/bed/	0.999
Cat$_{\text{STEM}}$	/panat/	1.0
pl	/-a/	1.0
dim	Red = σ_{μ}	0.999

Max-IO	9.90
Max-BR	11.84
Dep-IO	0.0
Dep-BR	1.12
Contig-BR	10.58
*MarginCluster	5.14
AlignRedL	3.53
AlignRedR	8.84
L-Anchor-BR	1.39
R-Anchor-BR	5.43
L-Anchor-IO	10.02
R-Anchor-IO	0.00
Red = X	26.66

Table 4.12: *The learned results for the templatic backcopying toy dataset*

How does the learned grammar implement templatic backcopying? Comparing the candidate in (57b) with the winner in (57a), we see that the substantial weight on Max-BR prevents any reduction in the size of the reduplicant. Additionally, Max-IO receives substantial weight to avoid further reducing both copies to a light syllable with a singleton vowel (see 57c). When comparing the total reduplicated candidate in (57d) with the winning templatic backcopying candidate in (57a), we can conclude that in this learned grammar, the motivation to backcopy the light syllable into the base is driven by both the BR-faithfulness and the urge to place the reduplicant as close to the left word edge as possible. Lastly, the candidate that back-copies part of the word-final syllable, as seen in (57e), loses to the winner. Although it satisfies the light syllable template and obeys BR-faithfulness, the substantial weight on L-Anchor-IO, which requires the leftmost material to be realized, penalizes this candidate.

(57) Templatic backcopying

/panat-RED/	RED =σ_{μ} 26.66	MAX-IO 9.90	L-ANCHOR-IO 10.02	MAX-BR 11.84	ALIGNREDL 3.53	$\mathcal{H}$	$P(s\,\vert\,u)$
a. [$p_1a_{2\text{base}}$-$p_1a_{2\text{red}}$]		3			2	36.76	0.999
b. [$p_1a_{2\text{base}}$-$a_{2\text{red}}$]		3		1	2	48.6	0.000
c. [$a_{2\text{base}}$-$a_{2\text{red}}$]		4			1	43.13	0.002
d. [$p_1a_2n_3a_4t_{5\text{base}}$-$p_1a_2n_3a_4t_{5\ \text{red}}$]	1				5	44.31	0.000
e. [$n_3a_{4\text{base}}$-$n_3a_{4\ \text{red}}$]		3	1		2	46.76	0.000

4.5.3 Underapplication

Opposite to overapplication, underapplication refers to the type of reduplication-phonology interaction where the phonological process fails to apply to the expected environment to maintain base-reduplicant faithfulness. As illustrated in Table 4.13, both copies restrain the voiced obstruent from devoicing. Unless there exists other independently motivated phonological factors, the space characterized by BRCT (McCarthy and Prince, 1995) lacks a grammar that can derive the categorical result reflected in Table 4.13, since there is no mechanism to balance out the preference of [bet] over [bed] and the preference of [bed$_{\text{base}}$-bed$_{\text{red}}$] over [bet$_{\text{base}}$-bet$_{\text{red}}$].

Meanings	*Forms*
CAT$_{\text{STEM}}$	bet
CAT-PL	bed-a
CAT-DIM	be**d**-be**d**
DOG$_{\text{STEM}}$	panat
DOG-PL	panat-a
DOG-DIM	panat-nat

Morphemes	*Intended hidden structures*
CAT$_{\text{STEM}}$	/bed/
DOG$_{\text{STEM}}$	/panat/
PL	/-a/
DIM	RED = $\sigma_{\mu\mu}$

Table 4.13: *Input data of the toy underapplication dataset (left) and the intended URs for each morpheme (right). Final devoicing + heavy syllable suffixation as the phonology. Final devoicing fails to apply to the motivated environments.*

The learned results are shown in Table 4.14. As before, the learner successfully learned the correct underlying representations for the stem CAT, and the right prosodic template for RED, namely, RED = $\sigma_{\mu\mu}$. The learned grammar enforces suffixing reduplication and

Morphemes	*Learned hidden structure*	p
Dog$_{\text{stem}}$	/bed/	0.999
Cat$_{\text{stem}}$	/panat/	1.0
pl	/-a/	1.0
dim	Red = $\sigma_{\mu\mu}$	0.999

*FinalVoicedObs	8.56
*VTV	1.16
Ident-IO(Voice)	8.56
Ident-BR(Voice)	15.92
Max-IO	11.2
Max-BR	6.70
Contig-BR	10.16
*MarginCluster	6.22
AlignRedL	0
AlignRedR	5.16
L-Anchor	0
R-Anchor	8.62
Red = X	21.18

Table 4.14: *The learned results for the underapplication toy dataset*

copying the word-final heavy syllable, consistent with the grammar in Table 4.5. Therefore, we omit the details here.

Let us focus on the constraints governing voicing alternations, as illustrated in the Tableau in (58). The substantial weight assigned to Ident-BR(Voice) indicates that the two copies in the output candidate must share the same voicing specification, thus favoring [bet$_{\text{base}}$-bet$_{\text{red}}$] and [bed$_{\text{base}}$-bed$_{\text{red}}$] but not [bed$_{\text{base}}$-bet$_{\text{red}}$], nor [bed$_{\text{base}}$-bet$_{\text{red}}$]. This time, the grammar assigns equal weights to Ident-IO(Voice) and *FinalVoicedObs, leading to 50% [bet] and 50% [bed] predicted for the isolation form of Cat. Similarly, for Cat-dim, the grammar predicts 50% [bet$_{\text{base}}$-bet$_{\text{red}}$] and 50% [bed$_{\text{base}}$-bed$_{\text{red}}$].

(58)

	*FinalVoicedObs 8.56	Ident-IO(Voice) 8.56	Ident-BR(Voice) 15.92	$\mathcal{H}$	$P(s\,\vert\,u)$
Input: /bed/					
a. [bed]	1			8.56	0.500
b. [bet]		1		8.56	0.500
Input: /bed-Red/					
a. [bed$_{\text{base}}$-bet$_{\text{red}}$]			1	15.92	0.000
b. [bed$_{\text{base}}$-bed$_{\text{red}}$]	1			8.56	0.500
c. [bet$_{\text{base}}$-bet$_{\text{red}}$]		1		8.56	0.500
d. [bet$_{\text{base}}$-bed$_{\text{red}}$]	1	1	1	33.04	0.000

This final learned grammar exhibits a frequency-matching behavior (Hayes et al., 2009): the training data contained one instance of the normal application of final devoicing in Cat$_{\text{STEM}}$ and one instance of non-application in Cat-dim. Thus, underapplication is predicted to lead to the learning of free variation for the interacting phonological process if no other independently motivated phonological factors are present. This result recapitulates the remarks by McCarthy and Prince (1995, p. 5) on the asymmetry between overapplication and underapplication (also see our discussion in Section 1.3) but from a learning perspective. Though it remains unclear how human learners acquire underapplication, some documented underapplication processes do seem to involve free variation (see, for example, Tagalog in Zuraw, 2002).

4.6 Learning the fixed templates in experiments

In the previous section, we have demonstrated that the proposed learner can handle various attested types of reduplication-phonology interactions with toy examples. In this section, we will discuss the learner's behaviors when prompted to participate in our experiment as a human subject. We will focus on three experiments where participants show a preference for certain fixed templates, namely Expt.1a (singular [dɔvɡə]; plural [dɔv-dɔvɡə]), Expt.2a (singular [pif]; plural [pif-pif]) and Expt.2c (singular [pif]; plural [pi-pif]). We start with an overview of the experimental results.

4.6.1 Review of experimental results: Size-based implications

For all artificial grammar learning experiments presented in Chapter 2, the design was straightforward. Participants were familiarized with a few pairs of singulars and their plurals, which were realized using different reduplicative rules. These familiarized pairs were highly ambiguous because they were compatible with multiple hypotheses of the reduplicative rules. Participants were tested with forms that could tease these hypotheses apart and were asked for their free-production responses. The experimental results revealed many emergent interactions with other segmental phonological processes, such as vowel reduction, cluster simplification, and so on. We leave the modeling task of these behaviors, as well as the great individual variations, to future work. In this section, we will only focus on one particular linguistic dimension as the first step, namely, the shapes of the reduplicant.

(59) The Familiarized patterns in each experiment

Experiment	*Examples*	*Singular*	*Reduplicant*	*Base*
1a	[ˈdɔv.gə] → [dɔvˈdɔv.gə]	ˈ$C_1V_2C_3.C_4V_5$	$C_1V_2C_3$	ˈ$C_1V_2C_3.C_4V_5$
2a	[ˈpif] → [ˈpifpif]	$C_1V_2C_3$	$C_1V_2C_3$	$C_1V_2C_3$
2c	[ˈpif] → ˈpipif]	$C_1V_2C_3$	C_1V_2	$C_1V_2C_3$

The familiarized patterns in each experiment are reviewed above in (59). The convergent results of these three experiments are that participants rapidly extracted reduplicative rules and they generalized in a manner that is sensitive to prosodic templates. In terms of the shapes, in Expt. 1a, participants predominately copied a heavy syllable, though with a considerable rate of CV copying. In Expt. 2a, most participants predominately extended total copying, with some participants exhibiting partial copying. In Expt. 2c, participants predominately extended light syllable copying, though they also exhibited variable shapes (heavy syllable, bisyllabic foot). These results reflected the size-based implications participants drew from the impoverished familiarization phase. In this section, we show that our learner is able to predict these patterns through different mechanics, sometimes relying on prosodic templates and other times solely on constraint interactions. This suggests an emergent account of "a-templatic" approaches (e.g., Gafos, 1998), which argue that the templatic effects do not depend on the specification of prosodic templates (see discussion in Section 1.3).

As before, we assumed that the learner already recognizes the effects of copying, namely identical sub-strings in the surface forms. At this stage, the learner needs to construct a hypothesis about the realization of copying by learning the correct reduplicative template and the right phonological grammar. It is reasonable to view the experimental setting as a morphological inflection task (see review in Kodner, 2022). In the current context, the learner learns to form the plural of a given stem. Different from the task before, we assume the learner already knows the underlying representation for each stem and is prompted by the realization of the particular affix copying.

4.6.2 Expt. 1a: Heavy syllable template

We simplified our learning task further and assumed that the learner has a way to determine that the reduplicant always occurs at the left side of the base,[16] and the base cannot be altered. Hence, for the base with 5 segments, we have only considered 32 candidates. Thus, we limited ourselves to the following constraints, with results shown in Table 4.15. A visualization of the learning trajectory of the parameter estimation is given in Appendix E.2.

Template	p
Wd	0.00
Ft	0.00
$\sigma_{\mu\mu}$	0.99
σ_{μ}	0.00

Max-BR	0.00
*MarginCluster	10.54
Contig-BR	10.54
L-Anchor-BR	10.53
R-Anchor-BR	0.00
Red = X	11.18

Table 4.15: *The learned results for Expt.1a*

When wug-testing the learned grammar with novel stems of varying first syllable shapes, such as [dɛbeɪ],[avdi] and [stæbgə], the learner produced the heavy syllable template: [dɛ-dɛbeɪ], [av-avdi] and [stæb-stæbgə] respectively.

4.6.3 Expt. 2a: Full reduplication

In a similar vein as Expt. 1a, we trained the learner on monosyllabic CVC copying, and the learned results are shown in Table 4.16. This learned grammar exhibits the insignificant role of the prosodic templates in the phonological grammar, as indicated by the relatively low weight on Red = X (w = 0.99). Moreover, the distribution is flat over the possible prosodic templates. Instead of making use of the templates, the learner learns to achieve total reduplication by banning deletion of any segment in the reduplicant, assigning a high weight to Max-BR. In other words, the learned grammar emerged to be an "atemplatic"

[16] See our approach in the previous section (Section 4.5).

account. Wug-testing the system here, any segment deletion leads to a candidate with more violations of the constraints. Hence, the total copy candidate always wins. We think such a learned result occurs because there is already an inherent "prosodic template" floating in the system, i.e. the notion of the base. In the context here, we assumed the base to always be the full stem itself. That said, "a-templatic" approaches are effective when the constraints suffice based on the phonological computation of a sequence of phonological material.

Prosodic Template	p
Wd	0.25
Ft	0.25
$\sigma_{\mu\mu}$	0.25
σ_{μ}	0.25

Max-BR	6.5
*MarginCluster	2.19
Contig-BR	2.18
L-Anchor	4.40
R-Anchor	1.10
Red = X	0.99

Table 4.16: *The learned results for Expt.2a*

4.6.4 Expt. 2c: Light syllable copying

We trained the learner on monosyllabic CV copying, with the learned results provided in Table 4.17. The results show that the learner learned to copy a light syllable and required the learned prosodic templates to enforce the reduplicant shape to always be the light syllable, as reflected by the substantial weight on Red = X. Additionally, it learned to avoid deleting extra segments for light syllable copying, thereby penalizing the possibility of less copying. This outcome aligned with the population-level preference for the reduplicant shape in Expt. 2c.

Prosodic Template	p
Wd	0.00
Ft	0.00
$\sigma_{\mu\mu}$	0.00
σ_{μ}	0.99

Max-BR	8.91
*MarginCluster	0.00
Contig-BR	0.00
L-Anchor-BR	0.00
R-Anchor-BR	0.00
Red = X	18.2

Table 4.17: *The learned results for Expt.2c*

4.7 Discussion and future research

In this chapter, we have introduced a reduplication learner within the framework of the EM-MaxEnt learner proposed by Wang and Hayes (resubmitted). The aim of our proposed learning mechanism is to treat reduplication as copying within the phonological module, allowing it to interact freely with other phonological processes. A key component of achieving this goal is the ability to treat candidate prosodic templates as hidden structures and learn (a). whether prosodic templates are needed in such a process and (b). if so, which prosodic unit is the correct one. Methodologically speaking, our approach is fully transparent and interpretable, employing constraints proposed by BRCT (McCarthy and Prince, 1995) that are theoretically well-motivated and empirically grounded. Empirically, it can handle different types of reduplication-phonology interactions, and account for the preferred reduplicant shapes observed in participants' responses in artificial grammar learning experiments. Although our evaluation of the proposed learner is still at a rather preliminary stage, playing with toy datasets, we are confident in its ability to scale up and ultimately meet the desiderata proposed at the beginning of this chapter.

There are a few areas where we remained somewhat vague during the introduction of our learner, which represents promising lines for future work. First, we did not address how to recognize the copying effects and perform morpheme segmentation for reduplicated strings. Yet this step is non-trivial as it is the first step for the functionality of the learner. We believe this is an ideal place to incorporate our proposal of finite-state buffered machines from Chapter 3 into the picture of learning. Moreover, we think different types of morphophonological processes, including listed allomorphy, reduplication, and segmental alternations, might be distinguishable by allomorph similarity during the process of learning. We are developing methods to unify the learning of these different types of processes into one framework. Lastly, it is important to scale up to more challenging datasets and to broader phenomena involving surface repetitions, such as Pima (Uto-Aztecan, central Arizona) in Riggle (2006) and aggressive reduplication in Zuraw (2002), to further evaluate this learner. The results of our learning experiments revealed not only the preferred reduplicant shapes but also a wide variety of individual grammars. We hypothesize that these individual grammars emerge from different biases over the constraints and hidden structures within the hypothesis space at an individual level. Random initial configurations may provide insight into this question, which we are actively exploring and will address in future work.

CHAPTER 5

General conclusion

5.1 Summary of the dissertation

This dissertation has investigated reduplication, the copying operation in morphophonology, focusing on its computational properties and its learning. Our studies examine a particular linguistic phenomenon through different perspectives, such as experimental studies, mathematical analysis, and computational learning. The artificial grammar learning experiments (Chapter 2) were situated within the typology of reduplicative patterns, supporting Prosodic Morphology (McCarthy and Prince, 1986) as well as the prosodic hierarchy. They further show that human learners are able to extrapolate unbounded copying from bounded inputs, which suggests that unbounded copying should be placed in the hypothesis space of morphophonology. Based on the converging evidence from typological findings and the experimental results, we proposed a computational model for morphophonological structures to include copying but exclude nesting dependencies (Chapter 3). We further examined the mathematical properties of this proposed formal device. Lastly, we introduced a learner capable of learning the hidden unobserved prosodic templates together with the learning of other morphophonological processes (Chapter 4). We discussed the implications of each study in situ. As a whole, this dissertation provides a unified account of different types of reduplication, and a unified account of reduplication with other morphophonological structures, both from the perspectives of the possible grammar and from the perspectives of a possible learner.

5.2 Future directions

There are several important questions that we have not been able to address in this dissertation, which we leave for future research.

In our experiments in Chapter 2, we have only collected experimental data from English speakers because English lacks productive reduplication in its morphophonology. Additionally, we have only explored certain dimensions of variation in reduplicative typology. Our findings largely reflect typological trends from a qualitative perspective. One key question that we are ultimately interested in addressing is the precise relationship between learning biases and typology. We believe that collecting a larger-scale corpus of experimental results would be beneficial for making quantitative predictions, which potentially require us to recruit more participants with more diverse language backgrounds and to investigate more linguistic dimensions.

Likewise, our formal framework (Chapter 2) serves as the foundation for possible variant proposals. It is currently formalized to handle only adjacent copying with two identical repetitions. We have briefly outlined several potential extensions to enrich our model and their formal implications. These include looking at non-adjacent copying, multiple reduplication, and imperfect copying. These extensions could help the formal framework better characterize natural language word sets and the hypothesis space of a human learner. We also discussed potential subclasses of the formal model based on the concept of "mode-determinism." Additionally, we considered how to detail the algorithmic steps involved in recognizing reduplication, with prosodic units providing a potential point of breakthrough. This suggests that a formal proposal based on finite-state buffered machines could serve as an effective parser for phonological strings. Future work should develop and formally investigate these possibilities, and if at all possible, empirically test these lines.

Lastly, the proposed morphophonological learner in Chapter 4 shows promise as a learner for reduplication. It opens new possible directions for future research, such as accounting for the variations observed in the experiments. We believe that this approach to handling reduplication could be applicable to other morphophonological processes, such as infixation and truncation. Additionally, modeling how *rapid* generalization of diverse reduplicative structures is at all possible could provide valuable insights for our phonological theory, theory of language learning, and theory of cognition in general.

APPENDIX A

	IO relation	BR relation
Max	Assign one violation for each segment in the input without a correspondent in the output base e.g., (/$p_1u_2s_3a_4$/, [$p_1u_2s_3$])	Assign one violation for each segment in the base without a correspondent in the reduplicant e.g., [$p_1u_2s_3$-$p_1u_2s_3a_4$]
Dep	Assign one violation for each segment in the output without a correspondent in the input e.g., (/$p_1u_2s_3$/, [$p_1u_2s_3a_4$])	Assign one violation for each segment in the reduplicant without a correspondent in the base e.g., [$p_1u_2s_3a_4$-$p_1u_2s_3$]
Ident[F]	Assign one violation for each pair of IO-corresponded segments without a correspondent e.g., (/$p_1u_2s_3a_4$/, [$b_1u_2s_3a_4$])	Assign one violation for each pair of BR-corresponded segments differ on feature F e.g., [$b_1u_2s_3a_4$-$p_1u_2s_3a_4$]
Contig [1]	Assign one violation if the output is not a contiguous string with respect to the input e.g., (/$p_1u_2s_3a_4$/, [$p_1s_3a_4$])	Assign one violation if the reduplicant is not a contiguous string with respect to the base e.g., [$p_1s_3a_4$-$p_1u_2s_3a_4$]
Anchor	{Right, Left}-Anchor-IO Assign one violation if the element standing at the right/left edge of input has no correspondent in the output e.g., (/$p_1u_2s_3a_4$/, [$u_2s_3a_4$])	{Right, Left}-Anchor-BR Assign one violation if the element standing at the right/left edge of the base has no correspondent in the reduplicant e.g., [$u_2s_3a_4$-$p_1u_2s_3a_4$]

Table A.1: *The commonly adopted families of faithfulness constraints and their definitions. Reduplicants are marked blue.*

[1] I-Contig and O-Contig are not the same (McCarthy and Prince, 1995, p.123): I-Contig indicates there is no skipping in the output while O-Contig means there is no intrusion in the output; /xyz/ → [xz] violates I-Contig but does not violate O-Contig; on the other hand, /xz/ → [xyz] violates O-Contig but does not violate I-Contig. Here, Contig-IO is used as a cover constraint for both configurations of mappings.

APPENDIX B

Supplementary materials to Chapter 2.

B.1 Experimental stimuli.

B.1.1 Experiment Series 1

B.1.1.1 Familiarized items

Nominal stem	**Experiment 1a**	**Experiment 1b**	**Experiment 1c**
	Perfect identity	Vowel reduction: [ə]	Vowel overwriting: [i]
ˈdɔv.gə	dɔv-ˈdɔv.gə	dəv-ˈdɔv.gə	div-ˈdɔv.gə
ˈdɛf.keɪ	dɛf-ˈdɛf.keɪ	dəf-ˈdɛf.keɪ	dif-ˈdɛf.keɪ
ˈtab.neɪ	tab-ˈtab.neɪ	təb-ˈtab.neɪ	tib-ˈtab.neɪ
ˈtæf.ku	tæf-ˈtæf.ku	təf-ˈtæf.ku	tif-ˈtæf.ku
ˈzap.moʊ	zap-ˈzap.moʊ	zəp-ˈzap.moʊ	zip-ˈzap.moʊ
ˈzɔv.gi	zɔv-ˈzɔv.gi	zəv-ˈzɔv.gi	ziv-ˈzɔv.gi
ˈʃæp.mə	ʃæp-ˈʃæp.mə	ʃəp-ˈʃæp.mə	ʃip-ˈʃæp.mə
ˈʃɛb.noʊ	ʃɛb-ˈʃɛb.noʊ	ʃəb-ˈʃæp.mə	ʃib-ˈʃæp.mə

Note that due to the many testing lists, we will refrain from listing the testing items here. They will be made publicly available once the journal article is published.

	Singular	**Expt. 2a & 2b**	**Expt. 2c**
list 1	ˈpif	ˈpif-pif	ˈpi-pif
	ˈbuv	ˈbuv-buv	ˈbu-buv
	ˈkæs	ˈkæs-kæs	ˈkæ-kæs
	ˈdɔz	ˈdɔz-dɔz	ˈdɔ-dɔz
list 2	ˈdiv	ˈdiv-div	ˈdi-div
	ˈpuk	ˈpuk-puk	ˈpu-puk
	ˈzæb	ˈzæb-zæb	ˈzæ-zæb
	ˈtɔf	ˈtɔf-tɔf	ˈtɔ-tɔf
list 3	ˈvib	ˈvib-vib	ˈvi-vib
	ˈkuf	ˈkuf-kuf	ˈku-kuf
	ˈzæd	ˈzæd-zæd	ˈzæ-zæd
	ˈpɔt	ˈpɔt-pɔt	ˈpɔ-pɔt
list 4	ˈkit	ˈkit-kit	ˈki-kit
	ˈpuk	ˈpuk-puk	ˈpu-puk
	ˈbæv	ˈbæv-bæv	ˈbæ-bæv
	ˈdɔz	ˈdɔz-dɔz	ˈdɔ-dɔz
list 5	ˈdiv	ˈdiv-div	ˈdi-div
	ˈzub	ˈzub-zub	ˈzu-zub
	ˈkæs	ˈkæs-kæs	ˈkæ-kæs
	ˈfɔt	ˈfɔt-fɔt	ˈfɔ-fɔt

Testing type	list 1	list 2	list 3	list 4	list 5
Familiar	ˈmʊb	ˈmʊt	ˈmʊs	ˈdʊv	ˈnʊg
	ˈnoʊg	ˈnoʊg	ˈtoʊk	ˈtoʊk	ˈvoʊd
	ˈdɪv	ˈgɪb	ˈgɪz	ˈnɪd	ˈbɪv
	ˈbeɪd	ˈpeɪf	ˈneɪp	ˈmeɪb	ˈmeɪb
Disyllabic CV	ˈdoʊ.gʌf	ˈdoʊ.sʌf	ˈgu.kʌp	ˈzi.vɪb	ˈsoʊ.fɪg
	ˈgeɪ.dʌz	ˈgi.bɪd	ˈteɪ.pʊs	ˈkoʊ.pʌs	ˈti.sɔp
	ˈku.pɪt	ˈkeɪ.pɛt	ˈbi.vɔd	ˈgeɪ.bɪd	ˈgeɪ.bɛz
	ˈti.kɛp	ˈtu.kɔs	ˈkoʊ.tʌf	ˈtu.sɔk	ˈpu.tʊs
Disyllabic CVC	ˈgus.pʌb	ˈsɔd.vʌb	ˈpiv.dɔk	ˈpʊt.fʌs	ˈkeɪd.bʌf
	ˈkɪp.tʊf	ˈtɪp.kɔf	ˈgæz.bʊd	ˈbiz.dɛv	ˈgɛz.dɪb
	ˈtoʊf.kʌs	ˈbʊz.dɛv	ˈtoʊk.pʌf	ˈdɪb.gɔz	ˈtʊp.sɛk
	ˈdeɪz.gɪv	ˈpif.tʌs	ˈdɛb.gɪv	ˈtɛf.kʊp	ˈdɑb.gɪv
Trisyllabic	ˈpɑ.bə.fɔd	ˈsɑ.pi.dɔv	ˈdɔ.bɛ.fɪk	ˈki.fæ.tʌs	ˈteɪ.poʊ.vɔk
	ˈgi.zɑ.bʌv	ˈtæ.kʊ.pɛf	ˈku.sɪ.tʊp	ˈtu.sɪ.fɔv	ˈgɔ.vi.dɛz
	ˈkoʊ.sæ.tɪf	ˈpi.su.fʌt	ˈbæ.dɑ.gɔz	ˈgɔ.bɑ.dʊz	ˈpoʊ.sɪ.tʌf
	ˈteɪ.fɛ.kəs	ˈdɛ.vɔ.gʊz	ˈti.fæ.pəs	ˈbæ.zɔ.gɪd	ˈdi.zeɪ.gɛb
Pentasyllabic	ˌboʊ.fɛ.ˈvu.pi.sɪk	ˌveɪ.kɛ.ˈfɑ.zʊ.bɪp	ˌfɪ.də.ˈzæ.kɑ.bɔt	ˌvɔ.boʊ.ˈfi.tæ.kɪs	ˌdoʊ.vi.ˈbɑ.zʊ.sɔk
	ˌteɪ.pɪ.ˈboʊ.gæ.kʊs	ˌdɛ.və.ˈgu.pɑ.zʌb	ˌbu.zɪ.ˈfɑ.soʊ.gʊd	ˌgɛ.zʊ.ˈbɑ.ki.vɪd	ˌgɑ.zu.ˈfæ.vi.dɪb
	ˌdu.væ.ˈfɑ.tɪ.gɛb	ˌgu.zi.ˈtoʊ.fɔ.vɛd	ˌpɛ.sɑ.ˈteɪ.dæ.kʌf	ˌkeɪ.sæ.ˈdɛ.gɪ.pɔt	ˌkɪ.pʊ.ˈzu.dɑ.tɛf
	ˌgɛ.zɑ.ˈseɪ.kɔ.dɪv	ˌtoʊ.fɑ.ˈveɪ.gɪ.sɪk	ˌti.pæ.ˈgɔ.fu.sɛk	ˌtɑ.ki.ˈzeɪ.vɛ.sʌp	ˌpi.sæ.ˈgoʊ.bɛ.kʊt

B.1.2 Experiment series 2

B.1.2.1 Familiarized items

B.1.2.2 Testing items

B.2 Bayesian multinomial logistic regression models

Possible response shape $\in \{l_1, l_2, l_3 \dots l_n\}$

$P(\text{Shape } k \mid \text{participant } i \text{ \& test type } j) = \text{softmax}(\beta_j^{shape} + \lambda_{ij}^{shape})_k$

Fixed effects

$$\beta_j^{shape} = [\beta_j^{shape}(l_1), \beta_j^{shape}(l_2), \beta_j^{shape}(l_3), \dots, \beta_j^{shape}(l_n) = 0]$$

$$\beta_{jk}^{shape} \sim \mathcal{N}(0, \sigma_\beta^{shape}) \qquad \sigma_\beta^{shape} \sim \text{Inv-Gamma}(2, 1)$$

Random effects

$$\lambda_{ij}^{shape} = [\lambda_{ij}^{shape}(l_1), \lambda_{ij}^{shape}(l_2), \lambda_j^{shape}(l_3), \dots, \lambda_j^{shape}(l_n) = 0]$$

$$\lambda_{ijk}^{shape} \sim \mathcal{N}(0, \sigma_\lambda^{shape}) \qquad \sigma_\lambda^{shape} \sim \text{Exponential}(1/2)$$

B.3 Monte Carlo simulations for Experiment Series 2

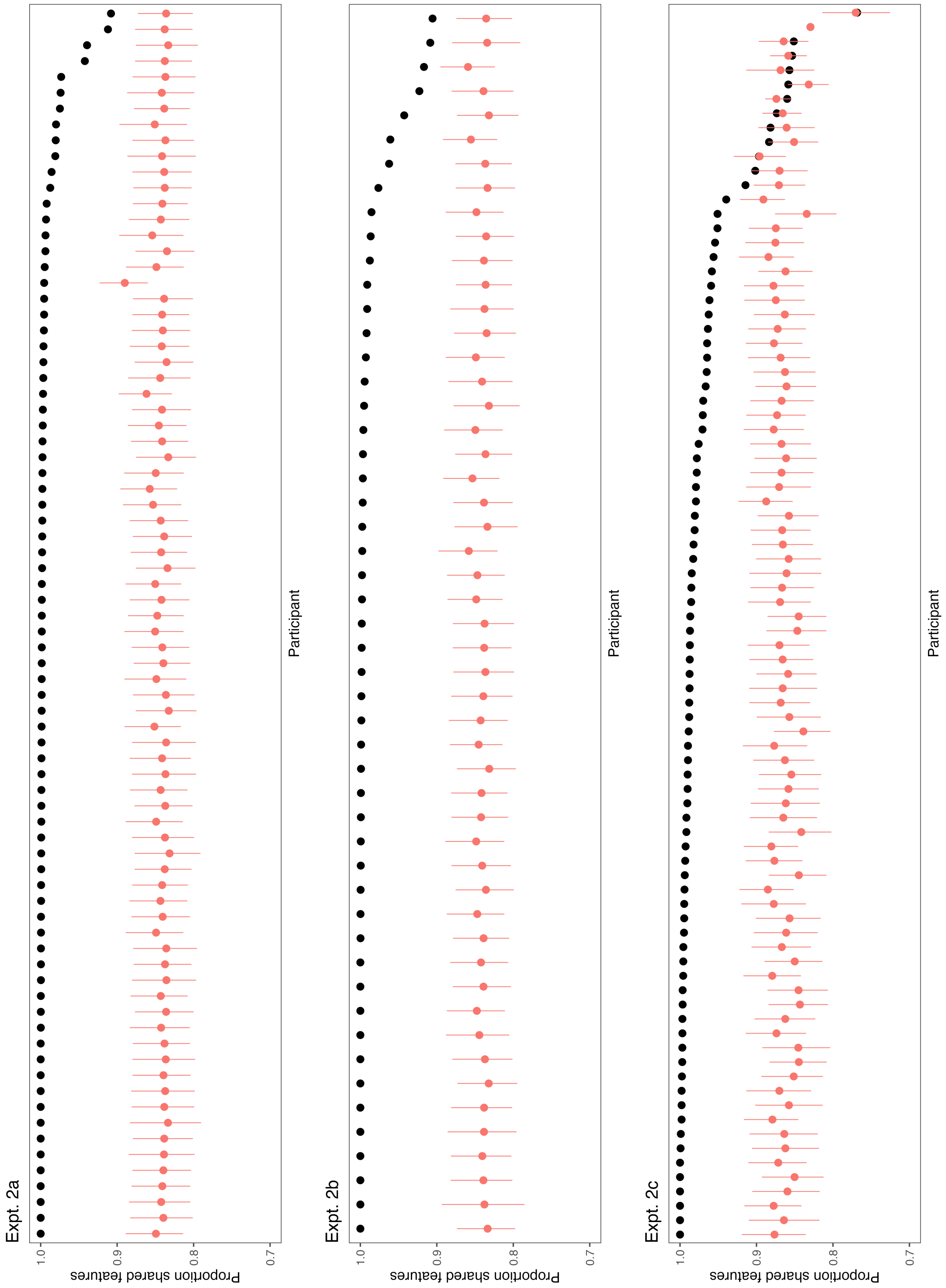
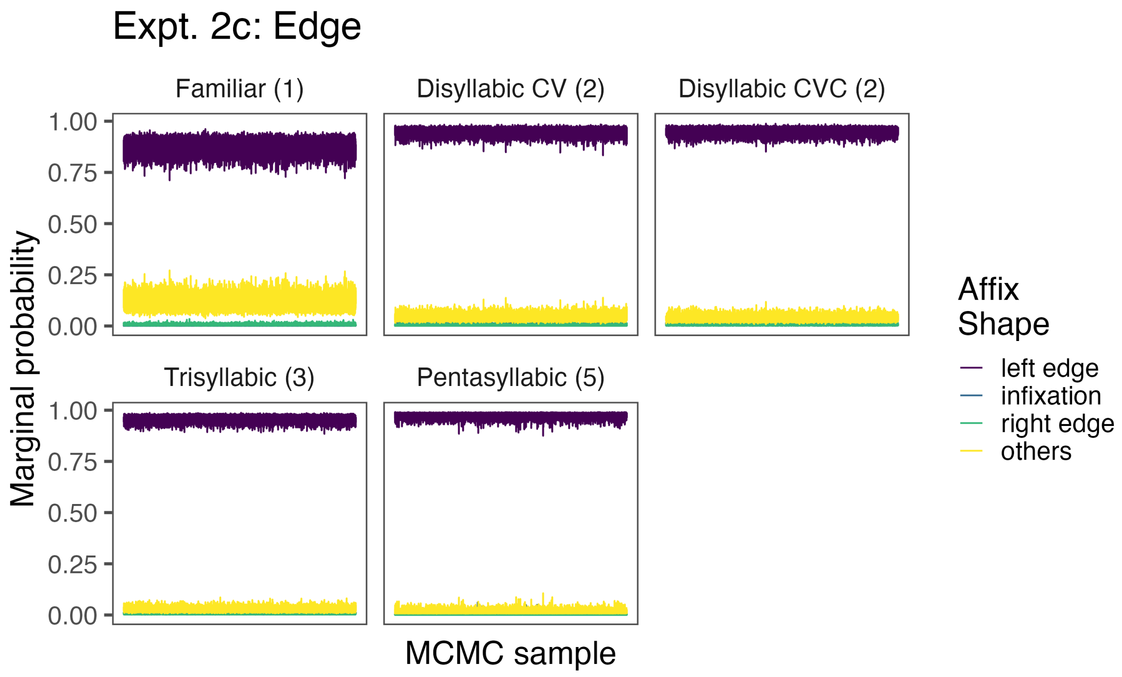


Figure B.1: *The average segment similarity between the affixes and the novel singular bases. Black dots: the observed responses. Red dots: the mean calculated in the Monte Carlo procedure (R = 10000) with bars indicating the 99% confidence interval of chance.*

B.4 Expt. 2c edge analysis



APPENDIX C

Proof of Theorem 1

Lemma 2. *For any string w, if $w \in L(M_1 \cap M_2)$, then $w \in L(M_1)$ and $w \in L(M_2)$.*

Proof. Assume $M_1 = \langle Q_1, \Sigma, I_1, F_1, G_1, H_1, \delta_1 \rangle$ and $M_2 = \langle Q_2, \Sigma, I_2, F_2, \delta_2 \rangle$.

Let the run on $M_1 \cap M_2$ that generates w be $D_0, D_1, \ldots, D_m$, where each $D_i = (u_i, (p_i, q_i, \mathbf{A}_i), v_i, m_i)$. We define a sequence $C_0, C_1, \ldots, C_m$ of configurations of M_1, and a sequence $B_0, B_1, \ldots, B_m$ of configurations of M_2, as follows:

$$C_i = (u_i, p_i, v_i, m_i)$$

$$B_i = \begin{cases} (v_i \backslash u_i, q_i) & \text{if } m_i = \textsc{b} \text{ and } (p_i, q_i, \mathbf{A}_i) \in (H_1 \times Q_2 \times \{\mathbf{0}\}) = H \\ (u_i, q_i) & \text{otherwise} \end{cases}$$

For the initial configuration $D_0 = (w, (p_0, q_0, \mathbf{A}_0), \epsilon, \textsc{n})$, we know that $(p_0, q_0, \mathbf{A}_0) \in I$, so $p_0 \in I_1$ and $q_0 \in I_2$. Therefore $C_0 = (w, p_0, \epsilon, n)$ is a valid starting configuration for a run of w on M_1, and $B_0 = (w, q_0)$ is a valid starting configuration for a run of w on M_2.

For the final configuration $D_m = (\epsilon, (p_m, q_m, \mathbf{A}_m), \epsilon, \textsc{n})$, we know that $(p_m, q_m, \mathbf{A}_m) \in F$, so $p_m \in F_1$ and $q_m \in F_2$. Therefore $C_m = (\epsilon, p_m, \epsilon, \textsc{n})$ is a valid ending configuration for a run on M_1, and $B_m = (\epsilon, q_m)$ is a valid ending configuration for a run on M_2.

To use the sequences $C_0, \ldots, C_m$ and $B_0, \ldots, B_m$ to establish that $w \in L(M_1)$ and $w \in L(M_2)$, we will show that, for every $i \in \{0, \ldots, m-1\}$, $C_i \vdash^*_{M_1} C_{i+1}$ and $B_i \vdash^*_{M_2} B_{i+1}$.

For each $i \in \{0, \ldots, m-1\}$, we know that $D_i \vdash_{M_1 \cap M_2} D_{i+1}$, so there are four cases to consider:

- Suppose $D_i \vdash_{\mathrm{N}} D_{i+1}$. Then $D_i = (xu_{i+1}, (p_i, q_i, \mathbf{A}_i),$
  $\epsilon, \mathrm{N})$ and $D_{i+1} = (u_{i+1}, (p_{i+1}, q_{i+1}, \mathbf{A}_{i+1}), \epsilon, \mathrm{N})$, with $((p_i, q_i, \mathbf{A}_i), x, (p_{i+1}, q_{i+1}, \mathbf{A}_{i+1})) \in \delta$, $(p_i, q_i, \mathbf{A}_i) \notin G$, and $(p_{i+1}, q_{i+1}, \mathbf{A}_{i+1}) \notin H$. Then $C_i = (xu_{i+1}, p_i, \epsilon, \mathrm{N})$, $C_{i+1} = (u_{i+1}, p_i, \epsilon, \mathrm{N})$, $B_i = (xu_{i+1}, q_i)$ and $B_{i+1} = (u_{i+1}, q_{i+1})$. We want to show that $C_i \vdash^*_{M_1} C_{i+1}$ and that $B_i \vdash^*_{M_2} B_{i+1}$.

  - Suppose the critical transition is in δ_{N}. Then $(p_i, x, p_{i+1}) \in \delta_1$ and $p_i \notin G_1$ and $p_{i+1} \notin H_1$, so $C_i \vdash_{\mathrm{N}} C_{i+1}$. Also either $(q_i, x, q_{i+1}) \in \delta_2$, or $x = \epsilon$ and $q_i = q_{i+1}$; so $B_i \vdash^* B_{i+1}$.

  - Suppose the critical transition is in $\delta_{\mathrm{N}\to\mathrm{B}}$. Then $x = \epsilon$, and $p_i = p_{i+1}$ and $q_i = q_{i+1}$. Therefore $C_i = C_{i+1}$ and $B_i = B_{i+1}$.

  - The critical transition cannot be in δ_{B}, because Lemma-1 implies that $\mathbf{A}_i = \mathbf{0}$.

  - The critical transition cannot be in $\delta_{\mathrm{B}\to\mathrm{N}}$, because Lemma-1 implies that $\mathbf{A}_i = \mathbf{0}$.

- Suppose $D_i \vdash_{\mathrm{N}\to\mathrm{B}} D_{i+1}$. Then $D_i = (u_i, (p_i, q_i, \mathbf{A}_i), \epsilon, \mathrm{N})$ and $D_{i+1} = (u_i, (p_i, q_i, \mathbf{A}_i), \epsilon, \mathrm{B})$, with $(p_i, q_i, \mathbf{A}_i) \in G$ and therefore $p_i \in G_1$. So $C_i = (u_i, p_i, \epsilon, \mathrm{N})$ and $C_{i+1} = (u_i, p_i, \epsilon, \mathrm{B})$, and therefore $C_i \vdash_{\mathrm{N}\to\mathrm{B}} C_{i+1}$. Furthermore $B_i = B_{i+1} = (u_i, q_i)$, since $\mathbf{A}_i = \mathbf{A}_\epsilon \neq \mathbf{0}$, so $B_i \vdash^* B_{i+1}$.

- Suppose $D_i \vdash_{\mathrm{B}} D_{i+1}$. Then $D_i = (xu_{i+1}, (p_i, q_i, \mathbf{A}_i), v_i, \mathrm{B})$ and $D_{i+1} = (u_{i+1}, (p_{i+1}, q_{i+1},$
  $\mathbf{A}_{i+1}),$
  $v_i x, \mathrm{B})$, with $((p_i, q_i, \mathbf{A}_i), x, (p_{i+1}, q_{i+1}, \mathbf{A}_{i+1})) \in \delta$, $(p_i, q_i, \mathbf{A}_i) \notin H$ and $(p_{i+1}, q_{i+1}, \mathbf{A}_{i+1}) \notin G$. So $C_i = (xu_{i+1}, p_i, v_i, \mathrm{B})$ and $C_{i+1} = (u_{i+1}, p_{i+1}, v_i x, \mathrm{B})$, but B_i and B_{i+1} will depend on the sub-cases below. There are four sub-cases to consider.

  - The critical transition cannot be in δ_{N}, since Lemma-1 implies that $\mathbf{A}_i = \mathbf{A}^{M_2}_{v_i} \neq \mathbf{0}$.

  - The critical transition cannot be in $\delta_{\mathrm{N}\to\mathrm{B}}$, since Lemma-1 implies that $\mathbf{A}_i = \mathbf{A}^{M_2}_{v_i} \neq \mathbf{0}$.

  - Suppose the critical transition is in δ_{B}. Then $(p_i, x, p_{i+1}) \in \delta_1$ and $p_i \notin H_1$ and $p_{i+1} \notin G_1$. Therefore $C_i \vdash_{\mathrm{B}} C_{i+1}$. Now consider B_i and B_{i+1}. Since $(p_i, q_i, \mathbf{A}_i) \notin H$ we know that $B_i = (xu_{i+1}, q_i)$. Also, we know $\mathbf{A}_{i+1} = \mathbf{A}_i \mathbf{A}_x \neq \mathbf{0}$, so $(p_{i+1}, q_{i+1}, \mathbf{A}_{i+1}) \notin H$ and $B_{i+1} = (u_{i+1}, q_{i+1})$. Finally, either $(q_i, x, q_{i+1}) \in \delta_2$, or $x = \epsilon$ and $q_i = q_{i+1}$; so in either case $B_i \vdash^* B_{i+1}$.

  - Suppose the critical transition is in $\delta_{\textsc{b}\to\textsc{n}}$. Then $x = \epsilon$ and $p_i = p_{i+1}$, so $C_i = C_{i+1}$. Also $\mathbf{A}_i \neq \mathbf{0}$, so $B_i = (u_{i+1}, q_i)$. Furthermore, $p_{i+1} \in H_1$ and $\mathbf{A}_{i+1} = \mathbf{0}$, so $B_{i+1} = (v_i \backslash u_{i+1}, q_{i+1})$. And we know that $v_i \backslash u_{i+1}$ is defined, because the configuration D_{i+1} is part of a successful run and its state $(p_{i+1}, q_{i+1}, \mathbf{A}_{i+1}) \in H$, so the step to D_{i+2} must involve matching an initial portion of the string u_{i+1} against the buffered string v_i. Finally, we also know from the definition of $\delta_{\textsc{b}\to\textsc{n}}$ that the (q_i, q_{i+1}) entry of $\mathbf{A}_i = \mathbf{A}^{M_2}_{v_i}$ is 1, so $q_{i+1} \in \delta_2^*(q_i, v_i)$. Therefore $B_i = (u_{i+1}, q_i) \vdash^*_{M_2} (v_i \backslash u_{i+1}, q_{i+1}) = B_{i+1}$.

- Suppose $D_i \vdash_{\textsc{b}\to\textsc{n}} D_{i+1}$. Then $D_i = (vu_{i+1}, (p_i, q_i, \mathbf{A}_i), v, \textsc{b})$ and $D_{i+1} = (u_{i+1}, (p_i, q_i, \mathbf{A}_i), \epsilon, \textsc{n})$, with $(p_i, q_i, \mathbf{A}_i) \in H$. Therefore $C_i = (vu_{i+1}, p_i, v, \textsc{b})$ and $C_{i+1} = (u_{i+1}, p_i, \epsilon, \textsc{n})$, and $p_i \in H_1$, so $C_i \vdash_{\textsc{b}\to\textsc{n}} C_{i+1}$. Since $(p_i, q_i, \mathbf{A}_i) \in H$, $B_i = (v \backslash vu_{i+1}, q_i) = (u_{i+1}, q_i)$. But also $B_{i+1} = (u_{i+1}, q_i)$. So $B_i = B_{i+1}$.

Therefore $C_0 \vdash^*_{M_1} C_m$, so $w \in L(M_1)$. Similarly, $B_0 \vdash^*_{M_2} B_m$, so $w \in L(M_2)$.

□

Lemma 3. *For any string w, if $w \in L(M_1)$ and $w \in L(M_2)$, then $w \in L(M_1 \cap M_2)$.*

Proof. Assume $w = x_1 x_2 x_3 \ldots x_n \in L_1$ and $w \in L_2$, N.T.S that $w \in L_M$.
$\because w \in L_1$ and $w \in L_2$
$\therefore$ there exists a sequence of configurations $C_0, C_1, C_2 \ldots . C_m$ with

- $C_0 = (w, p_0, \epsilon, \textsc{n})$ with $p_0 \in I_1$
- $C_m = (\epsilon, p_m, \epsilon, \textsc{n})$ with $p_m \in F_1$
- $\forall 0 \leq i < m$, $C_i \vdash_{M_1} C_{i+1}$

and there's a function $f : \textsc{suffix}(w) \to Q_2$ such that $f(w) \in I_2$ and $f(\epsilon) \in F_2$ and $\forall x \in \Sigma, v \in \Sigma^*$, $(f(xv), x, f(v)) \in \delta_2$.

For each $i \in \{0, \dots, m\}$, we take $C_i = (u_i, p_i, v_i, m_i)$, and define D_i to be a configuration of $M_1 \cap M_2$ as follows:

$$D_i = \begin{cases} (u_i, (p_i, f(u_i), \mathbf{0}), v_i, \mathrm{N}) & \text{if } m_i = \mathrm{N} \\ (u_i, (p_i, f(u_i), \mathbf{A}^{M_2}_{v_i}), v_i, \mathrm{B}) & \text{if } m_i = \mathrm{B} \end{cases}$$

First, notice that $D_0 = (w, (p_0, f(w), \mathbf{0}), \epsilon, \mathrm{N})$, where $p_0 \in I_1$ and $f(w) \in I_2$, so D_0 is a valid starting configuration for a run of w on $M_1 \cap M_2$. Similarly, $D_m = (\epsilon, (p_m, f(\epsilon), \mathbf{0}), \epsilon, \mathrm{N})$, where $p_m \in F_1$ and $f(\epsilon) \in F_2$, so D_m is a valid ending configuration for a run on $M_1 \cap M_2$. To show that $w \in L(M_1 \cap M_2)$, we will show that for each $i \in \{0, \dots, m-1\}$, $D_i \vdash^*_{M_1 \cap M_2} D_{i+1}$, which implies that $D_0 \vdash^*_{M_1 \cap M_2} D_m$.

For each $i \in \{0, \dots, m-1\}$, we know that $C_i \vdash_{M_1} C_{i+1}$, so there are four cases to consider.

- Suppose $C_i \vdash_{\mathrm{N}} C_{i+1}$. Then $C_i = (xu_{i+1}, p_i, \epsilon, \mathrm{N})$ and $C_{i+1} = (u_{i+1}, p_{i+1}, \epsilon, \mathrm{N})$ where $(p_i, x, p_{i+1}) \in \delta_1$ and $p_i \notin G$ and $p_{i+1} \notin H$. Therefore $D_i = (xu_{i+1}, (p_i, f(xu_{i+1}), \mathbf{0}), \epsilon, \mathrm{N})$ and $D_{i+1} = (u_{i+1}, (p_{i+1}, f(u_{i+1}), \mathbf{0}), \epsilon, \mathrm{N})$, with $(f(xu_{i+1}), x, f(u_{i+1})) \in \delta_2$. So $D_i \vdash_{\mathrm{N}} D_{i+1}$, since $(p_i, f(xu_{i+1}), \mathbf{0}) \notin G$ and $(p_{i+1}, f(u_{i+1}), \mathbf{0}) \notin H$.

- Suppose $C_i \vdash_{\mathrm{N} \to \mathrm{B}} C_{i+1}$. Then $C_i = (u, p, \epsilon, \mathrm{N})$ and $C_{i+1} = (u, p, \epsilon, \mathrm{B})$, where $p \in G_1$. Therefore $D_i = (u, (p, f(u), \mathbf{0}), \epsilon, \mathrm{N})$ and $D_{i+1} = (u, (p, f(u), \mathbf{A}^{M_2}_{\epsilon}), \epsilon, \mathrm{B})$, and we need to show that $D_i \vdash^*_{M_1 \cap M_2} D_{i+1}$.

  - Since $p \in G_1$, the automaton $M_1 \cap M_2$ has a transition $((p, f(u), \mathbf{0}), \epsilon, (p, f(u), \mathbf{A}^{M_2}_{\epsilon})) \in \delta_{\mathrm{N} \to \mathrm{B}}$. Therefore $D_i \vdash_{\mathrm{N}} (u, (p, f(u), \mathbf{A}^{M_2}_{\epsilon}), \epsilon, \mathrm{N})$.
  - Since $p \in G_1$, we know that $(p, f(u), \mathbf{A}^{M_2}_{\epsilon}) \in G$, and therefore $(u, (p, f(u), \mathbf{A}^{M_2}_{\epsilon}), \epsilon, \mathrm{N}) \vdash_{\mathrm{N} \to \mathrm{B}} (u, (p, f(u), \mathbf{A}^{M_2}_{\epsilon}), \epsilon, \mathrm{B}) = B_{i+1}$.

  Therefore $D_i \vdash^*_{M_1 \cap M_2} D_{i+1}$.

- Suppose $C_i \vdash_{\mathrm{B}} C_{i+1}$. Then $C_i = (xu_{i+1}, p_i, v_i, \mathrm{B})$ and $C_{i+1} = (u_{i+1}, p_{i+1}, v_i x, \mathrm{B})$, with $p_i \notin H_1$ and $p_{i+1} \notin G_1$. Therefore $D_i = (xu_{i+1}, (p_i, f(xu_{i+1}), \mathbf{A}^{M_2}_{v_i}), v_i, \mathrm{B})$ and $D_{i+1} = (u_{i+1}, (p_{i+1}, f(u_{i+1}), \mathbf{A}^{M_2}_{v_i x}), v_i x, \mathrm{B})$, with $(f(xu_{i+1}), x, f(u_{i+1})) \in \delta_2$. Since $p_i \notin H_1$ and

$p_{i+1} \notin G_1$ and $\mathbf{A}_{v_i}^{M_2} \neq \mathbf{0}$, the automaton $M_1 \cap M_2$ has a transition $((p_i, f(xu_{i+1}), \mathbf{A}_{v_i}^{M_2}), x,$ $(p_{i+1}, f(u_{i+1}), \mathbf{A}_{v_i}^{M_2} \mathbf{A}_x^{M_2})) \in \delta_{\text{B}}$.

- Suppose $C_i \vdash_{\text{B}\to\text{N}} C_{i+1}$. Then $C_i = (v_i u_{i+1}, p, v_i, \text{B})$ and $C_{i+1} = (u_{i+1}, p, \epsilon, \text{N})$, with $p \in H_1$. Therefore $D_i = (v_i u_{i+1}, (p, f(v_i u_{i+1}), \mathbf{A}_{v_i}^{M_2}), v_i, \text{B})$ and $D_{i+1} = (u_{i+1}, (p, f(u_{i+1}), \mathbf{0}),$ $\epsilon, \text{N})$, with $f(u_{i+1}) \in \delta_2^*(f(v_i u_{i+1})\ , v_i)$. We need to show that $D_i \vdash_{M_1 \cap M_2}^* D_{i+1}$.

  - Since $p \in H_1$ and the $(f(v_i u_{i+1}), f(u_{i+1}))$ entry of the matrix $\mathbf{A}_{v_i}^{M_2}$ must be 1, we know that the automaton $M_1 \cap M_2$ has a transition $((p, f(v_i u_{i+1}), \mathbf{A}_{v_i}^{M_2}), \epsilon, (p, f(u_{i+1}),$ $\mathbf{0})) \in \delta_{\text{B}\to\text{N}}$. Therefore $D_i \vdash_{\text{B}} (v_i u_{i+1}, (p, f(u_{i+1}), \mathbf{0}), v_i, \text{B})$.
  - Since $p \in H_1$, we know that $(p, f(u_{i+1}), \mathbf{0}) \in H$, and therefore $(v_i u_{i+1}, (p, f(u_{i+1}), \mathbf{0}),$ $v_i, \text{B}) \vdash_{\text{B}\to\text{N}} (u_{i+1}, (p, f(u_{i+1}), \mathbf{0}), \epsilon, \text{N}) = D_{i+1}$.

  Therefore $D_i \vdash_{M_1 \cap M_2}^* D_{i+1}$.

Therefore $D_0 \vdash_{M_1 \cap M_2}^* D_m$, i.e. $(w, (p_0, f(w), \mathbf{0}), \epsilon, \text{N}) \vdash_{M_1 \cap M_2}^* (\epsilon, (p_m, f(\epsilon), \mathbf{0}), \epsilon, \text{N})$, and so $w \in L(M_1 \cap M_2)$.

□

APPENDIX D

Equivalence of regular-copying expressions to FSBMs

We show here that RCEs and FSBMs are equivalent in terms of expressivity: namely, the languages accepted by FSBMs are precisely the languages denoted by RCEs. We prove this statement in two directions: 1) every RCE has a corresponding FSBM; 2) every language recognized by FSBMs can be denoted by an RCE.

Theorem 6. *Let R be a regular copying expression. Then, there exists an FSBM that recognizes $\mathcal{L}(R)$.*

Proof. We complete our proof by induction on the number of operators in R.

Base case: zero operators R must be ϵ, $\emptyset$, a for some symbol a in Σ. Then, standard method to construct corresponding FSAs, thus FSBMs, meet the requirements.

Inductive step: One or more operators In induction, we assume this theorem holds for RCEs with less than n operators with $n \geq 1$. Let R have n operators. There are two cases: 1): $R = R_1^C$; 2): $R \neq R_1^C$;

- Case 1: $R = R_1^C$. Then, we know R_1 must be a regular expression and we can construct an FSA for R_1. Assume there's an FSA $M_0 = \langle Q', \Sigma, I', F', \delta' \rangle$ that recognizes $L(R_1)$. Let $M = \langle Q, \Sigma, I, F, \delta, G, H \rangle$ with
  - $Q = Q' \cup \{q_0, q_f\}$
  - $G = I = \{q_0\}$
  - $H = F = \{q_f\}$
  - $\delta = \delta' \cup \{(q_0, \epsilon, q) \mid q \in I'\} \cup \{(q, \epsilon, q_f) \mid q \in F'\}$

As part of this construction, we add another initial state q_0 and a final state q_f and use them as the *only* initial and final states in the new machine. We add ϵ-arcs 1) from the new initial state q_0 to the previous initial states, and 2) from the previous final states to the new final state q_f. The key component is to add the copying mechanism: G, H, and special arcs. Let G contain only the initial state q_0, which would put the machine to B mode before it takes any transitions. Let H contain only the final state q_f, which stops the machine from buffering and sends it to string matching. Thus, if w is in $L(R_1)$, ww must be in the language accepted by this complete-path FSBM and nothing beyond. Figure D.1 shows such a construction. The proof showing L(M) = L(R) is suppressed here.

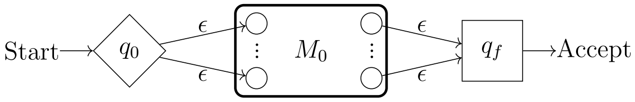


Figure D.1: *The construction used in converting the copy expression R_1^C to a finite-state buffered machine. $L(M_0) = L(R_1)$.*

- Case 2: when $R \neq R_1^C$ for some R_1, we know it has to be made out of the three operations: for some R_1 and R_2, $R = R_1 + R_2$, or $R = R_1R_2$ or $R = R_1^*$. Because R_1 and R_2 have operators less than i, from the induction hypothesis, we can construct FSBMs for R_1 and R_2 respectively. Using the constructions in Theorem 4, we can construct the new FSBM for R.

□

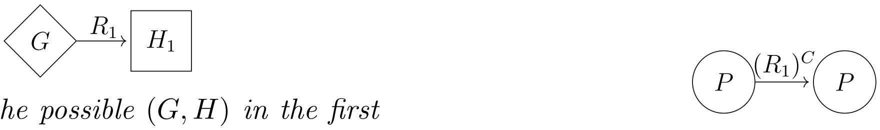
Theorem 7. *If a language L is recognized by an FSBM, then L could be denoted by a RCE.*

Instead of diving into proof details, we introduce the most crucial fragments to the full FSBM-to-RCE conversion: how the copying mechanism in a complete-path FSBM is converted into a copy expression. We leave out parts that use basic ideas of FSA-to-RE conversion, which can be found in Hopcroft and Ullman (1979, pp. 33–34).

The previous discussion on the realization of the copying mechanism in complete-path FSBMs concluded with three aspects 1) the specification of G states, 2) the specification of H states, and 3) the *completeness restriction* which imposes ordering requirements on G and H. Thus, to start with, we want to concentrate on the areas selected by G states and H states in a machine, as they are closely related to the copying mechanism.

The core is to treat any G state and H state pair as an small FSA: if the paths along the pair do not cross other special states, borrow the FSA-to-RE conversion to get a regular expression R_1, denoting the languages possible to be stored in the buffer temporarily. Importantly, there are only finitely many (G, H) pairs. Iterating through all possible paths between these two states and getting a general RE R_1 by union, we use two plain states with the RCE R_1 along the arc to denote the languages from that specific G to H. Then we plug them back into the starting FSBM.

All special states are eliminated. Thus, we get an intermediate representation with only plain states. Similar ideas as FSA-to-RE conversion could be applied again to get the final regular copying expression for this FSBM. The described conversion of the copying mechanism in a machine to a copy expression is depicted in Figure D.2.



(a) *Goal for the possible (G, H) in the first steps of the FSBM-to-RCE conversion*

(b) *Next step after Figure -D.2a*

Figure D.2: *The conversion of the copying mechanism in an FSBM to RCE. P represents the plain, non-H, non-G states*

APPENDIX E

E.1 One iteration of Pseudo-German

1. Step 1 (E-Step). Assume the URs for each morpheme are equiprobable, and the weights for the constraints are all 1. The learner calculates the expected "contribution" of each underlying representation for the surface observation.

	*FinalVoicedObs	Ident[Voice]	*VTV	$\mathcal{H}$	$P(s \mid u)$
	$w = 1$	$w = 1$	$w = 1$		
$P(/\text{bed}/ \mid \text{Cat}_{\text{stem}}) = 0.5$					
a. [bed]	1			1	0.5
b. [bet]		1		1	0.5
$P(/\text{bet}/ \mid \text{Cat}_{\text{stem}}) = 0.5$					
a. [bed]	1	1		2	0.119
b. [bet]				0	0.881
$P(/\text{bed-a}/ \mid \text{Cat}_{\text{pl}}) = 0.5$					
a. [bed-a]				0	0.881
b. [bet-a]		1	1	2	0.119
$P(/\text{bet-a}/ \mid \text{Cat}_{\text{pl}}) = 0.5$					
a. [bed-a]		1		1	0.5
b. [bet-a]			1	1	0.5

$$
\begin{aligned}
P([\text{bet}] \mid \textsc{Cat}_{\text{stem}}) &= P([\text{bet}], /\text{bed}/ \mid \textsc{Cat}_{\text{stem}}) + P([\text{bet}], /\text{bet}/ \mid \textsc{Cat}_{\text{stem}}) \\
&= P([\text{bet}] \mid /\text{bed}/) P(/\text{bed}/ \mid \textsc{Cat}_{\text{stem}}) \\
&+ P([\text{bet}] \mid /\text{bet}/) P(/\text{bet}/ \mid \textsc{Cat}_{\text{stem}}) \\
&= 0.5 * 0.5 + 0.881 * 0.5 \\
&= 0.6905 \\
P(/\text{bed}/ \mid [\text{bet}], \textsc{Cat}_{\text{stem}}) &= \frac{P([\text{bet}], /\text{bed}/ \mid \textsc{Cat}_{\text{stem}})}{P([\text{bet}] \mid \textsc{Cat}_{\text{stem}})} \\
&= \frac{0.5 * 0.5}{0.6905} \\
&= 0.362 \\
P(/\text{bet}/ \mid [\text{bet}], \textsc{Cat}_{\text{stem}}) &= \frac{P([\text{bet}], /\text{bet}/ \mid \textsc{Cat}_{\text{stem}})}{P([\text{bet}] \mid \textsc{Cat}_{\text{stem}})} \\
&= \frac{0.881 * 0.5}{0.6905} \\
&= 0.638
\end{aligned}
$$

$$
\begin{aligned}
P([\text{bed-a}] \mid \textsc{Cat-pl}) &= P([\text{bed-a}], /\text{bed-a}/ \mid \textsc{Cat-pl}) + P([\text{bed-a}], /\text{bed-a}/ \mid \textsc{Cat-pl}) \\
&= P([\text{bed-a}] \mid /\text{bed-a}/) P(/\text{bed-a}/ \mid \textsc{Cat-pl}) \\
&+ P([\text{bed-a}] \mid /\text{bet-a}/) P(/\text{bet-a}/ \mid \textsc{Cat-pl}) \\
&= 0.5 * 0.881 + 0.5 * 0.5 \\
&= 0.6905 \\
P(/\text{bed-a}/ \mid [\text{bed-a}], \textsc{Cat-pl}) &= \frac{P([\text{bed-a}], /\text{bed-a}/ \mid \textsc{Cat-pl})}{P([\text{bed-a}] \mid \textsc{Cat-pl})} \\
&= \frac{0.5 * 0.881}{0.6905} \\
&= 0.638 \\
P(/\text{bet-a}/ \mid [\text{bed-a}], \textsc{Cat-pl}) &= \frac{P([\text{bed-a}], /\text{bet-a}/ \mid \textsc{Cat-pl})}{P([\text{bed-a}] \mid \textsc{Cat-pl})} \\
&= \frac{0.5 * 0.5}{0.6905} \\
&= 0.362
\end{aligned}
$$

2. Step 2: Fill in the "incomplete" data with expected values, and perform regular maximization as in MaxEnt models for better weights estimation based on $E(u, s, \omega)$.

	*FinalVoicedObs $w = 1$	Ident[Voice] $w = 1$	*VTV $w = 1$	$\mathcal{H}$	$P(s \mid u)$	$E(u_\omega, s, \omega)$
$P($/bed/ $\mid$ Cat$_{\text{stem}}) = 0.5$						
a. [bed]	1			1	0.5	0
b. [bet]		1		1	0.5	0.362
$P($/bet/ $\mid$ Cat$_{\text{stem}}) = 0.5$						
a. [bed]	1	1		2	0.119	0
b. [bet]				0	0.881	0.638
$P($/bed-a/ $\mid$ Cat$_{\text{pl}}) = 0.5$						
a. [bed-a]				0	0.881	0.638
b. [bet-a]		1	1	2	0.119	0
$P($/bet-a/ $\mid$ Cat$_{\text{pl}}) = 0.5$						
a. [bed-a]		1		1	0.5	0.362
b. [bet-a]			1	1	0.5	0
$P($/panat/ $\mid$ Dog$_{\text{stem}}) =1$						
a. [panad]	1	1		2	0.119	0
b. [panat]				0	0.881	1
$P($/panat-a/ $\mid$ Dog$_{\text{pl}}) =1$						
a. [panad-a]		1		1	0.5	0
b. [panat-a]			1	1	0.5	1

The newly updated weights

*FinalVoicedObs	Ident[Voice]	*VTV
$w = 10.55$	$w = 5.93$	$w = 4.9$

3. Step 3 (E-step): Based on the newly updated weights, and the current UR probabilities, the learner re-calculates the expected "contribution" of each underlying representation for the surface observation.

	*FinalVoicedObs	Ident[Voice]	*VTV	$\mathcal{H}$	$P(s \mid u)$	$E(u, s, \omega)$
	$w = 10.55$	$w = 5.93$	$w = 4.9$			
$P(\text{/bed/} \mid \textsc{Cat}_{\text{stem}}) = 0.5$						
a. [bed]	1			10.55	0.01	0
b. [bet]		1		5.93	0.990	$\boxed{0.497}$
$P(\text{/bet/} \mid \textsc{Cat}_{\text{stem}}) = 0.5$						
a. [bed]	1	1		16.48	0	0
b. [bet]				0	1	$\boxed{0.503}$
$P(\text{/bed-a/} \mid \textsc{Cat}_{\text{pl}}) = 0.5$						
a. [bed-a]				0	1	$\boxed{0.793}$
b. [bet-a]		1	1	10.83	0	0
$P(\text{/bet-a/} \mid \textsc{Cat}_{\text{pl}}) = 0.5$						
a. [bed-a]		1		5.93	0.26	$\boxed{0.207}$
b. [bet-a]			1	4.9	0.74	0
$P(\text{/mot/} \mid \textsc{Dog}_{\text{stem}}) = 1$						
a. [mod]	1	1		16.48	0	0
b. [mot]				0	1	1
$P(\text{/mot-a/} \mid \textsc{Dog}_{\text{pl}}) = 1$						
a. [mod-a]		1		5.93	0.26	0
b. [mot-a]			1	4.9	0.74	1

$$
\begin{aligned}
P([\mathrm{bet}] \,|\, \textsc{Cat}_{\mathrm{stem}}) &= P([\mathrm{bet}], /\mathrm{bed}/ \,|\, \textsc{Cat}_{\mathrm{stem}}) \\
&+ P([\mathrm{bet}], /\mathrm{bet}/ \,|\, \textsc{Cat}_{\mathrm{stem}}) \\
&= P([\mathrm{bet}] \,|\, /\mathrm{bed}/) P(/\mathrm{bed}/ \,|\, \textsc{Cat}_{\mathrm{stem}}) \\
&+ P([\mathrm{bet}] \,|\, /\mathrm{bet}/) P(/\mathrm{bet}/ \,|\, \textsc{Cat}_{\mathrm{stem}}) \\
&= 0.990 * 0.5 + 1 * 0.5 \\
&= 0.995
\end{aligned}
$$

$$
\begin{aligned}
P(/\mathrm{bed}/ \,|\, [\mathrm{bet}], \textsc{Cat}_{\mathrm{stem}}) &= \frac{P([\mathrm{bet}], /\mathrm{bed}/ \,|\, \textsc{Cat}_{\mathrm{stem}})}{P([\mathrm{bet}] \,|\, \textsc{Cat}_{\mathrm{stem}})} \\
&= \frac{0.990 * 0.5}{0.995} \\
&= 0.497
\end{aligned}
$$

$$
\begin{aligned}
P(/\mathrm{bet}/ \,|\, [\mathrm{bet}], \textsc{Cat}_{\mathrm{stem}}) &= \frac{P([\mathrm{bet}], /\mathrm{bet}/ \,|\, \textsc{Cat}_{\mathrm{stem}})}{P([\mathrm{bet}] \,|\, \textsc{Cat}_{\mathrm{stem}})} \\
&= \frac{1 * 0.5}{0.995} \\
&= 0.503
\end{aligned}
$$

$$\begin{aligned}
P([\text{bed-a}] \mid \textsc{Cat-pl}) &= P([\text{bed-a}], /\text{bed-a}/ \mid \textsc{Cat-pl}) \\
&+ P([\text{bed-a}], /\text{bet-a}/ \mid \textsc{Cat-pl}) \\
&= P([\text{bed-a}] \mid /\text{bed-a}/) P(/\text{bed-a}/ \mid \textsc{Cat-pl}) \\
&+ P([\text{bed-a}] \mid /\text{bet-a}/) P(/\text{bet-a}/ \mid \textsc{Cat-pl}) \\
&= 1 * 0.5 + 0.26 * 0.5 \\
&= 0.63
\end{aligned}$$

$$\begin{aligned}
P(/\text{bed-a}/ \mid [\text{bed-a}], \textsc{Cat-pl}) &= \frac{P([\text{bed-a}], /\text{bed-a}/ \mid \textsc{Cat-pl})}{P([\text{bed-a}] \mid \textsc{Cat-pl})} \\
&= \frac{1 * 0.5}{0.63} \\
&= 0.793
\end{aligned}$$

$$\begin{aligned}
P(/\text{bet-a}/ \mid [\text{bed-a}], \textsc{Cat-pl}) &= \frac{P([\text{bed-a}], /\text{bet-a}/ \mid \textsc{Cat-pl})}{P([\text{bed-a}] \mid \textsc{Cat-pl})} \\
&= \frac{0.26 * 0.5}{0.63} \\
&= 0.207
\end{aligned}$$

4. Step 4 (M-step): based on the current expected values of each underlying representation, get a better parameter estimation for the UR parameters based on relative frequency estimation.

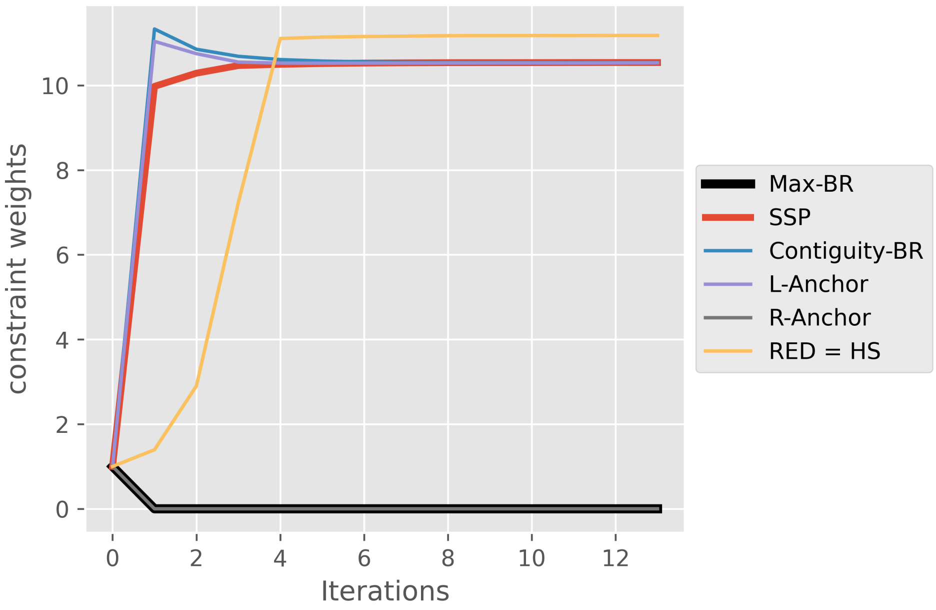
	$E(u_\omega, s, \omega)$
$P(\text{/bed/} \mid \textsc{Cat}_{\text{stem}}) = 0.5$	
a. [bed]	0
b. [bet]	0.497
$P(\text{/bet/} \mid \textsc{Cat}_{\text{stem}}) = 0.5$	
a. [bed]	0
b. [bet]	0.503
$P(\text{/bed-a/} \mid \textsc{Cat}_{\text{pl}}) = 0.5$	
a. [bed-a]	0.793
b. [bet-a]	0
$P(\text{/bet-a/} \mid \textsc{Cat}_{\text{pl}}) = 0.5$	
a. [bed-a]	0.207
b. [bet-a]	0

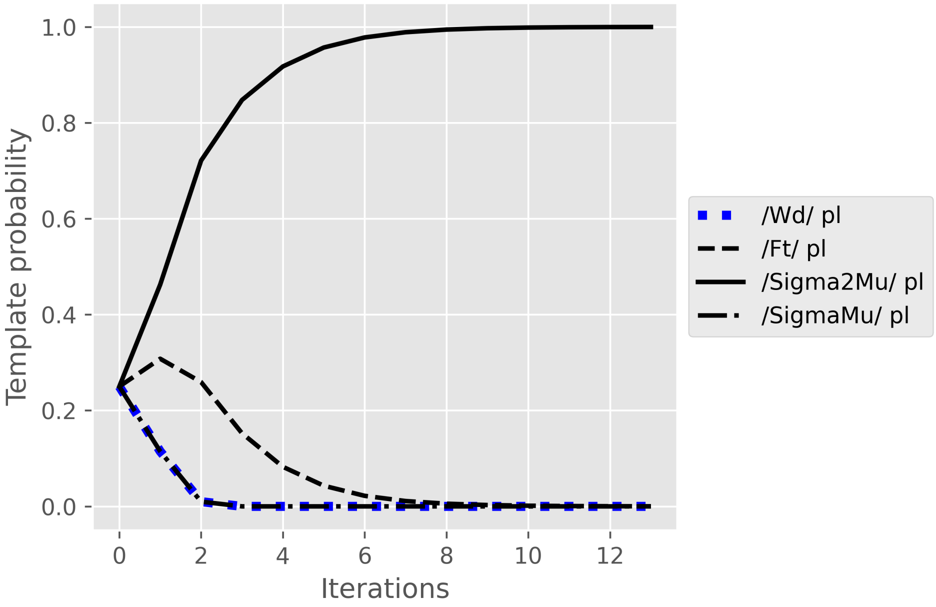
For $\textsc{Cat}$:

$$\begin{aligned}
\theta_{(\textsc{Cat},/\text{bed}/)} &= P(/\text{bed}/\,|\,\textsc{Cat}) \\
&= \frac{E(/\text{bed}/,[\text{bet}],\textsc{Cat}_{\text{stem}}) + E(/\text{bed-a}/,[\text{bed-a}],\textsc{Cat}_{\text{pl}})}{2} \\
&= \frac{(0.497 + 0.793)}{2} \\
&= 0.645 \\
\theta_{(\textsc{Cat},/\text{bet}/)} &= P(/\text{bet}/|\,\textsc{Cat}) \\
&= \frac{E(/\text{bet}/,\,[\text{bet}],\,\textsc{Cat}_{\text{stem}}) + E([/\text{bet-a}/,\,[\text{bed-a}],\textsc{Cat}_{\text{pl}})}{2} \\
&= \frac{(0.503 + 0.207)}{2} \\
&= 0.355
\end{aligned}$$

Note that it is the plural form which contributes more in informing the learner that the underlying representation (UR) for $\textsc{Cat}$ is /bed/ rather than /bet/.

E.2 The learning trajectory of Expt.1a





Bibliography

Natasha Abner. What you see is what you get.get: Surface transparency and ambiguity of nominalizing reduplication in American Sign Language. *Syntax*, 20(4):317–352, 2017.

Adam Albright and Bruce Hayes. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2):119–161, 2003.

Daniel M Albro. Evaluation, implementation, and extension of Primitive Optimality Theory. Master's thesis, University of California, Los Angeles, 1998.

Daniel M. Albro. Taking Primitive Optimality Theory beyond the Finite State. In *Proceedings of the Fifth Workshop of the ACL Special Interest Group in Computational Phonology*, pages 57–67, Centre Universitaire, Luxembourg, August 2000. International Committee on Computational Linguistics.

Daniel Matthew Albro. *Studies in computational Optimality Theory, with special reference to the phonological system of Malagasy.* PhD thesis, University of California, Los Angeles, 2005.

John Alderete, Jill Beckman, Laura Benua, Amalia Gnanadesikan, John McCarthy, and Suzanne Urbanczyk. Reduplication with Fixed Segmentism. *Linguistic Inquiry*, 30(3): 327–364, 1999.

Raquel G Alhama and Willem Zuidema. A review of computational models of basic rule learning: The neural-symbolic debate and beyond. *Psychonomic bulletin & review*, 26: 1174–1194, 2019.

Rajeev Alur and Pavol Černý. Expressiveness of streaming string transducers. In Kamal Lodaya and Meena Mahajan, editors, *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2010)*, volume 8 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 1–12, Dagstuhl, Germany, 2010. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Diana Apoussidou. On-line learning of underlying forms. *Rutgers Optimality Archive*, 835, 2006.

Diana Apoussidou. *The learnability of metrical phonology*, volume 148. LOT Utrecht, Neth., 2007.

Mark Aronoff. *Word formation in generative grammar.* Number 1. Linguistic Inquiry Monographs, 1976.

Peter Austin. *A grammar of Diyari, South Australia.* Cambridge University Press Cambridge, 1981.

R Harald Baayen, Richard Piepenbrock, and Hedderik van Rijn. The cel database. *Philadelphia, PA: Linguistic Data Consortium.[Release 2 (CD-ROM)*, 1995.

Bruce Bagemihl. The crossing constraint and 'backwards languages'. *Natural language & linguistic Theory*, 7(4):481–549, 1989.

Shraddha Barke, Rose Kunkel, Nadia Polikarpova, Eric Meinhardt, Eric Baković, and Leon Bergen. Constraint-based learning of phonological processes. In *Proceedings of the 2019 conference on EMNLP and the 9th IJCNLP*, pages 6176–6186, 2019.

Félix Baschenis, Olivier Gauwin, Anca Muscholl, and Gabriele Puppis. Untwisting two-way transducers in elementary time. In *2017 32nd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–12, 2017.

Outi Bat-El. Consonant identity and consonant copy: The segmental and prosodic structure of Hebrew reduplication. *Linguistic Inquiry*, 37(2):179–210, 2006.

Michael Becker, Nihan Ketrez, and Andrew Nevins. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language*, pages 84–125, 2011.

Michael Becker, Andrew Nevins, and Jonathan Levine. Asymmetries in generalizing alternations to and from initial syllables. *Language*, pages 231–268, 2012.

Kenneth R. Beesley and Lauri Karttunen. Finite-state non-concatenative morphotactics. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 191–198, Hong Kong, October 2000. Association for Computational Linguistics.

Gašper Beguš. Identity-based patterns in deep convolutional networks: Generative adversarial phonology and reduplication. *Transactions of the Association for Computational Linguistics*, 9:1180–1196, 2021.

Gašper Beguš and Alan Zhou. Interpreting intermediate convolutional layers of generative cnns trained on waveforms. *IEEE/ACM transactions on audio, speech, and language processing*, 30:3214–3229, 2022.

Caleb Belth. *Towards an Algorithmic Account of Phonological Rules and Representations*. PhD thesis, University of Michigan, 2023a.

Caleb Belth. Towards a learning-based account of underlying forms: A case study in Turkish. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 332–342, 2023b.

Ryan Bennett. Recursive prosodic words in Kaqchikel (Mayan). *Glossa a journal of general linguistics*, 3(1), 2018.

Ryan Thomas Bennett. *Foot-conditioned phonotactics and prosodic constituency*. PhD thesis, UC Santa Cruz, 2012.

Iris Berent. The phonological mind. *Trends in cognitive sciences*, 17(7):319–327, 2013.

Iris Berent, Colin Wilson, Gary F Marcus, and Douglas K Bemis. On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic inquiry*, 43(1):97–119, 2012.

Iris Berent, Outi Bat-El, Diane Brentari, Amanda Dupuis, and Vered Vaknin-Nusbaum. The double identity of linguistic doubling. *Proceedings of the National Academy of Sciences*, 113(48):13702–13707, 2016.

Iris Berent, Outi Bat-El, and Vered Vaknin-Nusbaum. The double identity of doubling: Evidence for the phonology–morphology split. *Cognition*, 161:117–128, 2017.

Jean Berko. The child's learning of English morphology. *Word*, 14(2-3):150–177, 1958.

Karen M. Booker. *Comparative Muskogean: Aspects of Proto-Muskogean Verb Morphology.* PhD thesis, University of Kansas, 1979.

Sami Boudelaa and William D Marslen-Wilson. Abstract morphemes and lexical representation: The CV-Skeleton in Arabic. *Cognition*, 92(3):271–303, 2004.

Mario Brdar. Adjective reduplication and diagrammatic iconicity. 2013.

Ellen Broselow and John McCarthy. A theory of internal reduplication. 1983.

Ellen I. Broselow. Salish double reduplications: Subjacency in morphology. *Natural Language & Linguistic Theory*, 1:317–346, 1983.

Gabriela Caballero. "Templatic backcopying" in Guarijio abbreviated reduplication. *Morphology*, 16:273–289, 2006.

Shira Calamaro and Gaja Jarosz. Learning general phonological rules from distributional information: A computational model. *Cognitive science*, 39(3):647–666, 2015.

Cezar Câmpeanu, Kai Salomaa, and Sheng Yu. Regex and extended regex. In Jean-Marc Champarnaud and Denis Maurel, editors, *Implementation and Application of Automata, 7th International Conference, CIAA 2002, Tours, France, July 3-5, 2002, Revised Papers*, volume 2608 of *Lecture Notes in Computer Science*, pages 77–84. Springer, 2002.

Cezar Câmpeanu, Kai Salomaa, and Sheng Yu. A formal study of practical regular expressions. *International Journal of Foundations of Computer Science*, 14(06):1007–1018, 2003.

Benjamin Carle and Paliath Narendran. On extended regular expressions. In *Language and Automata Theory and Applications: Third International Conference, LATA 2009, Tarragona, Spain, April 2-8, 2009. Proceedings 3*, pages 279–289. Springer, 2009.

Jill Louise Carrier. *The interaction of morphological and phonological rules in Tagalog: a study in the relationship between rule components in grammar.* PhD thesis, Massachusetts Institute of Technology, 1979.

Jane Chandlee. *Strictly local phonological processes.* PhD thesis, University of Delaware, 2014.

Jane Chandlee. Computational locality in morphological maps. *Morphology*, 27:599–641, 2017.

Jane Chandlee and Jeffrey Heinz. Bounded copying is subsequential: Implications for metathesis and reduplication. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 42–51, Montréal, Canada, June 2012. Association for Computational Linguistics.

Jane Chandlee, Rémi Eyraud, and Jeffrey Heinz. Learning strictly local subsequential functions. *Transactions of the Association for Computational Linguistics*, 2:491–504, 2014.

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

Alexander Clark and Ryo Yoshinaka. Distributional learning of parallel multiple context-free grammars. *Mach. Learn.*, 96(1–2):5–31, July 2014. ISSN 0885-6125.

Alexander Clark, Makoto Kanazawa, Gregory M Kobele, and Ryo Yoshinaka. Distributional learning of some nonlinear tree grammars. *Fundamenta Informaticae*, 146(4):339–377, 2016.

Yael Cohen-Sygal and Shuly Wintner. Finite-state registered automata for non-concatenative morphology. *Computational Linguistics*, 32(1):49–82, 2006.

Ryan Cotterell, Nanyun Peng, and Jason Eisner. Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association for Computational Linguistics*, 3:433–447, 2015.

Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4 (1):1–34, 2007.

Berthold Crysmann. Reduplication in a computational hpsg of hausa. *Morphology*, 27(4): 527–561, 2017.

Jennifer Culbertson. Artificial language learning. In *The Oxford Handbook of Experimental Syntax*. Oxford University Press, 03 2023. ISBN 9780198797722.

Jennifer Culbertson and David Adger. Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, 111(16):5842–5847, 2014.

Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. Learning biases predict a word order universal. *Cognition*, 122(3):306–329, 2012.

Christopher Culy. The complexity of the vocabulary of Bambara. *Linguistics and philosophy*, 8(3):345–351, 1985.

Robertx Daland, Bruce Hayes, James White, Marc Garellek, Andrea Davis, and Ingrid Norrmann. Explaining sonority projection effects. *Phonology*, 28(2):197–234, 2011.

Lisa Garnand Dawdy-Hesterberg and Janet Breckenridge Pierrehumbert. Learnability and generalisation of Arabic broken plural nouns. *Language, cognition and neuroscience*, 29 (10):1268–1282, 2014.

Colin Dawson and LouAnn Gerken. From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition*, 111 (3):378–382, 2009.

Aniello De Santo and Thomas Graf. Structure sensitive tier projection: Applications and formal properties. In Raffaella Bernardi, Greg Kobele, and Sylvain Pogodalla, editors, *Formal Grammar*, pages 35–50. Springer Berlin Heidelberg, 2019.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. *IEEE Transactions on pattern analysis and machine intelligence*, 19(4):380–393, 1997.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Robert M. W. Dixon. *The Dyirbal Language of North Queensland*, volume 9 of *Cambridge Studies in Linguistics*. Cambridge University Press, Cambridge, 1972.

Chuong B Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897–899, 2008.

Hossep Dolatian and Jeffrey Heinz. Learning reduplication with 2-way finite-state transducers. In Olgierd Unold, Witold Dyrka, and Wojciech Wieczorek, editors, *Proceedings of the 14th International Conference on Grammatical Inference*, volume 93 of *Proceedings of Machine Learning Research*, pages 67–80. PMLR, 05–07 Sep 2018a.

Hossep Dolatian and Jeffrey Heinz. Modeling reduplication with 2-way finite-state transducers. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 66–77, Brussels, Belgium, October 2018b. Association for Computational Linguistics.

Hossep Dolatian and Jeffrey Heinz. Redtyp: A database of reduplication with computational models. In *Proceedings of the Society for Computation in Linguistics*, volume 2, 2019. Article 3.

Hossep Dolatian and Jeffrey Heinz. Computing and classifying reduplication with 2-way finite-state transducers. *Journal of Language Modelling*, 8(1):179–250, 2020.

Laura J Downing. Morphological and prosodic constraints on Kinande verbal reduplication. *Phonology*, 17(1):1–38, 2000.

Laura J Downing and Sharon Inkelas. What is reduplication? Typology and analysis part 2/2: The analysis of reduplication. *Language and Linguistics Compass*, 9(12):516–528, 2015.

B Elan Dresher and Aditi Lahiri. The Germanic foot: Metrical coherence in Old English. *Linguistic Inquiry*, 22(2):251–286, 1991.

Markus Dreyer and Jason Eisner. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of the Conference on EMNLP*, pages 616–627, Edinburgh, 2011. Supplementary material (9 pages) also available.

Sarah Eisenstat. Learning underlying forms with MaxEnt. *Master's thesis, Brown University*, 2009.

Jason Eisner. Efficient generation in primitive Optimality Theory. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Madrid, Spain, 1997. Association for Computational Linguistics.

Jason Eisner. Expectation semirings: Flexible EM for finite-state transducers. In Gertjan van Noord, editor, *Proceedings of the ESSLLI Workshop on Finite-State Methods in Natural Language Processing (FSMNLP)*, Helsinki, August 2001. URL `http://cs.jhu.edu/~jason/papers/#eisner-2001-fsmnlp`. Extended abstract (5 pages).

Kevin Ellis, Armando Solar-Lezama, and Josh Tenenbaum. Unsupervised learning by program synthesis. *Advances in neural information processing systems*, 28, 2015.

Kevin Ellis, Adam Albright, Armando Solar-Lezama, Joshua B Tenenbaum, and Timothy J O'Donnell. Synthesizing theories of human language with bayesian program induction. *Nature communications*, 13(1):5024, 2022.

T. Mark Ellison. Phonological derivation in optimality theory. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94, page 1007–1013, USA, 1994. Association for Computational Linguistics.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Ansgar D Endress, Ghislaine Dehaene-Lambertz, and Jacques Mehler. Perceptual constraints and the learnability of simple grammars. *Cognition*, 105(3):577–614, 2007.

Joost Engelfriet and Hendrik Jan Hoogeboom. Mso definable string transductions and two-way finite state transducers, 1999.

Brock Ferguson and Casey Lew-Williams. Communicative signals support abstract rule learning by 7-month-old infants. *Scientific reports*, 6(1):1–7, 2016.

Sara Finley and William Badecker. Artificial language learning and feature-based generalization. *Journal of memory and language*, 61(3):423–437, 2009.

Sara Finley and Morten Christiansen. Multimodal transfer of repetition patterns in artificial grammar learning. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 33, 2011.

John Frampton. *Distributed reduplication*. MIT Press, 2009.

Michael C Frank and Joshua B Tenenbaum. Three ideal observer models for rule learning in simple languages. *Cognition*, 120(3):360–371, 2011.

Michael C Frank, Jonathan A Slemmer, Gary F Marcus, and Scott P Johnson. Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental science*, 12(4):504–509, 2009.

Robert Frank and Giorgio Satta. Optimality theory and the generative complexity of constraint violability. *Computational linguistics*, 24(2):307–315, 1998.

Dominik D Freydenberger and Markus L Schmid. Deterministic regular expressions with back-references. *Journal of Computer and System Sciences*, 105:1–39, 2019.

Diamandis Gafos. A-templatic reduplication. *Linguistic Inquiry*, pages 515–527, 1998.

Gillian Gallagher. Learning the identity effect as an artificial language: bias and generalisation. *Phonology*, 30(2):253–295, 2013.

Pedro Garcia, Enrique Vidal, and José Oncina. Learning locally testable languages in the strict sense. In *Proceedings of the Workshop on Algorithmic Learning Theory*, pages 325–338, 1990.

Michael Gasser. Acquiring receptive morphology: a connectionist model. *arXiv preprint cmp-lg/9405027*, 1994.

Gerald Gazdar and Geoffrey K Pullum. Computationally relevant properties of natural languages and their grammars. *New generation computing*, 3(3):273–306, 1985.

LouAnn Gerken. Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, 98(3):B67–B74, 2006.

LouAnn Gerken. Infants use rational decision criteria for choosing among models of their input. *Cognition*, 115(2):362–366, 2010.

Judit Gervain, Iris Berent, and Janet F Werker. Binding at birth: The newborn brain detects identity relations and sequential position in speech. *Journal of Cognitive Neuroscience*, 24 (3):564–574, 2012.

Jila Ghomeshi, Ray Jackendoff, Nicole Rosen, and Kevin Russell. Contrastive focus reduplication in English (the salad-salad paper). *Natural Language & Linguistic theory*, 22: 307–357, 2004.

David Gil. How to speak backwards in Tagalog. In *Pan-Asiatic Linguistics, Proceedings of the Fourth International Symposium on Language and Linguistics, January 8-10, 1996, Institute of Language and Culture for Rural Development, Mahidol University at Salaya*, volume 1, pages 297–306, 1996.

E Mark Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.

Sharon Goldwater and Mark Johnson. Learning ot constraint rankings using a maximum entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory*, volume 113, page 122, 2003.

Sharon Goldwater and Mark Johnson. Representational bias in unsupervised learning of syllable structure. In Ido Dagan and Daniel Gildea, editors, *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 112–119, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `https://aclanthology.org/W05-0615`.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Maria Gouskova. The reduplicative template in Tonkawa. *Phonology*, 24(3):367–396, 2007.

David W Gow Jr, Enes Avcu, Adriana Schoenhaut, David O Sorensen, and Seppo P Ahlfors. Abstract representations in temporal cortex support generative linguistic processing. *Language, Cognition and Neuroscience*, 38(6):765–778, 2023.

Thomas Graf. *Local and Transderivational Constraints in Syntax and Semantics*. PhD thesis, University of California, Los Angeles, 2013.

Coleman Haley and Colin Wilson. Deep neural networks easily learn unnatural infixation and reduplication patterns. *Proceedings of the Society for Computation in Linguistics*, 4 (1):427–433, 2021.

Jason D Haugen and Cathy Hicks Kennard. Base-dependence in reduplication. *Morphology*, 21:1–29, 2011.

Jason D Haugen, Adam Ussishkin, and Colin Reimer Dawson. Learning a typologically unusual reduplication pattern: An artificial language learning study of base-dependent reduplication. *Morphology*, 32(3):299–315, 2022.

Bruce Hayes. *Metrical stress theory: Principles and case studies.* University of Chicago Press, 1995.

Bruce Hayes. *Introductory phonology*, volume 7. John Wiley & Sons, 2009.

Bruce Hayes and May Abad. Reduplication and syllabification in ilokano. *Lingua*, 77(3-4): 331–374, 1989.

Bruce Hayes and Colin Wilson. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440, 2008.

Bruce Hayes, Péter Siptár, Kie Zuraw, and Zsuzsa Londe. Natural and unnatural constraints in Hungarian vowel harmony. *Language*, pages 822–863, 2009.

Phyllis M. Healey. *An Agta Grammar.* Bureau of Printing, Manila, 1960.

Jeffrey Heinz. *The Inductive Learning of Phonotactic Patterns.* PhD thesis, University of California, Los Angeles, 2007.

Jeffrey Heinz. String extension learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 897–906, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Jeffrey Heinz. The computational nature of phonological generalizations. In Larry Hyman and Frans Plank, editors, *Phonological Typology*, Phonetics and Phonology, chapter 5, pages 126–195. De Gruyter Mouton, 2018.

Jeffrey Heinz and William Idsardi. What complexity differences reveal about domains in language. *Topics in cognitive science*, 5(1):111–131, 2013.

Jeffrey Heinz, Chetan Rawal, and Herbert G Tanner. Tier-based strictly local constraints for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human language technologies*, pages 58–64, 2011.

Sean Quillan Hendricks. *Reduplication without template constraints: A study in bare-consonant reduplication.* The University of Arizona, 1999.

John E Hopcroft and Jeffrey D Ullman. Introduction to automata theory, languages, and computation. *Addison-Welsey, NY*, 1979.

Wenyue Hua and Adam Jardine. Learning input strictly local functions from their composition. In *International Conference on Grammatical Inference*, pages 47–65. PMLR, 2021.

Mans Hulden. *Finite-state Machine Construction Methods and Algorithms for Phonology and Morphology*. PhD thesis, University of Arizona, Tucson, USA, 2009.

Bernhard Hurch and Veronika Mattes. Typology of reduplication: The graz database. *The use of databases in cross-linguistic studies*, pages 301–328, 2009.

Riny Huybregts. The weak inadequacy of context-free phrase structure grammars. *Van periferie naar kern*, pages 81–99, 1984.

Sharon Inkelas. The dual theory of reduplication. 46(2):351–401, 2008.

Sharon Inkelas. *The interplay of morphology and phonology*, volume 8. Oxford University Press, 2014.

Sharon Inkelas and Laura J Downing. What is reduplication? Typology and analysis part 1/2: The typology of reduplication. *Language and linguistics compass*, 9(12):502–515, 2015.

Sharon Inkelas and Cheryl Zoll. *Reduplication: Doubling in morphology*, volume 106. Cambridge University Press, 2005.

Gerhard Jäger and James Rogers. Formal language theory: refining the chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1598):1956–1970, 2012.

Adam Jardine and Jeffrey Heinz. Learning tier-based strictly 2-local languages. *Transactions of the Association for Computational Linguistics*, 4:87–98, 2016.

Gaja Jarosz. *Rich lexicons and restrictive grammars: maximum likelihood learning in Optimality Theory.* PhD thesis, Johns Hopkins University, 2006.

Gaja Jarosz. Learning with hidden structure in Optimality Theory and Harmonic Grammar: Beyond robust interpretive parsing. *Phonology*, 30(1):27–71, 2013.

Gaja Jarosz. Expectation driven learning of phonology. *Ms. University of Massachusetts Amherst*, 2015.

Gaja Jarosz. Computational modeling of phonological learning. *Annual Review of Linguistics*, 5:67–90, 2019.

C Douglas Johnson. *Formal aspects of phonological description.* The Hague: Mouton, 1972.

Mark Johnson, Joe Pater, Robert Staubs, and Emmanuel Dupoux. Sign constraints on feature weights improve a joint model of word segmentation and phonology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 303–313, 2015.

Aravind K. Joshi. *Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions?*, page 206–250. Studies in Natural Language Processing. Cambridge University Press, 1985.

Aravind K Joshi, K Vijay Shanker, and David Weir. The convergence of mildly context-sensitive grammar formalisms. *Technical Reports (CIS)*, 1990.

Laura Kallmeyer. *Parsing Beyond Context-Free Grammars.* Cognitive Technologies. Springer, 2010.

Ronald M. Kaplan and Martin Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378, September 1994. ISSN 0891-2017.

Michael H Kelly. Word onset patterns and lexical stress in English. *Journal of Memory and Language*, 50(3):231–244, 2004.

Michael Kenstowicz and Charles Kisseberth. The problem of the abstractness of underlying representations. In Michael Kenstowicz and Charles Kisseberth, editors, *Topics in Phonological Theory*, pages 1–62. Academic Press, 1977.

Salam Khalifa, Sarah Payne, Jordan Kodner, Ellen Broselow, and Owen Rambow. A cautious generalization goes a long way: Learning morphophonological rules. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1793–1805, 2023.

Paul Kiparsky. Reduplication in Stratal OT. *Reality exploration and discovery: pattern interaction in language & life*, pages 125–142, 2010.

Gregory M. Kobele. *Generating Copies: An investigation into structural identity in language and grammar.* PhD thesis, University of California, Los Angeles, 2006.

Jordan Kodner. Computational models of morphological learning. In *Oxford Research Encyclopedia of Linguistics*. 2022.

Grzegorz Kondrak. *Algorithms for language reconstruction.* PhD thesis, University of Toronto, 2002.

Ágnes Melinda Kovács and Jacques Mehler. Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences*, 106(16):6556–6560, 2009a.

Ágnes Melinda Kovács and Jacques Mehler. Flexible learning of multiple speech structures in bilingual infants. *science*, 325(5940):611–612, 2009b.

Jennifer Kuo. *Phonological markedness effects in paradigm reanalysis.* PhD thesis, University of California, Los Angeles, 2023a.

Jennifer Kuo. Evidence for prosodic correspondence in the vowel alternations of Tgdaya Seediq. *Phonological Data and Analysis*, 5(3):1–31, 2023b.

Martin Kutrib, Andreas Malcher, and Matthias Wendlandt. *Queue Automata: Foundations and Developments*, pages 385–431. Springer International Publishing, Cham, 2018.

William Labov. A study of non-standard English. 1969.

Brenden M Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*, 2019.

Andrew Lamont. Serial reduplication is empirically adequate and typologically restrictive. *Linguistic Inquiry*, 54(4):797–839, 2023.

Karim Lari and Steve J Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer speech & language*, 4(1):35–56, 1990.

Juliette Levin. *A metrical theory of syllabicity.* PhD thesis, Massachusetts Institute of Technology, 1985.

Frantisek Lichtenberk. A grammar of Manam. *Oceanic linguistics special publications*, (18): i–647, 1983.

Geoffery MacLahlan and David Peel. Finite mixture models. *John & Sons*, 2000.

Alexis Manaster-Ramer. Copying in natural languages, context-freeness, and queue grammars. In *Proceedings of the 24th Annual Meeting on Association for Computational Linguistics*, ACL '86, page 85–89, USA, 1986. Association for Computational Linguistics.

Alec Marantz. Re reduplication. *Linguistic inquiry*, 13(3):435–482, 1982.

G. F. Marcus, S. Vijayan, S. Bandi Rao, and P. M. Vishton. Rule learning by seven-month-old infants. *Science*, 283(5398):77–80, 1999.

Gary F Marcus. *The algebraic mind: Integrating connectionism and cognitive science.* MIT press, 2003.

Gary F. Marcus, Keith J. Fernandes, and Scott P. Johnson. Infant rule learning facilitated by speech. *Psychological Science*, 18(5):387–391, 2007.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W. H. Freeman and Company, 1982.

Alexander Martin and James White. Vowel harmony and disharmony are not equivalent in learning. *Linguistic Inquiry*, 52(1):227–239, 2021.

Elisa Mattiello. *Extra-grammatical morphology in English: Abbreviations, blends, reduplicatives, and related phenomena*, volume 82. Walter de Gruyter, 2013.

Connor Mayer and Travis Major. A challenge for tier-based strict locality from Uyghur backness harmony. In Annie Foret, Greg Kobele, and Sylvain Pogodalla, editors, *Formal Grammar 2018*, pages 62–83. Springer Berlin Heidelberg, 2018.

Connor Mayer, Adeline Tan, and Kie Zuraw. Introducing maxent. ot: an R package for Maximum Entropy constraint grammars. *Phonological Data & Analysis*.

John J McCarthy. OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 17(2): 207–263, 1986.

John J McCarthy. The gradual path to cluster simplification. *Phonology*, 25(2):271–319, 2008.

John J McCarthy and Abigail Cohn. Alignment and parallelism in Indonesian phonology. 1998.

John J McCarthy and Alan Prince. Faithfulness and identity in prosodic morphology. In Harry van der Hulst Rene Kager and Wim Zonneveld, editors, *The prosody-morphology interface*, pages 218–309. Cambridge University Press, 1999.

John J McCarthy and Alan Prince. Faithfulness and identity in prosodic morphology. *Optimality Theory in Phonology: A Reader*, pages 77–98, 2004.

John J. McCarthy and Alan S. Prince. Prosodic morphology. Ms., University of Massachusetts, Amherst, and Brandeis University, Waltham, Mass., 1986.

John J McCarthy and Alan S Prince. Foot and word in prosodic morphology: The arabic broken plural. *Natural Language & Linguistic Theory*, 8(2):209–283, 1990.

John J. McCarthy and Alan S. Prince. The emergence of the unmarked: Optimality in prosodic morphology. Amherst, MA, 1994. Graduate Linguistic Student Association, Dept. of Linguistics, University of Massachusetts.

John J. McCarthy and Alan S. Prince. Faithfulness and reduplicative identity. In *Papers in Optimality Theory*, Amherst, MA, 1995. GLSA (Graduate Linguistic Student Association), Dept. of Linguistics, University of Massachusetts.

John J McCarthy and Alan S Prince. Prosodic morphology. *The Handbook of Morphology*, pages 281–305, 2017.

John J McCarthy, Wendell Kimper, and Kevin Mullin. Reduplication in Harmonic Serialism. *Morphology*, 22(2):173–232, 2012.

John J McCarthy, Amel Khalfaoui, and Matthew Tucker. How to delete. *Perspectives on Arabic linguistics XXX*, pages 7–32, 2018.

Donka Minkova. Ablaut reduplication in English: The criss-crossing of prosody and verbal art. *English Language & Linguistics*, 6(1):133–169, 2002.

Edith A. Moravcsik. Reduplicative constructions. In Greenberg, J. H., and et al., editors, *Universals of Human Language. Volume 3: Word Structure*, pages 297–334. Stanford University Press, Stanford, 1978.

Elliott Moreton. Analytic bias and phonological typology. *Phonology*, 25(1):83–127, 2008.

Elliott Moreton and Katya Pertsova. Implicit and explicit processes in phonological concept learning. *Phonology*, pages 1–53, 2024.

Elliott Moreton, Brandon Prickett, Katya Pertsova, Joshua Fennell, Joe Pater, and Lisa Sanders. Learning reduplication, but not syllable reversal. In Ryan Bennett, Richard Bibbs, Mykel Loren Brinkerhoff, Stephanie Rich Max J. Kaplan, Amanda Rysling, Nicholas Van Handel, and Maya Wax Cavallaro, editors, *Supplemental Proceedings of the 2020 Annual Meeting on Phonology*, 2021.

David George Nash. *Topics in Warlpiri grammar.* PhD thesis, Massachusetts Institute of Technology, 1980.

Max Nelson. Segmentation and UR acquisition with UR constraints. *Proceedings of the Society for Computation in Linguistics*, 2(1):60–68, 2019.

Max Nelson, Hossep Dolatian, Jonathan Rawski, and Brandon Prickett. Probing RNN encoder-decoder generalization of subregular functions using reduplication. *Society for Computation in Linguistics*, 3(1), 2020.

Nicole Nelson. Wrong side reduplication is epiphenomenal: Evidence from yoruba. In Bernhard Hurch, editor, *Studies on Reduplication*, pages 135–160, Berlin, Boston, 2005. De Gruyter Mouton.

Andrew Nevins and Bert Vaux. Metalinguistic, shmetalinguistic: The phonology of shm-reduplication. In *Proceedings from the annual meeting of the Chicago Linguistic Society*, volume 39, pages 702–721. Chicago Linguistic Society, 2003.

Taishin Y. Nishida and Shigeko Seki. Grouped partial et0l systems and parallel multiple context-free grammars. *Theoretical Computer Science*, 246(1):131–150, 2000.

Richard Nivens. Reduplication in four dialects of West Tarangan. *Oceanic Linguistics*, pages 353–388, 1993.

Charlie O'Hara. How abstract is more abstract? Learning abstract underlying representations. *Phonology*, 34(2):325–345, 2017.

Mitsuhiko Ota and Barbora Skarabela. Reduplicated words are easier to learn. *Language Learning and Development*, 12(4):380–397, 2016.

Mitsuhiko Ota and Barbora Skarabela. Reduplication facilitates early word segmentation. *Journal of Child Language*, 45(1):204–218, 2018.

Mary Paster. *Phonological conditions on affixation.* PhD thesis, University of California, Berkeley, 2006.

Joe Pater and Elliott Moreton. Structurally biased phonology: Complexity in learning and typology. *Journal of the English and Foreign Languages University, Hyderabad*, 3(2):1–44, 2012.

Joe Pater, Robert Staubs, Karen Jesney, and Brian Cantwell Smith. Learning probabilities over underlying representations. In *Proceedings of the twelfth meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 62–71, 2012.

Sharon Peperkamp, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, 101(3):B31–B41, 2006.

Janet Pierrehumbert. Dissimilarity in the Arabic verbal roots. In *Proceedings of NELS*, volume 23, pages 367–381. University of Massachusetts Amherst, 1993.

Brandon Prickett, Aaron Traylor, and Joe Pater. Learning reduplication with a neural network that lacks explicit variables. 10(1):1–38, 2022.

Douglas Pulleyblank. Patterns of Reduplication in Yoruba. In *The Nature of the Word: Studies in Honor of Paul Kiparsky*. The MIT Press, 12 2008.

Hugh Rabagliati, Brock Ferguson, and Casey Lew-Williams. The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental science*, 22 (1):e12704, 2019.

Eric Raimy. The phonology and morphology of reduplication. In *The Phonology and Morphology of Reduplication*. De Gruyter Mouton, 2000.

Eric Raimy. Reduplication. In Oostendorp, Marc van, Colin J. Ewen, Elizabeth Hume, and Keren Rice, editor, *The Blackwell Companion to Phonology*. Wiley-Blackwell, 2011.

Ezer Rasin and Roni Katzir. On evaluation metrics in Optimality Theory. *Linguistic Inquiry*, 47(2):235–282, 2016.

Ezer Rasin and Roni Katzir. Learning abstract underlying representations from distributional evidence. In *Proceedings of NELS*, volume 48, pages 283–290, 2018.

Ezer Rasin and Roni Katzir. A conditional learnability argument for constraints on underlying representations. *Journal of Linguistics*, 56(4):745–773, 2020.

Ezer Rasin, Iddo Berger, Nur Lan, Itamar Shefi, and Roni Katzir. Approaching explanatory adequacy in phonology using Minimum Description Length. *Journal of Language Modelling*, 9(1):17–66, 2021.

Jonathan Rawski, Hossep Dolatian, Jeffrey Heinz, and Eric Raimy. Regular and polyregular theories of reduplication. *Glossa: a journal of general linguistics*, 8(1), 2023.

Tomas Riad. *The phonology of Swedish.* Phonology of the World's Langu, 2014.

Caitlin Richter. *Alternation-Sensitive Phoneme Learning: Implications for Children's Development and Language Change.* PhD thesis, University of Pennsylvania, 2021.

Jason Riggle. Nonlocal reduplication. In *Proceedings of the 34th Meeting of the North-East Linguistics Society (NELS 34)*, page 485–496, USA, 2004a. GLSA, University of Massachusetts.

Jason Riggle. Infixing reduplication in Pima and its theoretical consequences. *Natural Language & Linguistic Theory*, 24:857–891, 2006.

Jason Alan Riggle. *Generation, recognition, and learning in finite state Optimality Theory.* University of California, Los Angeles, 2004b.

Brian Roark and Richard Sproat. *Computational approaches to morphology and syntax*, volume 4. Oxford University Press, 2007.

Carl Rubino. *Reduplication: Form, function and distribution*, pages 11–30. De Gruyter Mouton, Berlin, Boston, 2005.

Carl Rubino. Reduplication. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

Jesse Saba Kirchner. *Minimal Reduplication*. PhD thesis, University of California, Santa Cruz, 2010.

Jesse Saba Kirchner. Minimal reduplication and reduplicative exponence. *Morphology*, 23 (2):227–243, 2013.

Jenny R Saffran, Seth D Pollak, Rebecca L Seibel, and Anna Shkolnik. Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, 105(3):669–680, 2007.

Edward Sapir and Harry Hoijer. *The Phonology and Morphology of the Navaho Language*. Number v. 50-51. University of California Press, 1967.

Walter Savitch. A formal model for context-free languages augmented with reduplication. *Computational Linguistics*, 15(4):250–261, 1989.

Walter J. Savitch. Why it might pay to assume that languages are infinite. *Annals of Mathematics and Artificial Intelligence*, 8:17–25, 1993.

Markus L Schmid. Characterising REGEX languages by regular languages equipped with factor-referencing. *Information and Computation*, 249:1–17, 2016.

Hiroyuki Seki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. On multiple context-free grammars. *Theoretical Computer Science*, 88(2):191–229, 1991.

Elisabeth Selkirk. Prosodic domains in phonology: Sanskrit revisited. 1980a.

Elisabeth Selkirk. The role of prosodic categories in English word stress. *Linguistic inquiry*, 11(3):563–605, 1980b.

Elisabeth Selkirk. On the major class features and syllable theory. *Language sound structure*, 1984.

Stuart M Shieber. Evidence against the context-freeness of natural language. In *Philosophy, Language, and Artificial Intelligence*, pages 79–89. Springer, 1985.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216, 1990.

Paul Smolensky and Alan Prince. Optimality theory: Constraint interaction in generative grammar. *Optimality Theory in phonology*, 3, 1993.

Philip Spaelti. *Dimensions of variation in multi-pattern reduplication.* PhD thesis, University of California, Santa Cruz, 1997.

Richard William Sproat. *Morphology and computation.* MIT press, 1992.

Edward P. Stabler. Varieties of crossing dependencies: Structure dependence and mild context sensitivity. *Cognitive Science*, 93(5):699–720, 2004.

Juliet Stanton. Learnability shapes typology: The case of the midpoint pathology. *Language*, pages 753–791, 2016.

Juliet Stanton and Sam Zukoff. Prosodic identity in copy epenthesis: Evidence for a correspondence-based approach. *Natural Language & Linguistic Theory*, 36:637–684, 2018.

Donca Steriade. *Greek prosodies and the nature of syllabification.* PhD thesis, Massachusetts Institute of Technology, 1982.

Donca Steriade. Reduplication and syllable transfer in Sanskrit and elsewhere. *Phonology*, 5(1):73–155, 1988.

Donca Steriade. Paradigm uniformity and the phonetics-phonology boundary. *Papers in laboratory phonology*, 5:313–334, 2000.

Jan-Olof Svantesson, Anna Tsendina, Anastasia Karlsson, and Vivan Franzén. *The phonology of Mongolian.* OUP Oxford, 2005.

Adeline Tan. Concurrent hidden structure & grammar learning. *Proceedings of the Society for Computation in Linguistics*, 5(1):55–64, 2022.

Bruce Tesar. *Output-driven phonology: Theory and learning.* Number 139. Cambridge University Press, 2014.

Bruce Tesar and Paul Smolensky. Learnability in optimality theory. *Linguistic Inquiry*, 29 (2):229–268, 1998.

Bruce Tesar, John Alderete, Graham Horwood, Nazarré Merchant, Koichi Nishitani, and Alan S Prince. Surgery in language learning. 2003.

Erik D Thiessen. Effects of inter-and intra-modal redundancy on infants' rule learning. *Language Learning and Development*, 8(3):197–214, 2012.

Simon Todd, Annie Huang, Jeremy Needle, Jennifer Hay, and Jeanette King. Unsupervised morphological segmentation in a language with reduplication. In Garrett Nicolai and Eleanor Chodroff, editors, *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–22, Seattle, Washington, July 2022. Association for Computational Linguistics.

Suzanne Urbancyzk. *Patterns of reduplication in Lushootseed.* PhD thesis, University of Massachusetts, Amherst, 1996.

Suzanne Urbancyzk. *Patterns of reduplication in Lushootseed.* Psychology Press, 2001.

Harry van der Hulst. Issues in foot typology. *Issues in phonological structure*, pages 95–127, 2000.

Clara Vania and Adam Lopez. From characters to words to in between: Do we capture morphology? In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Rachelle Waksler. Cross-linguistic evidence for morphological representation in the mental lexicon. *Brain and Language*, 68(1):68–74, 1999.

Markus Walther. Finite-state reduplication in one-level prosodic morphology. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000.

Yang Wang. Recognizing reduplicated forms: Finite-state buffered machines. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Online, August 2021a. Association for Computational Linguistics.

Yang Wang. Regular languages extended with reduplication: Formal models, proofs and illustrations. Master's thesis, University of California, Los Angeles, 2021b.

Yang Wang and Bruce Hayes. Learning underlying representations: The role of abstractness. *Linguistic Inquiry*. Resubmitted for review.

Yang Wang and Tim Hunter. On regular copying languages. *Journal of Language Modelling*, 11(1):1–66, 2023.

Wei Wei and Rachel Walker. A Lookahead Effect in Mbe Reduplication: Implications for Harmonic Serialism. *Linguistic Inquiry*, 51(4):845–859, 10 2020.

James White. Evidence for a learning bias against saltatory phonological alternations. *Cognition*, 130(1):96–115, 2014.

James White. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language*, pages 1–36, 2017.

Ronnie Bring Wilbur. *The phonology of reduplication.* University of Illinois at Urbana-Champaign, 1973.

Colin Wilson. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, 30(5):945–982, 2006.

Colin Wilson. Modeling morphological affixation with interpretable recurrent networks: sequential rebinding controlled by hierarchical attention. In *CogSci*, 2018.

Colin Wilson. Re(current) reduplication: Interpretable neural network models of morphological copying. In *Proceedings of the Society for Computation in Linguistics*, volume 2, 2019. Article 56.

Colin Wilson. Learning morphology with inductive bias: Evidence from infixation. In *Proceedings of the 46th annual Boston University Conference on Language Development*, 2022.

Samantha Wray, Linnaea Stockall, and Alec Marantz. Early Form-Based Morphological Decomposition in Tagalog: MEG Evidence from Reduplication, Infixation, and Circumfixation. *Neurobiology of Language*, 3(2):235–255, 02 2022.

Hongzhi Xu, Jordan Kodner, Mitch Marcus, and Charles Yang. Modeling morphological typology for unsupervised learning of language morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6672–6681, 2020.

Lily Xu, Elizabeth Solá-Llonch, Huilei Wang, and Megha Sundara. A meta-analytic review of morphological priming in Semitic languages. *The Mental Lexicon*, 2023.

Meng Yang and Deborah JM Wong. Malay verbal reduplication with the məŋ-prefix. *JSEALS*, page 118, 2020.

Yifan Yang. Rapa nui: A case for correspondence in reduplication. *Linguistic Inquiry*, 54 (2):395–412, 2023.

Sora Heng Yin and James White. Neutralization and homophony avoidance in phonological learning. *Cognition*, 179:89–101, 2018.

Moira Yip. Repetition and its avoidance: The case of Javanese. In *Proceedings of the South Western Optimality Theory workshop 1995.Arizona Phonology Conference Volume 5,*, pages 238—-262, Tucson, AZ, 1995. University of Arizona.

Alan CL Yu. *The morphology and phonology of infixation*. University of California, Berkeley, 2003.

Alan CL Yu. Toward a typology of compensatory reduplication. In *West Coast Conference on Formal Linguistics (WCCFL)*, volume 24, page 397. Citeseer, 2005.

Alan CL Yu. *A natural history of infixation*, volume 15. OUP Oxford, 2007.

Kristine Yu. Computational perspectives on phonological constituency and recursion. *Catalan Journal of Linguistics*, 20:77–114, 2021.

Fujimura Yuko. Reduplication in standard Malay and Japanese. *Journal of Modern Languages*, 13(1):65–92, 2001.

Draga Zec. The syllable. *The Cambridge Handbook of Phonology*, pages 161–194, 2007.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, 23(4):550–560, 1997.

Eva Zimmermann. Database of the DFG project Multiple Reduplication: Typology and theory, 2019. online available at `http://www.evazimmermann.org/multiple-reduplication-data.html`.

Eva Zimmermann. Two is too much…in the phonology! *The Linguistic Review*, 38(3): 537–572, 2021a.

Eva Zimmermann. Faded copies: Reduplication as distribution of activity. *Glossa: a journal of general linguistics*, 6(1), 2021b.

Cheryl Zoll. Ghost segments and optimality. In *Proceedings of WCCFL*, volume 12, pages 183–199, 1993.

Cheryl Zoll. Subsegmental parsing: Floating features in Chaha and Yawelmani. 1994.

Sam Zukoff. *Indo-European reduplication:Synchrony, diachrony, and theory.* PhD thesis, Massachusetts Institute of Technology, 2017.

Sam Zukoff. Reduplicant shape alternations in Tawala: Re-evaluating base-dependence. M.S.

Kie Zuraw. Floating phonotactics: Infixation and reduplication in Tagalog loanwords. Master's thesis, University of California, Los Angeles, 1996.

Kie Zuraw. Aggressive reduplication. *Phonology*, 19(3):395–439, 2002.

Kie Zuraw. Vowel reduction in Palauan reduplicants. In *Proceedings of the 8th Annual Meeting of the Austronesian Formal Linguistics Association [AFLA 8]*, pages 385–398, 2003.

Kie Ross Zuraw. *Patterned exceptions in phonology.* PhD thesis, University of California, Los Angeles, 2000.

Arnold M Zwicky and Geoffrey K Pullum. Plain morphology and expressive morphology. In *Annual Meeting of the Berkeley Linguistics Society*, pages 330–340, 1987.