

The complexity and learnability of phonological patterns

Simulations, experiments, typology

This dissertation investigates the relation between the complexity of phonological patterns, their learnability by simulated and human learners, and their typological distributions.

The research presented in this book is divided into three parts. The first part is dedicated to computational evidence and aims to show how the building blocks of phonological patterns emerge in a neural network as it is exposed to auditory and lexical distributions. Taking a diachronic approach, I show that sound systems evolve towards stable states, regardless of their initial distribution.

The second part describes two experiments with human learners, establishing the learnability of a number of phonological patterns of various degrees of complexity. The results suggest that more complex patterns are more difficult to learn, and that learners reduce the complexity of their input if possible.

The third part assesses attested sound changes and plosive inventories in terms of their complexity. The typological data show that sound change does not necessarily reduce complexity, and that plosive inventories differ greatly in terms of their complexity. The tension between the experimental and typological data sheds light on the interaction of different forces – cognitive, auditory, and articulatory – that shape phonological typology.

ISBN 978-94-6093-369-1

DOI <https://dx.medra.org/10.48273/LOT0584>



Klaas Seinhorst

The complexity and learnability of phonological patterns

Simulations, experiments, typology

THE COMPLEXITY AND LEARNABILITY
OF PHONOLOGICAL PATTERNS

SIMULATIONS, EXPERIMENTS, TYPOLOGY

Published by:
LOT
Kloveniersburgwal 48
1012 CX Amsterdam
The Netherlands

+31 20 525 2461
email: lot@uva.nl
www.lotschool.nl

Cover illustration: [https://commons.wikimedia.org/wiki/File:The Sounds of Earth Record Cover - GPN-2000-001978.jpg](https://commons.wikimedia.org/wiki/File:The_Sounds_of_Earth_Record_Cover_-_GPN-2000-001978.jpg). Source: National Aeronautics and Space Administration (NASA). Public domain.

ISBN: 978-94-6093-369-1
DOI: <https://dx.medra.org/10.48273/LOT0584>
NUR: 616

Copyright © Klaas Seinhorst, 2021. All rights reserved.

THE COMPLEXITY AND LEARNABILITY
OF PHONOLOGICAL PATTERNS

SIMULATIONS, EXPERIMENTS, TYPOLOGY

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. K.I.J. Maex

ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit

op vrijdag 19 februari 2021, te 14:00 uur

door

Klaas Tijmen Seinhorst

geboren te Kampen

Promotiecommissie

Promotor:	prof. dr. P.P.G. Boersma	Universiteit van Amsterdam
Copromotor:	dr. S.R. Hamann	Universiteit van Amsterdam
Overige leden:	prof. dr. J.B. Pierrehumbert	University of Oxford
	dr. S. Moran	Universität Zürich
	prof. dr. C.C. Levelt	Universiteit Leiden
	prof. dr. J.E. Rispens	Universiteit van Amsterdam
	prof. dr. E.O. Aboh	Universiteit van Amsterdam
	dr. J.K. Szymanik	Universiteit van Amsterdam

Faculteit der Geesteswetenschappen

Dankwoord en opdracht

De voltooiing en verdediging van dit proefschrift rust(t)en op ten minste zes verschillende, maar allemaal even onmisbare pijlers.

Paul en Silke, mijn *Doktoreltern*. Ik heb in zo veel dankwoorden een zin gelezen met de strekking “zonder mijn begeleiders had dit proefschrift er heel anders uitgezien” dat die zin elke betekenis leek te verliezen; jullie hebben me laten zien hoeveel waarheid erin kan schuilen. Ik heb waanzinnig veel van jullie geleerd, en kan nog waanzinnig veel meer van jullie leren. Jullie waren geduldig, scherp, en altijd betrokken, zowel binnen de muren van de universiteit als erbuiten. Ik had het niet beter kunnen treffen.

I'm greatly indebted to the members of the reading committee for their willingness to assess my dissertation; I look forward to our discussion.

Een cruciaal onderdeel van dit proefschrift wordt gevormd door de inspanning van de 256 proefpersonen in mijn experimenten (inclusief afvallers en deelnemers aan pilots).

Voor de financiering van dit project ben ik de Nederlandse belastingbetaler zeer erkentelijk, en ik vind het waardevol dat de resultaten van mijn onderzoek openbaar zijn, zonder embargo van partijen met winstoogmerk. Ik heb geprobeerd de Nederlandse samenvatting (p. 201) zo toegankelijk mogelijk te maken voor niet-specialisten.

Precies vanwege die openbaarheid dank ik alle anderen die me in dit traject tot steun zijn geweest hier slechts in deze algemene bewoordingen (maar daarom niet minder oprecht); ik hoop de meesten van jullie ook persoonlijk te kunnen bedanken.

Een uitzondering op de vorige regel vormen mijn grootouders. Mijn grootvader werd als boerenzoon geboren in de Achterhoek, maar had geen aspiraties om het bedrijf van zijn ouders voort te zetten; hij wilde veel liever vergelijkende taalwetenschap studeren. Zijn ouders konden echter geen vervolgopleiding betalen en vroegen de rijken van het dorp om hun gunsten, maar die stelden als voorwaarde dat hij plantenziektekunde zou gaan studeren, zodat de boerengemeenschap uiteindelijk baat zou hebben bij hun steun. Mijn grootvader stemde hiermee in, en heeft zich buitengewoon verdienstelijk gemaakt in zijn onverkozen werkveld. Desalniettemin bleef hij toch ook taalwetenschapper, en zijn voorliefde heeft mij ongetwijfeld beïnvloed; jaren later besloot ik taalwetenschap te gaan studeren — gelukkig had ik die vrijheid wel. Mijn grootvader heeft die beslissing niet meer meegemaakt, maar dankzij mijn grootmoeder, zijn weduwe, heb ik mijn studie kunnen voltooien en aan dit promotietraject kunnen beginnen. Ik draag dit proefschrift op aan hen beiden.

Table of contents

Chapter 1: Introduction	1
1.1 Phonemes and features	1
1.1.1 <i>Distinctivity</i>	2
1.1.2 <i>Features at the level of the language system</i>	3
1.1.3 <i>Features at the level of the language user</i>	4
1.2 Typological tendencies and their explanations	5
1.2.1 <i>Universality and innateness</i>	5
1.2.2 <i>Functional explanations</i>	6
1.2.3 <i>Cultural evolution</i>	7
1.2.4 <i>Learning mechanisms in language acquisition</i>	9
1.3 Typological tendencies in sound systems	10
1.3.1 <i>Nativism and functionalism</i>	11
1.3.2 <i>Markedness</i>	13
1.3.3 <i>Typological diversity</i>	14
1.4 Outline of this dissertation	15
PART I: SIMULATIONS	
Chapter 2: Emergent phonological features in a symmetric neural network	21
2.1 Feature induction in human learners	21
2.1.1 <i>Distributional learning</i>	21
2.2 A bidirectional model of feature induction	23
2.2.1 <i>A bidirectional theoretical framework: the BiPhon model</i>	23
2.2.2 <i>Some properties of the neural network and the input</i>	24
2.2.3 <i>A bidirectional learning algorithm: the inoutstar rule</i>	26
2.3 Emergent features	26
2.3.1 <i>The first stage: distributional learning</i>	27
2.3.2 <i>The second stage: lexicon-driven learning</i>	29
2.3.3 <i>The division of categories across nodes</i>	31
2.4 Auditory versus lexical learning	31
2.4.1 <i>Testing ground: a bimodal distribution</i>	32
2.5 An emergent lexicon	34
2.5.1 <i>Different paces of lexical acquisition</i>	35
2.6 Robustness against variation in word learning	37
2.6.1 <i>Mismatches between meaning and sound</i>	37
2.6.2 <i>Reduced activities in the lexicon</i>	39
2.6.3 <i>Delayed lexical development</i>	40

2.6.4	<i>Mishearings</i>	42
2.7	Phonemes and features (again)	44
2.7.1	<i>First layout: two separate surface layers</i>	45
2.7.2	<i>Second layout: a single surface layer</i>	45
2.8	Grounding, valency, and underspecification	47
2.8.1	<i>Grounding</i>	47
2.8.2	<i>Valency</i>	48
2.8.3	<i>Underspecification</i>	49
Chapter 3: The cultural evolution of auditory contrast		51
3.1	Explanations of auditory dispersion	51
3.2	Speech production in the neural network	52
3.3	A “standard” initial distribution	54
3.4	A skewed, exaggerated initial distribution	56
3.5	A distribution with one bimodally distributed category	60
PART II: EXPERIMENTS		
Chapter 4: Phonological pattern learning (1): a 3×2 parameter space		69
4.1	Pattern learning of non-phonological feature combinations	70
4.2	Complexity measures: feature economy and logical complexity	72
4.2.1	<i>Feature economy</i>	73
4.2.2	<i>Logical (Boolean) complexity</i>	75
4.2.3	<i>Regularity</i>	76
4.3	Pattern learning of phonological feature combinations	77
4.3.1	<i>Category structures of plosive inventories</i>	78
4.3.2	<i>Feature economy, logical complexity, and gaps</i>	80
4.3.3	<i>Regularity (again)</i>	83
4.4	Learning experiments: stimuli, method, and analysis	83
4.4.1	<i>Stimuli</i>	84
4.4.2	<i>Method</i>	87
4.4.3	<i>Average misestimation as a measure of learnability</i>	88
4.4.4	<i>Match as a measure of learnability</i>	89
4.4.5	<i>Analysis</i>	90
4.5	Experiment A: task A1 (implicit learning of handshapes)	90
4.5.1	<i>Error scores</i>	90
4.5.2	<i>Matches</i>	92
4.5.3	<i>Complexity differences and regularisation</i>	92
4.6	Experiment B: task B1 (implicit learning of handshapes)	94
4.6.1	<i>Error scores</i>	94
4.6.2	<i>Matches</i>	96
4.6.3	<i>Complexity differences and regularisation</i>	96

4.7	Experiment B: task B2 (implicit learning of speech)	97
4.7.1	<i>Error scores</i>	97
4.7.2	<i>Matches</i>	98
4.7.3	<i>Complexity differences and regularisation</i>	99
4.8	Feature economy versus logical complexity	100
4.8.1	<i>Error scores</i>	100
4.8.2	<i>Matches</i>	101
4.9	The effect of modality	102
Chapter 5: Phonological pattern learning (2): a 3×3 parameter space		103
5.1	Expanding the parameter space: two ternary features	103
5.2	Experiment C: task C1 (implicit learning of handshapes)	106
5.2.1	<i>Matches and regularisation</i>	107
5.2.2	<i>Reaction times</i>	108
5.2.3	<i>Logical complexity versus feature economy (again)</i>	109
5.3	Experiment C: task C2 (regularisation in a diffusion chain)	110
5.3.1	<i>Regularisation</i>	112
5.3.2	<i>Reaction times</i>	115
5.4	Experiment C: task C3 (classification of Sets)	115
5.4.1	<i>Matches</i>	118
5.4.2	<i>Reaction times</i>	118
5.5	Experiment C: task C4 (production of Sets)	119
5.5.1	<i>Matches</i>	119
5.5.2	<i>Reaction times</i>	119
5.6	Comparison of the implicit-learning tasks (A1/B1 and C1)	119
PART III: TYPOLOGY		
Chapter 6: Complexity in sound changes		125
6.1	Complexity in language change	125
6.1.1	<i>Complexity-reducing linguistic change: compressibility</i>	126
6.1.2	<i>Complexity-increasing linguistic change: communication</i>	127
6.1.3	<i>Sound change</i>	128
6.2	Phonological change: inductive biases only	129
6.2.1	<i>Inductive biases only: cohort B1</i>	129
6.2.2	<i>Inductive biases only: cohort A1</i>	132
6.2.3	<i>Inductive biases only: general predictions</i>	133
6.3	Practice: attested phonological sound changes	133
6.4	Old English to Modern English	135
6.4.1	<i>Obstruents</i>	135
6.4.2	<i>Vowels</i>	136

6.5	The First Germanic Consonant Shift	138
6.6	Zulu	140
6.7	Evaluation	142
	6.7.1 <i>The role of phonetics</i>	143
Chapter 7: Complexity in sound systems		145
7.1	Definitions of complexity	145
7.2	Feature-based analyses of sound systems	146
7.3	Complexity indices of plosive inventories in UPSID	147
	7.3.1 <i>The role of language family</i>	149
	7.3.2 <i>Feature economy versus logical complexity</i>	150
7.4	Category structures	151
	7.4.1 <i>Comparison with experimental data</i>	152
7.5	Inductive biases versus phonetics	153
	7.5.1 <i>Places of articulation</i>	153
	7.5.2 <i>Systems with one gap</i>	153
	7.5.3 <i>Regularisation versus irregularity</i>	153
Chapter 8: Discussion and implications		157
8.1	Computer simulations	157
	8.1.1 <i>The lexicon and speaker normalisation</i>	157
	8.1.2 <i>The lexicon and diachronic merger</i>	158
	8.1.3 <i>Sequential information and diachronic split</i>	158
	8.1.4 <i>Regularisation in the neural network</i>	159
8.2	Complexity measures	159
	8.2.1 <i>Computing complexity indices</i>	159
	8.2.2 <i>Correlated measures of complexity</i>	162
	8.2.3 <i>Feature economy versus logical complexity (one last time)</i>	163
	8.2.4 <i>Gestural economy and perceptual warping</i>	165
8.3	Levels of representation	165
	8.3.1 <i>Phonemes and allophones</i>	165
	8.3.2 <i>Learnability in other domains versus concrete reality</i>	166
8.4	Limitations of the data	167
	8.4.1 <i>The input to the neural network</i>	167
	8.4.2 <i>Modality</i>	168
	8.4.3 <i>Adult versus child learners</i>	168
	8.4.4 <i>Regularisation and cognitive load</i>	169
	8.4.5 <i>Effect sizes in the Markov chains</i>	169
	8.4.6 <i>Analyses of typological data</i>	170

References	173
Summary (in English)	199
Samenvatting (in het Nederlands)	201
Curriculum vitae	207

Author contributions

Klaas Seinhorst (KS) wrote the proposal for the PhD project that resulted in this dissertation, and all chapters in this dissertation. Paul Boersma (PB) and Silke Hamann (SH) co-supervised the project and provided feedback on earlier versions of all chapters.

The computer simulations in **Chapters 2 and 3** were run by KS, using a script that was created by PB and expanded by KS. These chapters expand on two earlier publications. The first is Boersma, Benders and Seinhorst (2020), specifically §6; this section was written by KS, with feedback from PB and Titia Benders. The second is Seinhorst, Boersma and Hamann (2019), which was written by KS with feedback from PB and SH.

The experiments in **Chapter 4 and 5** were designed by KS, with input from PB and SH. Chapter 4 expands on two articles that have appeared earlier: the data from task A1 were published as Seinhorst (2017), and the data from task B1 were published as Seinhorst (2016a). The introduction of Chapter 4 expands on the introduction from Seinhorst (2017). KS implemented and conducted the experiment from Chapter 4; he implemented the experiment from Chapter 5, and conducted it with the help of a student assistant, Floor van de Leur (FvdL). KS analysed the data from both experiments, consulting PB for statistical advice. KS wrote both papers mentioned above, with feedback from PB and SH.

The data from the Markov chain based on task B1, presented in **Chapter 6**, were part of Seinhorst (2016a). The sound change data in this chapter appeared as Seinhorst (2016b). KS wrote this paper, with feedback from SH. The data about the sound systems in **Chapter 7** were collected by FvdL as part of her BA thesis project, supervised by KS. These data, along with a comparison of the complexity measures, will appear in Seinhorst and Van de Leur (under review). This paper was written by KS, with feedback from FvdL.

1

Introduction

The photo on the cover of this dissertation shows the casing of a golden phonograph record that was sent into outer space by NASA, aboard the Voyager spacecraft, in 1977. The record is intended to illustrate the diversity of life and culture on Earth to any kind of extraterrestrial life form that may encounter it, by means of various pictures and sounds: spoken greetings in 55 languages, a short excerpt from Igor Stravinsky's *The Rite of Spring*, a piece of gamelan music, a page from a book by Sir Isaac Newton, a recording of human brain waves, and a photograph of the Golden Gate Bridge, just to name a few. All these pictures and sounds have been divided into four sections, one of which, named "Sounds of Earth", contains short fragments of sounds like thunder, animal cries, the noises of various vehicles, and the motto *per aspera ad astra* ("through hardships to the stars") in Morse code.

Much like the golden record, this dissertation is concerned with the sounds of Earth, albeit a specific subset of them: the linguistic sounds of Earth – that is, the units of speech that the spoken languages on this planet use to distinguish between different meanings – and with the way that languages organise these units into inventories. (Unlike the golden record, this dissertation is not aimed at an audience of extraterrestrials.) More specifically, the research presented in this dissertation explores the questions how the elements of sound systems emerge, and what roles the notions of complexity and learnability play in the acquisition, change and typology of phonological systems. To do so, I collect and evaluate computational, experimental and typological evidence. In this introductory chapter, I sketch the backdrop against which the research described in this dissertation is set.

1.1 Phonemes and features

In order to convey meaning, spoken languages combine units that are in themselves meaningless. In English, the sounds [b], [æ] and [t] are semantically empty in isolation, but when they are concatenated in this order, they refer to a specific category of objects, namely a certain kind of mammal. The sounds cannot be scrambled at random without consequences: if the same set of sounds is concatenated in reverse order, it refers to an entirely different category of objects. All natural languages, spoken as well as signed, combine a relatively small number of meaningless elements into a potentially unlimited number of meaningful forms, a property called "double articulation" (Martinet 1960) or "duality of patterning" (Hockett 1963).

1.1.1 Distinctivity

If we replace the [t] from [bæt] with a [d], the resulting word has a different meaning; we are no longer referring to a category of objects, but to an attribute of objects. This simple observation provides an insight into the organisation of the sound system of the English language: it tells us that the difference between [t] and [d] is **DISTINCTIVE** (or **CONTRASTIVE**) in English, that is, this difference distinguishes between meanings — after all, being a *bat* is not the same thing as being *bad*. By definition, then, [t] and [d] are **PHONEMES** of English: abstract phonological categories with a lexically contrastive function. In Korean, the sounds {t d} both occur as well, but they are not contrastive; rather, a single phoneme [t] is posited, which predictably changes to /d/ intervocalically (Cho, Jun & Ladefoged 2002). Compounding *su* [su] ‘water’ and *to* [to] ‘way’, for instance, yields *sudo* /sodo/ ‘waterway, waterworks’ (Hyman 2018: 4). In Korean, then, /t/ and /d/ are not phonemes, but **ALLOPHONES**, or positional variants, of one phoneme. The relations between phonemes and allophones in a sound system can be manifold: in Dutch, for instance, the phoneme [d] usually surfaces as /d/, but it may undergo a phonological process (such as voicing assimilation or final devoicing) and surface as /t/; the phoneme [t] usually surfaces as /t/, but it may undergo a phonological process (such as voicing assimilation) and surface as /d/. The words [çeyt+bɛd] ‘out of bed’, for instance, surface as /çeydbet/.

In the previous sentences, I used the verb “surface”, which suggests different possible depths of representation; these differences are also reflected in the various symbols between which I notated segments in the previous paragraph. The property of distinctivity is inherently related to the lexicon, so in terms of phonological representations, phonemes are defined in a so-called **UNDERLYING FORM** (Chomsky & Halle 1968; Prince & Smolensky 1993/2004), in which lexical material receives its phonological form. Allophones are not distinctive and are therefore defined on a different level of representation: this level is commonly named the **SURFACE FORM**, and it is structured in terms of prosodic constituents such as phrases, feet and syllables. Throughout this dissertation I write underlying forms in [pipes], surface forms in /slashes/, and auditory forms in [square brackets]. Sets of segments or syllables are written in {curly brackets}.

The term “phoneme” was first used in 1855 by the Bulgarian philosopher Petar Beron, in the meaning ‘sound sequence’ (Mugdan 2014); in the early 1860s, the French merchant and amateur phonetician Antoni Dufriche-Desgenettes used the term to signify “an element of a language-specific or universal sound inventory” (Mugdan 2014: 186). The usage of the term in its current meaning, i.e. a semantically contrastive phonological category, was developed by the Polish linguists Jan Baudouin de Courtenay and Mikołaj Kruszewski in the late 19th century; this definition has since been widely adopted (Jones 1919, 1931; Sapir 1921; Jakobson, Karcevsky & Trubetzkoy 1928; Bloomfield 1933; Trubetzkoy 1939; and many others).

Most phonologists agree that phonemes and allophones can be regarded as bundles of atomic properties. In structuralist phonology, such a property is called a phonological FEATURE (Jakobson, Karcevsky & Trubetzkoy 1928; Jakobson 1941), a primitive that is usually assumed to be grounded in auditory or articulatory properties, such as voicing, stridency, or place of articulation. For instance, the symbol /t/ denotes a segment that, regardless of its phonemic status, simultaneously possesses the properties of coronal place, voicelessness, and plosive manner. The notion of distinctivity applies to features as well: a feature is said to be distinctive (or contrastive) if it distinguishes between minimal pairs. For instance, the voicing feature is distinctive in English, because [t] and [d] are phonemes of this language, and these phonemes differ only with respect to their voicing. Chapter 2 investigates the emergence of phonological features in the individual language learner; I further discuss the issue of their phonetic grounding in §2.8.1 (p. 47).

1.1.2 Features at the level of the language system

Sound systems are usually visualised in matrices where the rows and columns correspond to the constituent distinctive features. By way of example, Table 1.1 shows the obstruent system of Bisa (Maddieson 1984: 286), a language spoken in Burkina Faso.

Table 1.1. *The obstruents of Bisa.*

		labial	coronal	dorsal
plosive segments	voiceless	p	t	k
	voiced	b	d	g
fricative segments	voiceless	f	s	
	voiced	v	z	

Phonologists use features not only in the analysis of phoneme inventories, but also in the formalisation of phonological processes. These tend to target natural CLASSES, groups of segments that can be exhaustively described in terms of one or more features: for instance, /p b f v/ is a natural class in Bisa (namely the class of labial obstruents), but /p b s z/ is not. The tendency of processes to affect natural classes holds both synchronically and diachronically. For instance, the process of final devoicing in Toba Batak only takes place in voiced stops; only back vowels undergo umlaut in German; and in the First Germanic Consonant Shift, each of the three changes targeted different natural classes: voiceless stops in the first, voiced stops in the second, and voiced aspirated stops in the last (cf. §6.5, p. 138). This is not to say that all phonologically active classes, that is, classes that participate in phonological processes, can be neatly characterised in terms of features: in a typological survey of 6077 classes, Mielke (2008: 118) found this to be true in only 75.35% of the cases.

Natural classes tend to be SYMMETRICAL or REGULAR, that is, they tend to disprefer gaps. Trask (2000: 326) illustrates this with the fricatives of Old English, which had {f θ s ʃ x}. The introduction of the voiced fricatives {v ʒ} in loanwords from French yielded an asymmetric inventory, which, according to Trask, was resolved in two steps: first, |ð| came into existence, so that |θ| would not be left without a voiced counterpart; second, |x| disappeared, likely because a contrastively voiced |ɣ| was absent. These changes resulted in a symmetrical inventory. Trask does not explain why the two gaps were resolved in different ways; he also neglects to mention that /ɣ/ existed as an intervocalic allophone of |x|. The notion of symmetry and its role in phoneme inventories takes centre stage in Chapters 4–7.

1.1.3 Features at the level of the language user

The fact that phonologists are able to observe the role of features in phonological systems and processes could be explained by assuming that features have psycholinguistic reality in the brain of the individual speaker-listener. Although this hypothesis was discarded as “mentalist” and unfalsifiable by some (Bloomfield 1933; Twaddell 1935), it was shared by many other structuralist phonologists, such as Sapir (1921, 1933), Sommerfelt (1928), and Benni (1929: 36), who characterises the phoneme as “ein Psychophon”. Since the cognitive revolution and the subsequent rise of psycholinguistic research in the 1950s and ’60s, it has become possible to study the cognitive nature of linguistic representations, such as the feature. In a classic study, Liberman et al. (1957) tested listeners’ discrimination of pairs of sounds that differed on a single auditory continuum: they found that listeners had more trouble discriminating between stimuli if these belonged to the same phonological category than if they belonged to different categories, even if the acoustic differences between the stimuli were the same. This phenomenon, known as CATEGORICAL PERCEPTION, suggests that auditory perception is mediated by phonological categories.¹ Chládková (2014, ch. 2) conducted an elegantly designed categorisation experiment showing that Czech listeners perceive synthesised tokens from the central part of the acoustic vowel space, where no category exists in Czech, in terms of features rather than phonemes. Also, the F₁ trajectories of the Dutch close diphthongs [ei], [øy] and [ou] are identical for front and back vowels (Van der Harst 2011: 11), suggesting that their production is modulated by the same height feature specifications. More recently, neurophysiological methods, such as electroencephalography (EEG), have been applied to explore the nature of phonological representations in the brain of the individual speaker-listener (a.o. Chládková 2014, ch. 4; Schluter, Politzer-Ahles & Almeida 2016; McCloy & Lee 2019).

¹ Liberman et al. assume that the mediating categories are phonemes instead of features, to which they refer as “the smallest unit of speech” (p. 358, footnote 3); however, the stimuli in their experiment differ with respect to the direction and extent of the second formant transition, which they take to be the main auditory correlate of the phonological place feature.

1.2 Typological tendencies and their explanations

This dissertation is not just concerned with speech sounds, but more specifically with the typology of sound systems. Typological research is centered around two main questions: (i) what kind of diversity exists in the cross-linguistic distribution of traits such as word order, stress patterns, and so on? and (ii) how can this diversity be explained? This section provides a brief introduction to some explanations of general typological patterns; I address phonological typology in §1.3 (p. 10).

1.2.1 Universality and innateness

Hauser, Chomsky and Fitch (2002) assert that if extraterrestrial aliens were to land on earth, they would surely think that all humans speak a single language, their mutually unintelligible lexicons aside. Indeed, there are many similarities between the spoken languages of the world, and these have led linguists to formulate linguistic universals, such as Hockett's (1963) design features (including duality of patterning, mentioned in §1.1), Greenberg's (1963) morphosyntactic universals, and Pinker and Bloom's (1990) substantive universals. Chomsky (1965) famously devised a theory of Universal Grammar (UG), among other things assuming that syntactic categories such as nouns and verbs are universal. Moreover, considering the fairly small amount of input that children seem to need to acquire language (the "poverty of the stimulus" argument), as well as children's ability to create novel utterances, Chomsky postulated that such rich linguistic knowledge must be part of the human genetic make-up, and therefore present at birth. The framework of Principles and Parameters (Chomsky 1981; Chomsky & Lasnik 1993) posits that languages have only a small number of innate parameters at their disposal; in this view, the task of the language-acquiring child is to find the appropriate parameter settings for her native language(s). By 2002, Chomsky had reduced the list of innate properties of language to recursion only (Hauser, Chomsky & Fitch 2002).

However, of all languages that have ever existed on earth, only a fragment is still currently spoken; of this subset, only a subset has been or will ever be documented, making it available for typological study. Considering this small sample of languages, it seems imprudent to posit a theory of putative universals. Moreover, typologists have discovered a degree of diversity that is not predicted by UG, which has called the notion of universality into question. For instance, languages can mark tense on nouns (Nordlinger & Sadler 2004); Straits Salish seems to lack a noun vs. verb distinction altogether (Jelinek 1995; cf. also Haspelmath 2012 about the pitfalls of comparing word classes between languages); and the allegedly universal CV syllable is absent in Arrernte (Breen & Pensalfini 1999); for more examples, see Evans and Levinson (2009). In the face of such typological diversity, the assumption of highly specified innate linguistic knowledge seems hard to maintain.

1.2.2 Functional explanations

An alternative approach to typological tendencies starts from an ecological perspective, stressing that all human languages serve the purpose of communication (a.o. Martinet 1962; Dik 1989; Hengeveld & Mackenzie 2008): interlocutors use it to share propositions, gather information, instruct others, and so on. In this view, typological tendencies are a result of the communicative function that all languages share. For instance, all 31 languages in Dingemanse, Torreira and Enfield's (2013) sample have strategies to signal a breakdown in communication; Roberts and Levinson (2017) argue that the process of turn-taking in conversation may have an effect on typological tendencies in word order; and only 1 out of 955 languages in Dryer's (2013) sample does not use any morphosyntactic or phonological strategy to distinguish between statements and questions.

The interactive nature of language displays itself in the way in which communication systems adapt to their environment: language is a fundamentally modality-neutral system that can be transmitted through whatever means is most appropriate or convenient. Speech is the most common modality, but by no means the only one available: sign languages are used if the auditory modality is unavailable for any reason, for instance if one or more interlocutors are deaf, or if loud machines in a factory make spoken conversation impossible (Meissner, Philpott & Philpott 1975); some tone languages use drums or whistling instead of speech if larger distances need to be spanned (cf. Seifart et al. 2018 on Bora, an Amazonian language in which communication may proceed through drumming; and Rialland 2005 on whistled languages); tactile signing can be used if neither the visual nor auditory modalities are available (Mesch 2001); and so on.

Furthermore, all humans are endowed with a roughly identical speech and hearing apparatus, and with the same basic cognitive capacities. For instance, typically developed individuals possess a tongue, lips, teeth, a larynx, lungs, cochleas, and auditory nerves; they also possess the abilities to create and maintain mental representations of objects, and they are able to memorise, process and categorise incoming stimuli. This machinery likely plays a role in explaining certain cross-linguistic tendencies, even though it is also used for things other than language: for instance, "heavy" constituents are often clause-final, purportedly because of memory and processing constraints (a.o. Hawkins 1994; Haspelmath et al. 2001). In a similar vein, some theories of language acquisition stress the interactional nature of language: Tomasello (2003, 2009), for instance, argues that children learn language because they are able to recognise intentions and goals in interaction, as well as find patterns in language. Chater and Christiansen (2018) regard language acquisition as learning the skill of successful interaction, not so much the induction of a grammar, although they do not deny the existence and relevance of grammatical knowledge.

1.2.3 Cultural evolution

Any biases in human capabilities to acquire, process, perceive and produce language – such as inductive biases, memory limitations, and perceptuomotor biases – may be amplified as time passes: language is acquired by many consecutive generations of learners, each of whom may unwittingly change the language system in accordance with their biases. This development eventually gives rise to apparent universals, because languages, independently from one another, reach the same “solutions” (Hurford & Kirby 2002; Christiansen & Chater 2008; Chater & Christiansen 2010); the emergence of linguistic structure at the level of the language system is an unintended consequence of interactions between individuals, somewhat similar to the invisible hand effect in economics. Rather, languages have evolved to become learnable because they have to pass through the so-called “transmission bottleneck”: a learner has to construct a grammar on the basis of a small fraction of all possible utterances their teacher could possibly produce, and this process will proceed more faithfully if the language is more learnable. Herein also lies a reply to the poverty of the stimulus argument: children acquire language with apparent ease because it has become tuned to their learning preferences.

This kind of evolution is called CULTURAL EVOLUTION, in order to distinguish it from biological, genetically encoded evolution. Cultural evolution occurs not only in language, but in all sorts of learnt behaviour, such as tool use, social norms, music, religion, and so on; cross-cultural similarities seem to have developed in these areas as well (cf. Strenski 2006 for religion; Savage et al. 2015, Mehr et al. 2019 for music). For instance, musical rhythm tends to be isochronous, that is, music usually has a regular beat; melodic lines tend to be short, likely because of memory constraints; and in musical scales, the octave often plays a central role, probably because humans perceive tones that are an octave apart as somehow the same. This ability, called “octave generalisation”, has already been attested in human infants (Demany & Armand 1984), and also in white rats (Blackwell & Schlosberg 1943) and rhesus monkeys (Wright et al. 2000).

A paradigm called ITERATED LEARNING mimics the process of cultural evolution: in this paradigm, the output of one participant constitutes the learning input to the next participant. This way, a chain of learners, a so-called diffusion chain, is created. Each participant in an iterated learning study is taken to represent a generation of learners, similar to the way children learn from their caregivers or teachers, thus providing ecological validity to the framework. Although the process of cultural evolution has been studied most extensively over the last two decades or so, the framework is much older: an example of much earlier work is Bartlett (1932), who tracked which elements and properties of stories and pictures were preserved throughout a diffusion chain.

The iterated learning paradigm is used in computer simulations as well as behavioural experiments, with linguistic as well as non-linguistic stimuli. Kalish,

Griffiths and Lewandowsky (2007), for instance, investigated inductive biases in the learning of mathematical functions, finding a bias towards a simple linear function $f(x) = ax$; Ravnani et al. (2018) had learners repeat rhythmic patterns that were random initially but evolved into isochronous patterns. Kirby, Cornish and Smith (2008) defined a semantic space in which three shapes existed (square, circle, triangle), each of which had one of three colours (grey, red, blue) and moved in one of three ways (straight line, circle, wave). For each of the resulting $3^3 = 27$ combinations, they made up a random, monomorphemic word. In a diffusion chain, most of the semantic features eventually became reliably encoded, i.e. each property of the stimuli would come to be expressed with a dedicated morpheme; after ten generations of learners, the language would often consist of nine different morphemes that could be combined to express any of the 27 meanings. This result suggests that such reliable meaning-to-form mappings exist in natural languages because they are more learnable than a lexicon of random, unrelated words. Smith and Wonnacott (2010) created a semi-artificial language in which two nominal plural markers existed, but each noun could occur with either marker. In the initial language, the choice for the marker was unpredictable, but such unpredictable variation was eliminated in a diffusion chain. In line with these findings, Kirby et al. (2015) argue that cultural evolution results in reduced complexity, and that languages have evolved to strike a balance between maximal expressivity and minimal complexity; see also §6.1.1 (p. 126).

The transmission of information between generations is called “vertical transmission”; “horizontal transmission” involves the repeated interaction of learners with members of their own generation. This process of negotiation, too, appears to result in the emergence of structure (De Boer 2000, 2001; Selten & Warglien 2007; Vogt & Divina 2007; Moulin-Frier et al. 2015). According to Dahl (2004: 60), horizontal transmission plays a bigger role in language change than vertical transmission, because children tend to speak like their peers, not like their parents.

Well-documented real-life cases of the emergence of linguistic structure are found in Al-Sayyid Bedouin Sign Language (ABSL), a sign language that emerged some seventy years ago in an isolated community with a high prevalence of deafness, and Nicaraguan Sign Language (NSL), a sign language that originated in Deaf schools in Managua in the 1970s and 80s. The second generation of ABSL signers developed a preference for clause-final predicates (Sandler et al. 2005); NSL developed such properties as verb agreement, and structured encoding of motion events (Senghas, Kita & Özyürek 2004). Both languages are the topic of vigorous debate, since proponents as well as opponents of UG regard them as support for their theoretical stance: proponents consider them as evidence of a “language-ready brain” (Kegl 2002), while opponents stress the roles of communication and cultural evolution (Sandler et al. 2005). The creolisation of pidgins is often mentioned as another

example of emergent linguistic structure (e.g. McWhorter 2001), but the relation between pidgins and creole languages is not straightforward (Mufwene 2002; Muysken & Smith 2008): not all pidgins evolve into creoles, and not all creoles have evolved out of pidgins.

A few central questions in language evolution research are how and when human language originated, what communicative skills other animal species possess, and how the necessary neurological and genetic substrates for these skills evolved, both in humans and in other species. Many examples of communication systems exist in other animals: songbirds sing to mark their territory and attract mates; bees convey the location of a food source through dance; numerous species use vocalisations to signal danger; and so on. Social learning, an important prerequisite for human language, is also found in several species of animal: bumblebees are capable of learning a simple trick to obtain food by observing conspecifics who have already learnt this trick (Alem et al. 2016), as are octopuses (Fiorito & Scotto 1992); lizards who are shown a video of another lizard using its legs to access a food source can imitate this behaviour (Siviter et al. 2017); social transmission of tool use occurs in chimpanzees, both in captivity (Horner et al. 2006) and in the wild (Hobaiter et al. 2014). Nevertheless, human language seems to be the most elaborate and expressive communication system found in nature.

1.2.4 Learning mechanisms in language acquisition

Over the last decades, much research has been done into the learning mechanisms that language-acquiring infants have at their disposal.

One such learning mechanism involves the extraction of sequential information. A major problem that the learners face is segmentation: they need to learn to divide a continuous stream of speech into discrete segments and individual words. Experimental evidence suggests that children are capable of tracking transitional probabilities between sounds, both within and between words: if the child has learned that a sequence of two sounds is unlikely to occur within a word, she will assume a word boundary to intervene between these sounds. This ability has been attested in eight-months-old infants, even after an exposure of only two minutes (Saffran, Aslin & Newport 1996). Marcus et al. (1999) found that seven-month-olds can keep track of transitional probabilities between words as well, in order to extract morphosyntactic rules; they do so better with speech than with other stimuli, such as pure tones or animal sounds (Marcus et al. 2007), likely because they recognise the communicative function of speech (Hoff 2006; Ferguson & Lew-Williams 2014).

Another learning mechanism is DISTRIBUTIONAL LEARNING, in which the infant learns from probability distributions in her environment. For instance, by six months of age, infants are still able to discriminate between auditory contrasts that do not occur in their native language; by the time they are twelve months old, they

have lost this ability and instead have tuned in to the contrasts that are distinctive in their language (Werker et al. 1981; Werker & Tees 1984). The same perceptual narrowing happens in deaf children who acquire sign language (Palmer et al. 2012). Wanrooij, Boersma and Van Zuijlen (2014a) already found an effect of distributional learning in two- to three-month-olds; Wanrooij, Boersma and Van Zuijlen (2014b) found that distributional training is less effective for adults than for infants. Distributional learning is one of the crucial learning mechanisms in the computer simulations of the emergence of phonological features in Chapters 2 and 3 of this dissertation.

1.3 Typological tendencies in sound systems

As Hyman (2008) points out, it is not straightforward to establish universals in sound systems: many issues, both practical and theoretical, face the typologist. Should she look at directly observable surface segments, or at underlying segments? In the latter case, the segments need to be inferred, a process that may cause disagreement and that strongly depends on the theoretical stance of the phonologist. For instance, Everett (1986) controversially analyses the surface segment /k/ in Pirahã as an allophone of an underlying |hi| sequence. Another problem is that phonological categories are language-specific, and therefore cannot be compared directly between languages; remember from §1.2.1 that Haspelmath (2012) raised a similar issue about the comparison of word classes between languages. Nevertheless, it seems possible to draw a number of broad generalisations over the sound systems of spoken languages: for instance, all spoken languages that have been described so far alternate consonants and vowels, distinguish at least two vowel heights, and use plosive segments.

Languages also seem to strongly prefer certain phonemes over others, as becomes obvious when we look at databases of sound systems. Figure 1.1 shows the relative frequencies of the 250 most frequent phonemes in the PHOIBLE database (Moran & McCloy 2019), containing the phoneme inventories of 2186 languages. The frequency rank is shown on the horizontal axis, and the number of languages in which the phoneme is found on the vertical axis. The relative frequencies follow an exponential distribution: a small number of phonemes occurs in a large number of languages (although no phoneme is found in all languages in the database), and many phonemes are only found in a handful of languages. The broad transcriptions of the ten most frequent phonemes in PHOIBLE, in descending order of incidence, are |m i k j u a p w n t|, and the relative frequencies of these sounds range from 96% for |m| to 68% for |t|. The figure clearly shows how long the tail of this frequency distribution is: for instance, the 100th most frequent phoneme occurs in only 5% of the languages in the database.

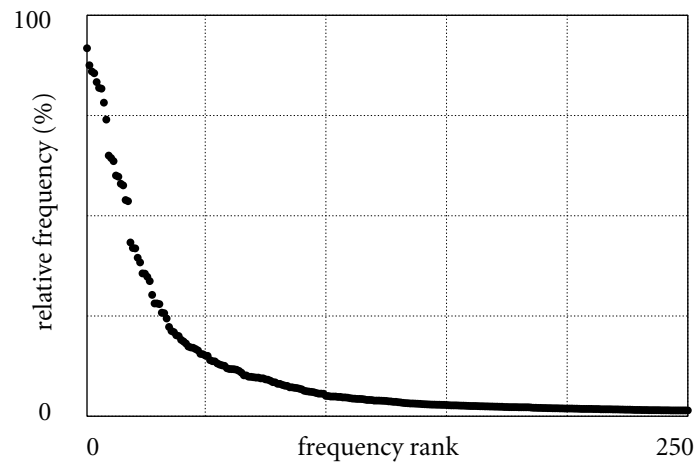


Figure 1.1. *Relative frequency distribution of the 250 cross-linguistically most frequent phonemes in PHOIBLE (Moran & McCloy 2019).*

1.3.1 Nativism and functionalism

In a nativist view, tendencies in sound systems are explained by assuming that languages draw a subset of phonological features from a finite, universal set, hypothesised to be part of our innate linguistic knowledge (a.o. Chomsky & Halle 1968; Prince & Smolensky 1993/2004; Clements & Hume 1995). Learners need to discover which features are relevant to describe the phonemes and allophones in their ambient language(s), based on minimal pairs and phonological alternations; the mappings between the features and their auditory and articulatory correlates are assumed to be automatic and universal (Chomsky & Halle 1968; Chomsky & Lasnik 1993).

In a functionalist approach to phonology and phonetics (a.o. Passy 1890; Martinet 1955, 1960), typological tendencies are ascribed to general cognitive and perceptuomotor biases, where the word “general” means ‘not unique to language’. More specifically, two forces are assumed to be at play: on the one hand, a preference for maximal AUDITORY DISTINCTIVENESS, and on the other hand, a propensity towards maximal ARTICULATORY EASE. This means that speakers prefer unambiguous auditory cues that are least likely to be perceived by listeners as any category other than the intended one, while also aiming to expend as little muscle activity as necessary. These forces oppose each other, since the least ambiguous cues tend to be auditorily peripheral, and the production of such tokens usually requires more articulatory effort than the production of more central tokens. For instance, a highly distinct token of [i] would have a very high second formant, requiring an extreme tongue body position, probably with additional lip spreading to further increase the second formant (F_2).

The interaction of auditory and articulatory factors manifests itself in various ways. These factors may explain why all languages seem to have plosives: these sounds are articulatorily simple, requiring only a ballistic closure within the oral cavity, with maximally salient auditory effect, namely going from the loudest possible sound (a vowel) to silence and back (Ohala 1996; Boersma 1998). The observation that [k] seems to be the most frequent plosive in PHOIBLE may be attributed to its velar place, since only a small body of air needs to be trapped behind its constriction, requiring relatively little articulatory effort as compared to plosives at other places, for the same intensity of the burst.² Also, such processes as reduction (Koopmans-Van Beinum 1980; Ernestus 2000; Johnson 2004) and grammaticalisation (Bybee, Perkins & Pagliuca 1993; Hopper & Traugott 2003) are explained fairly straightforwardly: once a word has become predictable and hence expected (either in a specific context, or in the entire language), the speaker will expend less articulatory effort to pronounce it; the listener is able to recognise reduced pronunciations of the word as long as context is provided (Kemps et al. 2004); in a priming experiment, Van de Ven, Tucker and Ernestus (2011) found that once a reduced prime has been semantically processed, unreduced and reduced primes facilitate the recognition of upcoming words equally well. Regarding the structure of sound systems, Quantal Theory (Stevens 1972, 1989; Stevens & Keyser 2010) posits that such systems prefer sounds in which auditory cues are robust to articulatory imprecisions. In Chapters 2 and 3 I provide an explicit model of the interaction of auditory and articulatory forces, both at the level of the individual learner and at the level of the language system.

The tendency toward efficiency is not only at work in phonology and phonetics: animals (including humans) tend to choose the least effortful way to achieve their goal in any kind of behaviour (Ferrero 1894; Zipf 1935, 1949). Favaro et al. (2020), for instance, found that in the vocalisations of African penguins, the most frequent syllable is also the shortest one. Another example can be found in the desire paths that pedestrians may create to shorten their route: they want to get from point A to point B, but they will invest as little effort as possible. From a functionalist perspective, it is not surprising that such efficiency is found outside of language as well, because language is assumed to share its cognitive, perceptual and articulatory machinery with other kinds of behaviour.

² The size of this advantage decreases as the location of the closure moves towards the front of the oral cavity. Nevertheless, in PHOIBLE, [p] is the second most frequent plosive after [k]; this may be related to the labiality of [p], which makes it visually more salient than a non-labial plosive (cf. McGurk & MacDonald 1976 about the relevance of visual information in speech perception).

1.3.2 Markedness

An important concept in phonological theory is MARKEDNESS (a.o. Jakobson 1941; Chomsky & Halle 1968; Prince & Smolensky 1993/2004). Some feature values, combinations of feature values, or phonotactic properties, are assumed to be marked, while others are unmarked. Marked and unmarked elements stand in an implicational-hierarchical relationship: the presence of a marked element in a language implies the presence of the concomitant unmarked element. Markedness relations may relate to several sources of evidence, such as order of acquisition, or typological distributions: unmarked properties are those that are acquired early or that are typologically frequent. For instance, syllables without onsets are considered marked: they are cross-linguistically less common than syllables with onsets; languages that allow onsetless syllables tend to allow syllables with onsets too (with the exception of Arrernte), whereas the opposite is not true; and syllables with onsets tend to be acquired before syllables without onsets (Levelt, Schiller and Levelt 1999 found that Dutch children learn to produce CV syllables before V ones, and CVC syllables before VC ones). Similarly, if a language has voiced obstruents, it is predicted to have voiceless obstruents too, as [voiced] is the marked feature value.³

In a nativist view, a trait is typologically frequent and acquired early exactly *because* it is unmarked: the markedness is prime. In a functionalist perspective, when a linguistic trait is typologically frequent or acquired early, this is explained by articulatory and auditory factors: for instance, all spoken languages probably have plosives because they are easy to produce, and because their low sonority contrasts maximally with vowels. A correlation exists between the amount of spectral energy and the discharge rate of the neurons in the auditory nerve (Kiang 1980), meaning that the neurons fire most frequently during the syllable nucleus; this entails that sequences of segments with a large difference in sonority are preferred, because this sonority difference makes the transition between the segments more prominent (Delgutte 1982, 1997). This neurophysiological fact of auditory perception explains the observation that languages tend to prefer onsets, so that the syllables goes from lower to higher intensity; this difference is largest if the onset is a plosive.

On the basis of typological evidence, Boersma (1998: 454–457) argues that the sonority scale is not innate, but instead is shaped by any function it happens to be used for; he also argues that it is not necessary to assume that phonological features, representations, and constraints are innate, because the learner is able to induce them from sufficient input. Boersma (2008) provides explicit formalisations of

³ Yidj, an Australian Aboriginal language, has only voiced plosives (Dixon 1977); if we would like to adhere to the assumption that the [voiced] feature value is marked, we could analyse the plosives as voiceless underlyingly. Not only is such an analysis circular, but for Yidj, Dixon (p. 32) argues against it because “[i]t is, in fact, normal for the glottis to be vibrating throughout the articulation of a Yidj word”. See also the discussion of the plosive inventory of Bandjalang in §7.4 (p. 151).

phonological acquisition in which markedness emerges because of properties of the input, again showing that the assumption of innateness is not strictly necessary. In Chapter 2 and 3, I follow this idea by modelling phonological features as emergent, rather than innate, categories.

1.3.3 Typological diversity

In §1.2.1 I listed a few examples of linguistic traits that display a considerable degree of typological variation in the languages of the world. Another example can be found in the size of sound systems: on the smaller end, we find languages like Pirahã, spoken in the Amazon, with ten phonemes (plus two lexical tones) (Everett 1986), and Rotokas, spoken in part of Papua New Guinea, with eleven phonemes (Firchow & Firchow 1969). On the other hand, Taa (also known as !Xóõ), spoken in parts of Botswana and Namibia, is famous for being found at the opposite end of the spectrum; estimates of its size range up to 163 phonemes (Traill 1985). According to Maddieson (2013), the consonant inventories in 2662 the languages described in the World Atlas of Language Structures (WALS; Dryer & Haspelmath 2013) on average contain 22.7 segments; the average size of the vowel inventories of these languages is just under 6.

The creation of large databases with descriptions of languages, such as Glottolog (Hammarström, Forkel & Haspelmath 2019), PHOIBLE (Moran & McCloy 2019), and the WALS mentioned above, has made it easier to gauge the extent of linguistic diversity; moreover, such databases also provide a sufficiently large and diverse sample for researchers to test correlations between typological facts and possible explanatory factors, either linguistic or non-linguistic. For instance, Lupyan and Dale (2010) performed a search across 2236 languages and found significant correlations between social structure and several measures of morphological complexity, positing that languages adapt to the environment in which they are used: in languages with large groups of adult learners, certain traits are more likely to be lost, because late L2 acquisition tends to proceed imperfectly (a.o. Meisel 2011; see also §6.1.1). However, some of Lupyan and Dale's effects are only significant in a model that does not group observations from the same language family together (cf. §7.3.1); these effects disappear when language family is added to the model (cf. Lupyan and Dale's Table 1).

Another strategy to investigate explanations of linguistic diversity is computer modelling. Traill (1985: 101) observes that Khoisan speakers differ from of non-Khoisan speakers because four out of five Khoisan informants “do not have an alveolar ridge”, where non-Khoisan speakers do; a computer model by Moisik and Dediu (2017) suggests that this anatomical difference makes clicks easier to articulate for Khoisan speakers, and therefore more frequent in their languages.

A second, quite striking, example of the interplay between anatomy and language was put forward by Blasi, Moran et al. (2019), who argue that the production

of labiodentals only became possible when the human bite configuration changed to an overbite after the Neolithic, as a result of changing diets after the rise of agriculture. Blasi, Moran et al. estimate that labiodentals occur 27% less frequently in sound systems of languages currently spoken in hunter-gatherer societies than in languages currently spoken in food-producing societies; they also estimate that labiodentals were less likely to have occurred in early time periods of the Indo-European language family than in the current time period. Another interesting example of a non-linguistic factor explaining linguistic diversity is described by Butcher (2006): he ascribes the tendency of Australian Aboriginal languages to lack high vowels as well as fricative obstruents to the observation that chronic *otitis media* is highly prevalent in the Aboriginal population. This condition has a strong negative effect on listener's perception of frequencies under 400 Hz, impeding the perception of the first formant in high vowels, and above 5000 Hz, the part of the spectrum in which the noise of some fricative sounds is concentrated.

1.4 Outline of this dissertation

On the first page of this chapter, I promise to “collect and evaluate computational, experimental and typological evidence”. In order to present the different sources of empirical data in a clear-cut, systematic manner, I have divided the data in this dissertation into three parts.

The first part (Chapters 2 and 3) is concerned with computer simulations of auditory and lexical acquisition. Chapter 2 investigates the induction of phonological features in a neural network that learns from auditory distributions and lexical information; I show that phonological features emerge robustly, regardless of the pace of lexical acquisition, and also when the learner makes mistakes. These results suggest that the phonological feature, which plays a central role in this dissertation, is a psychologically plausible primitive. Chapter 3 explores the parts that auditory distinctiveness and articulatory ease play in the evolution of various sound systems. The results show that all distributions eventually reach the same stable state, thus lending credence to the idea that typological tendencies may arise because of cultural evolution, and stressing the importance of a diachronic perspective.

The second part (Chapters 4 and 5) is dedicated to experimental evidence. In these two chapters, I present results from several tasks investigating the relation between different measures of complexity, inventory size, and learnability in the acquisition of phonological patterns. After all, the most feasible way to answer the question what determines the learnability of a system is to controlledly manipulate one or more properties of that system, and see if and how these manipulations affect the learning process in human learners.

In the third part (Chapters 6 and 7), I present two typological studies, and relate them to the experimental results from the previous chapters: are the systems

that humans in the lab find easy to learn also more frequently attested in the languages of the world? Are there any discrepancies between the two sources of information, and if so, how can we explain them? In Chapter 6, I again take a diachronic perspective by investigating a small number of sound changes; Chapter 7 presents an analysis of a database of 317 attested plosive systems in the UPSID database.

In Chapter 8, I further discuss some assumptions and findings, as well as their implications for further research: among other things, I discuss the problems of establishing complexity indices, correlated complexity measures, the limits of regularising behaviour, and issues with typological databases.

All data files – the script from the computer simulations, the tables with the experimental and typological data, and the R markdown files for the statistical analyses – are compiled in a project in the Open Science Framework online repository at <https://osf.io/jf8rx>.

*I'm wired to the world
That's how I know everything
I'm super brain
That's how they made me*

(Goldfrapp — Utopia)

Part I:
SIMULATIONS

2

Emergent phonological features in a symmetric neural network

In this chapter, I present results from computer simulations of phonetic, phonological and lexical acquisition in a neural network, in order to investigate the induction of phonological features using both bottom-up (auditory) and top-down (lexical) information.

2.1 Feature induction in human learners

Many generative frameworks, such as Optimality Theory (OT), make reference to innate phonological categories; such frameworks are therefore incompatible with the hypothesis that categories are emergent (but see Boersma, Escudero & Hayes 2003 for an OT model of category learning). For this reason, Benders (2013), Chládková (2014), and Boersma, Benders and Seinhorst (2020) used neural networks to model the emergence of phonological features: these features are not hard-wired into the model, but induced from probability distributions in the environment of the learner.

A neural network is a computational model of information processing. It consists of nodes that are organised into layers; these nodes are linked to each other by connections that can be either excitatory (if their weight is positive) or inhibitory (if their weight is negative). Connections may connect nodes within the same layer as well as nodes within different layers. Activity can flow through these connections, increasing or decreasing the activities of the nodes, much as it happens within the mammalian brain; two nodes that are connected with a strong connection exert a larger influence on each other than two nodes that are connected with a weak connection. The knowledge of the network is stored in the weights of its connections, and learning proceeds by adjusting these weights.

2.1.1 Distributional learning

An important mechanism in feature induction, and a crucial ingredient of the simulations in this chapter, is DISTRIBUTIONAL LEARNING, the process in which learners induce categories by being exposed to distributions of input tokens. This ability has repeatedly been attested in the lab (cf. §1.2.4), although a meta-analysis suggests that the effect of distributional learning remains somewhat unclear (Cristia 2018).

This chapter is an extended version of §§1–3 from Seinhorst, Boersma and Hamann (2019).

Auditory distributional learning studies are based on the rationale that in natural language, different meaning categories have different auditory realisations, or in other words, the probability distributions of their auditory correlates are distinct; this entails that, in order to successfully acquire her ambient language(s), the language learner would be wise to create a phonological category for every peak in the probability distribution of auditory tokens. Experiments are usually designed as follows. The participants are divided into two groups: one group is exposed to tokens from a monomodal probability distribution of auditory cues, that is, a distribution with one peak; the other group is exposed to tokens from a bimodal probability distribution, that is, a distribution with two peaks. After the learning phase, participants perform a discrimination task with the stimuli at the intersections of these distributions. Learners' ability to discriminate between the stimuli can be established with a behavioural task, or with a neurolinguistic measure, such as mismatch negativity responses (MMN). Both groups of participants have heard the stimuli at the intersections of the distributions equally often, so any differences between groups cannot be due to different input frequencies of the tested token; however, in the monomodal condition, the two auditory values belong to the same peak, whereas they belong to different peaks in the bimodal condition. Learners that have undergone the monomodal condition are not expected to hear the difference between the two auditory values (remember the phenomenon of categorical perception from §1.1.3), but learners that have undergone the bimodal condition are, exactly because the two groups have induced different numbers of categories. Figure 2.1 shows an example of monomodal and bimodal frequency distributions; arrows indicate the tokens at the intersections of the distributions, on which participants are tested.

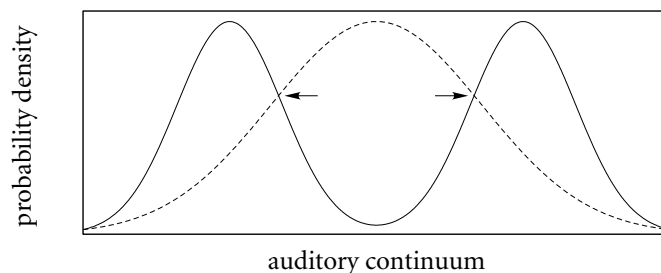


Figure 2.1. *Probability densities of auditory distributions in monomodal (dashed line) and bimodal (solid line) experimental conditions. Learners are tested on their ability to discriminate the tokens indicated by the arrows.*

If the auditory tokens in the experiment are accompanied by some sort of label (such as a word or an image), this is called “supervised learning”; if learners are merely exposed to auditory tokens, learning is “unsupervised”. Supervised learning may aid auditory distributional learning: Ter Schure (2016) finds that combinations of visual and auditory stimuli increase infants’ attention during learning.

2.2 A bidirectional model of feature induction

In this section I describe two crucial ingredients of the neural network model of feature induction: the theoretical framework, and the learning algorithm that is used to update the weights in the network. Both components are necessary to ensure the bidirectionality of the network, which turns out to be indispensable for the emergence of auditory dispersion in Chapter 3.

2.2.1 A bidirectional theoretical framework: the BiPhon model

The theoretical framework in which the computer simulations in this chapter are embedded is Boersma's model of bidirectional phonology and phonetics, or the BiPhon model (see Boersma 2011 for an overview). This model assumes (at least) two phonological representations, namely the underlying and surface forms. Additionally, it posits two phonetic representations: an auditory form, specifying all auditory events in the speech signal (e.g. formant frequencies, plosive release bursts, frication noise, etc.), and an articulatory form, specifying all muscle gestures in the speech utterance. Figure 2.2 shows the architecture of the model. The speaker starts from a meaning they wish to express, and moves down through the figure, ending up at an articulatory form; the listener travels up through the figure, deducing an intended meaning from auditory input.

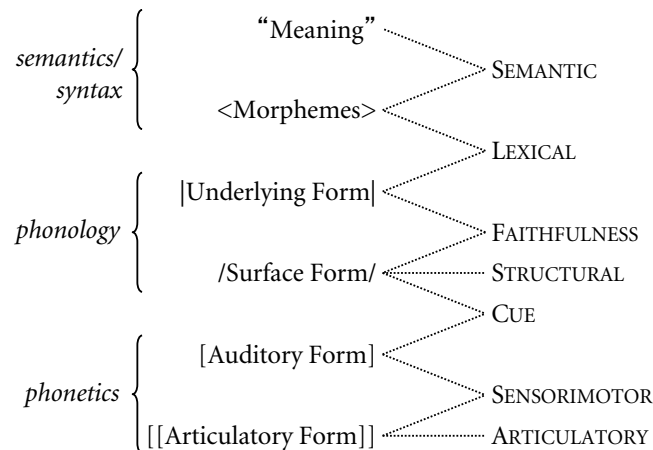


Figure 2.2. *The architecture of the BiPhon model (Boersma 2011: 33).*

The relations between these representations, as well as some representations themselves, are evaluated by different kinds of language-specific knowledge. The relation between the underlying form and the surface form is evaluated by knowledge of faithfulness (McCarthy & Prince 1995); an example of faithfulness knowledge would be that every segment in the underlying form must have a correspondent in the surface form. The phonotactic well-formedness of the surface form is constrained by

structural restrictions (Prince & Smolensky 1993/2004); an example of such a restriction would be that a syllable must have an onset. The relation between auditory events and phonological structure is evaluated by language-specific cue knowledge (Escudero & Boersma 2004; Boersma 2009); an example of cue knowledge would be that in English, a phonetically long vowel cues a following syllable-final voiced obstruent. The relation between auditory events and articulatory gestures is described by sensorimotor knowledge (Boersma 2006); an example of sensorimotor knowledge would be to raise the mandible to lower F_1 in vowels. The amount of effort that is required to produce an articulatory form is evaluated by articulatory knowledge, militating against gestures for which the speaker needs to expend much muscle effort (Boersma 1998; Kirchner 1998/2001).

In the BiPhon model, the speaker–listener uses their knowledge bidirectionally, that is, in production as well as perception: for instance, the same knowledge that prevents them from producing a fricative with a plosive release burst prevents them from perceiving a plosive release burst as belonging to a fricative. Bidirectionality is argued for by Smolensky (1996), Tesar (1997), and Tesar and Smolensky (2000); it is a crucial assumption in, for instance, Blutner (2000), Jäger (2003), and in exemplar models of speech perception and production (Pierrehumbert 2001; Wedel 2006).

2.2.2 Some properties of the neural network and the input

The computer simulations presented here follow Benders (2013), Chládková (2014), Boersma (2019), and Boersma, Benders and Seinhorst (2020) by using the architecture of the BiPhon model, with neural networks as a decision mechanism (“BiPhon-NN”). The simulations are run in the computer programme Praat (Boersma & Weenink 2018). In these simulations, the auditory representation (AudF) consists of two separate layers, each corresponding to an auditory continuum; these may be spectral centre of gravity in sibilants (CoG, expressed in e.g. ERB or Hertz), and periodicity, simplifyingly assuming that these are the single cues to a phonological sibilant contrast.⁴ The number of nodes in each AudF layer is set here at 48. The SF level consists of two layers as well, here consisting of six nodes each, connected to one of the auditory continua.

All AudF nodes are connected to all SF nodes that belong to the same continuum through excitatory cue connections. Each node within an SF layer is connected to every other node in that same layer by an inhibitory connection with a weight of -0.05 (this low number makes these connections difficult to see in the figures). These connections embody the concept of competitive learning (Grossberg 1976, 1987; Rumelhart & Zipser 1985): when a node gradually becomes activated, it simultaneously reduces the activities of the other nodes through these inhibitory connect-

⁴ If the auditory cues are commensurable, such as F_1 and F_2 in vowels, a single auditory layer suffices (Benders 2013; Chládková 2014; Boersma 2019).

ions. The weights of these connections do not change throughout the simulation. The network is symmetric: only a single connection exists between two nodes A and B, and the influence of node A on node B is equal to that of node B on node A.

A neural network in its initial state is shown in Figure 2.3. At this point, the network consists of only two levels of representation: the surface form /SF/ and the auditory form [AudF]. The excitatory cue connections that connect these levels have random, small weights at this point. They are drawn in black; stronger connections are indicated by thicker lines. The nodes have random activities at this point; more strongly activated nodes are drawn with larger circles inside the nodes.

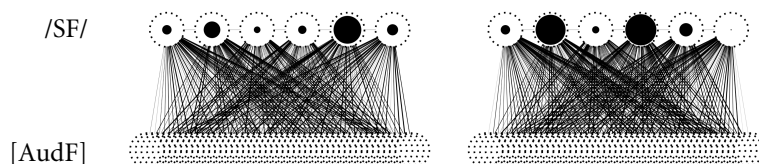


Figure 2.3. A neural network before the first learning step.

The network learns from distributions of auditory tokens. It is trained using artificial data, not natural stimuli. Figure 2.4 shows a possible probability distribution for a single continuum; the dotted lines in the figure indicate the probability distributions of the auditory tokens corresponding to two phonological categories, and the solid line indicates the pooled frequency distribution. This is the sum of the two dotted lines, and it constitutes the auditory environment of the learner. In this figure, the peaks of the distributions lie at 35% and 65% of the continuum, with the standard deviations equaling $1/14^{\text{th}}$ of the continuum.

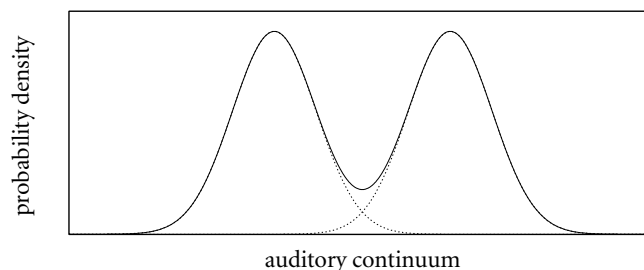


Figure 2.4. The input to the learner: frequency distributions of the lexical categories (dotted lines), and the pooled frequency distribution (solid line).

For each lexical category, the input probability of each of the 48 individual AudF nodes is computed from the continuous function in Figure 2.4. Following the assumption that the auditory continua are periodicity and spectral centre of gravity in sibilants, an inventory with binary contrasts on both continua is probably $\{s \int z\}$, as found in English, French, and many other languages.

2.2.3 A bidirectional learning algorithm: the inoutstar rule

The learning algorithm according to which the weights in the network are updated during learning is the INOUTSTAR learning rule (Boersma, Benders & Seinhorst 2020), which combines properties of the outstar (Grossberg 1969) and instar (Grossberg 1969, 1976; Rumelhart & Zipser 1985) algorithms. Both these algorithms are directional, meaning that an input layer and an output layer need to be defined to compute the amount of the weight update in a learning step. In outstar learning, the weight of a connection comes to reflect the probability that an output node is on given that an input node is on; in instar learning, the weight of a connection comes to reflect the probability that an input node is on given that an output node is on. The effect of directionality in learning can be illustrated by considering the probability distributions in Figure 2.4, specifically the edges of the auditory continuum. If we consider a token at the left edge of the continuum, we can be certain that this token corresponds to the phonological category that is realised as this leftmost peak; this means that if learning happens with AudF as the input and SF as the output, outstar learning will create strong connections between the left periphery of the continuum and the nodes of the corresponding category at SF. However, the shape of the auditory distribution shows that it is quite unlikely that this category will be realised as a token at the edge of the continuum, so instar learning will create weak connections between the nodes at SF and the left periphery of the continuum.

Formula (2.1) specifies the amount of a weight update in a learning step with the inoutstar rule (this is formula (20) in Boersma, Benders & Seinhorst 2020). In the formula, Δw_{ij} is the amount of the weight update of the connection between nodes i and j , η_w is the learning rate, a_i and a_j are the activities of nodes i and j , and w_{ij} is the weight of the connection between nodes i and j .

$$(2.1) \quad \Delta w_{ij} = \eta_w \left(a_i a_j - \frac{w_{ij}(a_i + a_j)}{2} \right)$$

The learning rule is insensitive to the direction of processing: in the formula, it does not matter whether node i or j is the input or output node, because $a_i a_j = a_j a_i$, $a_i + a_j = a_j + a_i$, and, because of the symmetry of the network, $w_{ij} = w_{ji}$. This insensitivity makes the rule suited for bidirectional use.

2.3 Emergent features

In this section, I investigate the emergence of phonological features under various circumstances. The learning process consists of two stages: the first stage involves feature induction through auditory distributional learning, without any involvement of the lexicon (cf. Benders 2013; Boersma, Benders & Seinhorst 2020); the second stage involves lexicon-supervised learning. The auditory environment to the network

in this section is the probability distribution from Figure 2.4, with peaks at 35% and 65% of both continua.

2.3.1 The first stage: distributional learning

In the first learning stage, the network is only exposed to an auditory distribution; the learner is not yet aware of any meanings that these sounds represent.

A learning step in the first stage proceeds as follows. For both auditory continua independently, an AudF node is drawn from the probability distribution: low-probability tokens will be selected less often than high-probability tokens. As an example, let us imagine that AudF node 25 has been selected in a given learning step. To this number, transmission noise is added (Boersma & Hamann 2008), representing noise that originates in the transmission channel itself, such as articulatory imprecisions on the part of the speaker, background noise, or noise in the listener's perception process. As a result, the input node may no longer be exactly 25; it is drawn from a Gaussian distribution with the original input node as its mean and a standard deviation of 5% of the auditory continuum (so in this case 2.4 nodes). This distribution is shown in Figure 2.5. Let us imagine that in this learning step, the resulting AudF node is 27.61 instead of 25; an arrow indicates the displacement.

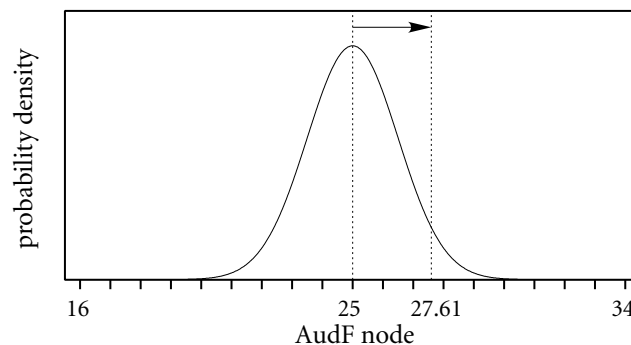


Figure 2.5. *The probability distribution of input nodes after transmission noise if node 25 was originally selected. In this example, the selected node after noise is 27.61.*

On the basilar membrane in the inner ear, this auditory token will excite those hair cells that are sensitive to the frequency of the token, as well as adjacent hair cells (Moore & Glasberg 1983). In the neural network, this is implemented as a bell-shaped activation pattern with a standard deviation of 1.5 nodes: in this example, the mean of this distribution is node 27.61. This activation pattern is computed for each individual AudF node, as shown in Figure 2.6: since node 27.61 is slightly closer to node 28 than to node 27, node 28 is activated slightly more strongly than node 27.

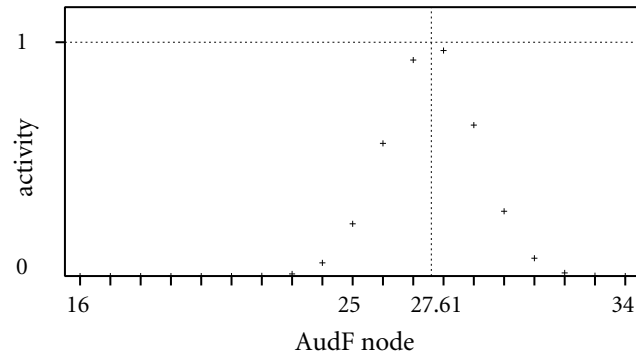


Figure 2.6. *The activities in the individual AudF nodes follow a bell-shaped excitation pattern with node 27.61 as its mean.*

If the addition of transmission noise causes the input node to fall outside the continuum, the learning step is terminated, and a new learning step begins.

At this point, the nodes at the auditory layers are clamped, meaning that their activities are fixed, and the activities are spread from the AudF nodes to the surface layer through the cue connections. This activation spreading happens in 100 time increments, allowing the activities in the network to reach an equilibrium (see §3 in Boersma, Benders & Seinhorst 2020). As a node at SF becomes active, it simultaneously deactivates other nodes within the same layer through the lateral inhibitory connections. The tenth learning step in this stage is shown in Figure 2.7.

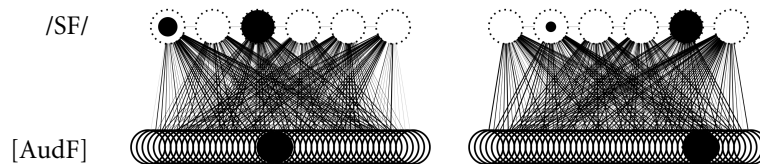


Figure 2.7. *The tenth step in the distributional learning stage.*

Once activity spreading is complete, the weights of the cue connections are updated. These steps are performed 8000 times, after which the first learning stage is complete. Figure 2.8 shows the same network at this point.

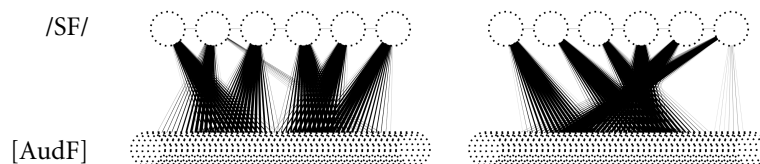


Figure 2.8. *The same network after 8000 learning steps.*

The figure shows that nodes at SF have become connected to certain parts of the auditory continua: on the leftmost SF, nodes 1, 2 and 3 are connected to the left side of the continuum, while nodes 4, 5 and 6 are connected to the right side. On the rightmost SF, nodes 1, 5 and 6 have specialised in the left side of the continuum, and nodes 2, 3 and 4 have become connected to the right side. This means that if we pace through the left auditory continuum, activate an AudF node and spread the activity up to SF, nodes 1, 2 and 3 will be activated for the left side of the continuum, and nodes 4, 5 and 6 will switch on for the right side. The network thus shows two kinds of behaviour, meaning that a phonological contrast has emerged. The network shows categorical behaviour on the right continuum too; nodes 1, 5 and 6 will become activated when an AudF node on the left side of this auditory continuum is activated, and nodes 2, 3 and 4 when the right half of the continuum is activated. The phonological categories refer to contiguous parts of the auditory continuum; because of the Gaussian shape of the activations at AudF during learning, every learning step strengthens the connections between not just one AudF node and the SF layer, but between a small group of AudF nodes and the SF layer.

SF node 2 on the leftmost continuum and SF node 6 on the rightmost continuum are still somewhat connected to the peripheries of their corresponding auditory continua: these cue connections have retained their random, initial weights because very few input tokens occurred at those peripheries, and no learning has taken place there. When such peripheral tokens are activated, and activity is spread upwards to the surface layer, this layer will display very little activity; such behaviour can be interpreted as insecurity on how to categorise such peripheral sounds.

2.3.2 The second stage: lexicon-driven learning

When the second learning stage begins, the learner has become aware that different concepts exist in the external world. This awareness has been implemented in the network as the creation of a third layer of nodes, representing the lexical categories; these concepts are distinguished by different sounds, and in this second stage the network will acquire the mappings between sound and meaning. Every lexical category consists of a group of four nodes; each node in each category is connected to all SF nodes through excitatory connections with small, random weights.

In the theoretical model from Figure 2.2 (p. 23), the lexical layer represents different meanings in the external world; however, I assume that different meanings also correspond to different phonemic categories, making the added layer a mix between a semantic and an underlying representation. This is a simplification: just like the surface form, an underlying form should be thought of as an emergent representation, which the learner constructs by positing a representation that abstracts away from morphophonological alternations. No such alternations have been implemented in the simulations in this dissertation; Chládková (2014) shows how alternations may facilitate the induction of features in BiPhon-NN.

In the second learning stage, the input to the network consists of sound–meaning pairs, as in Chládková (2014): the learner hears a sound, and knows to which lexical category this sound belongs. The tenth learning step in the second stage, so the 8010th learning step overall, is shown in Figure 2.9:

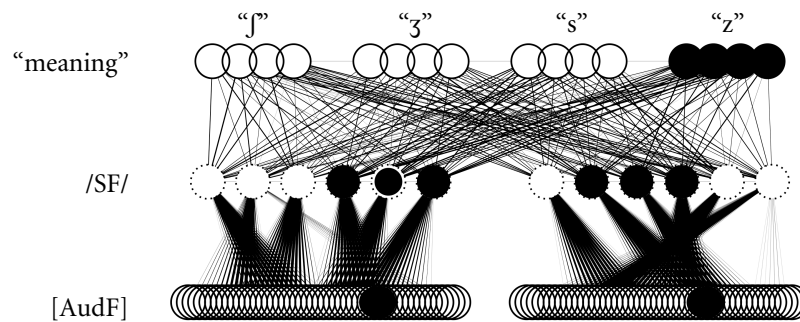


Figure 2.9. *The tenth learning step in the second learning stage.*

A learning step proceeds as follows. A lexical category is selected at random, and the nodes that belong to the selected category are activated. The auditory tokens are now drawn from the individual distributions in Figure 2.4 (drawn with dotted lines), no longer from the pooled distribution (drawn with a solid line). Again, transmission noise is added to the selected token, and a bell-shaped activity pattern is applied to AudF. All nodes at the lexical and AudF layers are clamped, while the nodes at SF remain unclamped. Activity is allowed to spread through the network in 100 time increments; since the activation spreads through SF, the lexical learning process is necessarily mediated by the features that emerged in the first learning stage. Once the activity spreading is done, the weights of the connection in the network are updated.

These steps are carried out 8000 times; the learning process is then done. The same network after a total of 16000 learning steps is shown in Figure 2.10.

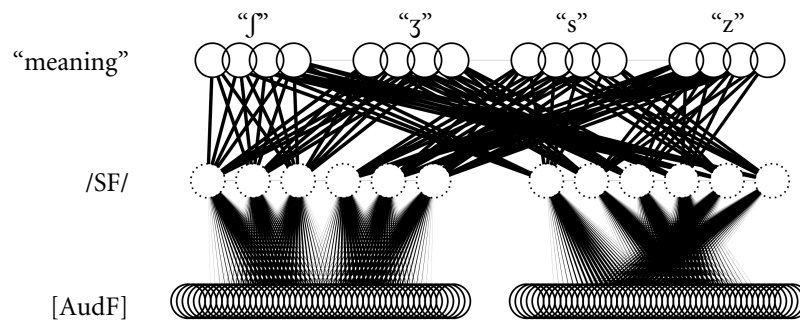


Figure 2.10. *The same network after a total of 16000 learning steps.*

2.3.3 The division of categories across nodes

The six nodes in each SF layer are usually divided evenly between the two categories, that is, each feature value is usually represented by three SF nodes; in a minority of cases, the division is 4–2 or 2–4, and in a few cases a single node has not been recruited, yielding a 3–2 or 2–3 division. Figure 2.11 shows the divisions of phonological categories across SF nodes for ten runs of simulations. Black and white nodes belong to contrasting categories. In these ten runs, all nodes were recruited by a category, with a 3–3 division occurring 15 out of 20 times, and a 4–2 division the remaining 5 times. The figure also shows that the distribution of nodes across categories differs randomly between runs; the SF–AudF interface looks different in every simulation, but categorical behaviour emerges every time, on both continua.

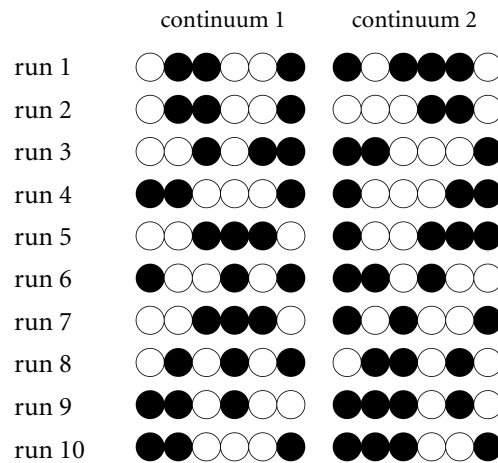


Figure 2.11. *Distributions of phonological categories across SF nodes over ten runs: contrasting feature values are drawn in contrasting colours. In every run, categorical behaviour emerges on both continua.*

2.4 Auditory versus lexical learning

The observant reader may have noticed that after lexicon-driven learning, the SF–AudF interface looks somewhat different than it did after distributional learning alone. Comparing the network from §2.3 after the distributional learning stage (Figure 2.8, p. 28) to the same network after lexical learning (Figure 2.10, p. 30), we can see two differences in the interface between AudF and SF. Firstly, the peripheral connections that still remained after distributional learning have disappeared after 8000 additional pieces of input, probably because tokens from these peripheries, which have a low probability of being selected in a learning step, have now appeared in the input. Secondly, the lexical learning stage has caused all SF nodes that belong

to a phonological category to become connected to AudF in the exact same way, which was not necessarily true after distributional learning alone. Consider, for instance, node 2 on the leftmost SF in Figure 2.8, which seems to be the only node to have become specialised in the leftmost five or ten AudF nodes at this point; top-down information from the lexicon has taught the network that those AudF nodes belong to a single feature value, distributed across multiple SF nodes, so all these SF nodes have become connected to the auditory continuum in identical ways.

2.4.1 Testing ground: a bimodal distribution

The effect of top-down information on category creation can be investigated by training the network on a language in which the number of peaks in the auditory distribution is not equal to the number of lexical categories, such as a language in which the auditory realisations of two or more lexical categories have identical probability distributions, or a language with one or more bimodally distributed categories. Figure 2.12 shows the probability distribution of a system with two lexical categories, contrasting in their spectral centre of gravity: one category is drawn with a solid line, the other with a dashed line. The dashed category is unimodal, with its peak at 50% of the auditory continuum; the other category is bimodal, with peaks at 20% and 80% of the continuum. For all three peaks, the standard deviation of the distribution is $1/14^{\text{th}}$ of the continuum, as before.

Assuming that all sounds in this inventory are voiceless, this three-way CoG contrast should likely be transcribed as $\{\text{ʃ} \text{ ɛ} \text{ s}\}$: these are the sibilants found in Chuvash, Mandarin Chinese, and Polish, although all three categories have phonemic status in those languages (and Polish also has their voiced counterparts $\{\text{ʒ} \text{ z} \text{ z}\}$). The left peak of the bimodal category then corresponds to $[\text{ʃ}]$, its right peak to $[\text{s}]$, and the monomodal category to $[\text{ɛ}]$.

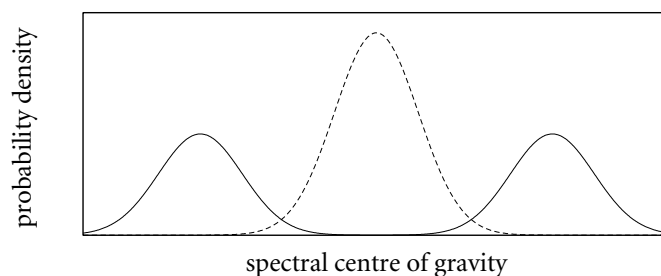


Figure 2.12. *The probability distributions on the CoG continuum of lexical categories in a language with one bimodal category.*

Because the pooled probability distribution is trimodal, that is, the auditory environment of the learner has three peaks, the learner will infer the existence of three categories during distributional learning. A network after 8000 learning steps is

shown in Figure 2.13; because I assume all segments to be voiceless, I disregard the periodicity continuum for now. SF nodes 4 and 6 are connected to the left part of the CoG continuum, nodes 1 and 3 to its centre, and nodes 2 and 5 to the right part.

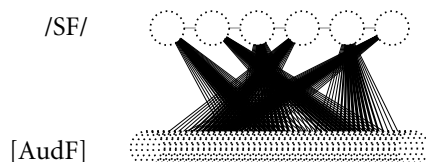


Figure 2.13. A neural network after distributional learning of a language with one bimodal category.

The same network after lexical learning is shown in Figure 2.14:

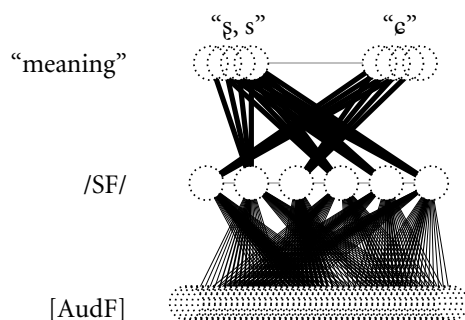


Figure 2.14. The same neural network after lexicon-driven learning of a language with one bimodal category.

The network has now learned that the outer peaks on the auditory continuum belong to the same meaning category, and as a result nodes 2, 4, 5 and 6 have all become connected to both edges of the continuum. The categories that emerged after distributional learning, which were audition-based, now reflect lexical information too. Such bimodal distributions do not seem to occur in natural languages, and therefore the representation in Figure 2.14 is probably hypothetical; Boersma and Hamann (2008) provide an explicit formalisation of the demise of a bimodally distributed category within Optimality Theory, and I do the same in neural networks in §3.5 (p. 60).

This reorganisation of the SF layer after lexicon-driven learning happens robustly, as visualised in Figure 2.15 (next page), which compares the divisions of categories across SF nodes on the CoG continuum before and after lexical learning for ten runs of simulations. Different categories are drawn in black, white and grey. In two cases, an SF node was not recruited; these nodes are drawn with a dotted edge. After distributional learning alone, a ternary contrast emerged in every run;

however, the knowledge that the peripheral peaks belong to the same lexical category caused the nodes that had become specialised in those peaks (i.e. the grey and black nodes in the left part of Figure 2.15) to become subsumed under one bigger category (drawn in black).

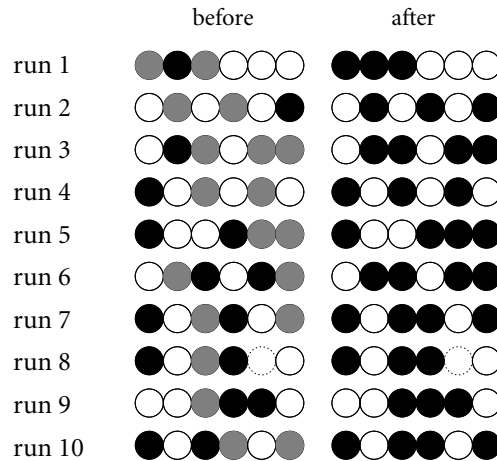


Figure 2.15. Distributions of phonological categories across SF nodes over ten runs, before and after lexical learning. The three-way contrast is replaced by a binary contrast.

2.5 An emergent lexicon

In the previous two sections, there was a clear cut-off between the two stages of learning: in the first stage, which ended after the 8000th learning step, the lexicon did not exist, while in the second stage, starting in the 8001st learning step, it was in place in a single stroke, ready to supervise the learning process. Such a sudden transition between stages, and such an abrupt inception of the lexicon in the language-acquiring child, may not be very realistic: the lexicon probably unfolds in a more gradual and less smooth fashion. In this section, I explore the behaviour of the model when the lexicon becomes active gradually. In these simulations, the network has a lexical layer from the onset of learning; the activities at this layer are small at first, but they increase throughout the learning process. More specifically, the activities at the lexical layer develop according to a logistic curve as a function of the learning step. The formula of the logistic functions explored here is given as (2.2):

$$(2.2) \quad a(s) = \frac{1}{1 + e^{k(s-m)}}$$

In this formula, $a(s)$ is the activity in each node of the selected lexical category in learning step s , e is the base of the natural logarithm, k is the steepness of the curve, and m is the learning step in which the logistic curve reaches its inflection point, so

$a(s) = 0.5$. Figure 2.16 shows three logistic curves with different degrees of steepness, that is, with different values of k : for the dotted line, $k = -0.005$; for the solid line, $k = -0.001$; for the dashed line: $k = -0.0005$). In all three curves, m lies exactly halfway between the first and last learning step.

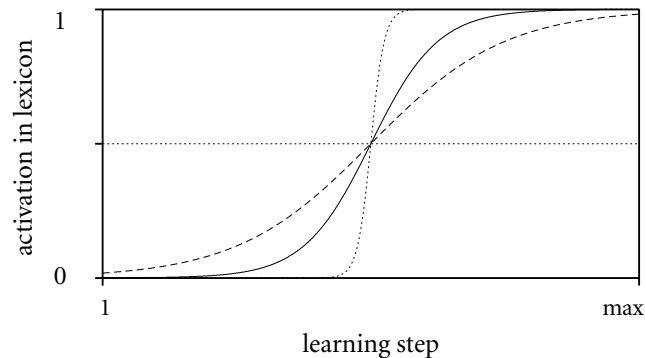


Figure 2.16. Activations at the lexical layer as a function of learning step, for three different values of k .

More negative values of k yield steeper curves, approximating the abrupt transition between prelexical and lexical learning from the two-stage model from §§2.3–4 more closely.

2.5.1 Different paces of lexical acquisition

Even when the lexical layer is present at the onset of learning, the activities at this layer are initially fairly small, so auditory distributional learning is still the predominant strategy at this point; the connections between lexical and phonological categories only become strengthened later in the acquisition process, at which point lexical information may influence the AudF–SF interface as it did in §2.4. Therefore, an emergent lexicon does not seem to have an influence on phonological category creation at SF: categorical behaviour emerges in every run, irrespective of the steepness of the curve, and the division of categories across SF nodes is identical to the division found in §2.3.3. This division can be seen in Figure 2.17, for ten runs, for different degrees of steepness. Again, as in the two-stage simulations from §2.3, this division is usually 3–3, namely in 43 out of 60 cases; in ten cases, it is 4–2. In seven cases, one SF node has not been recruited by any category, resulting in a 3–2 division.

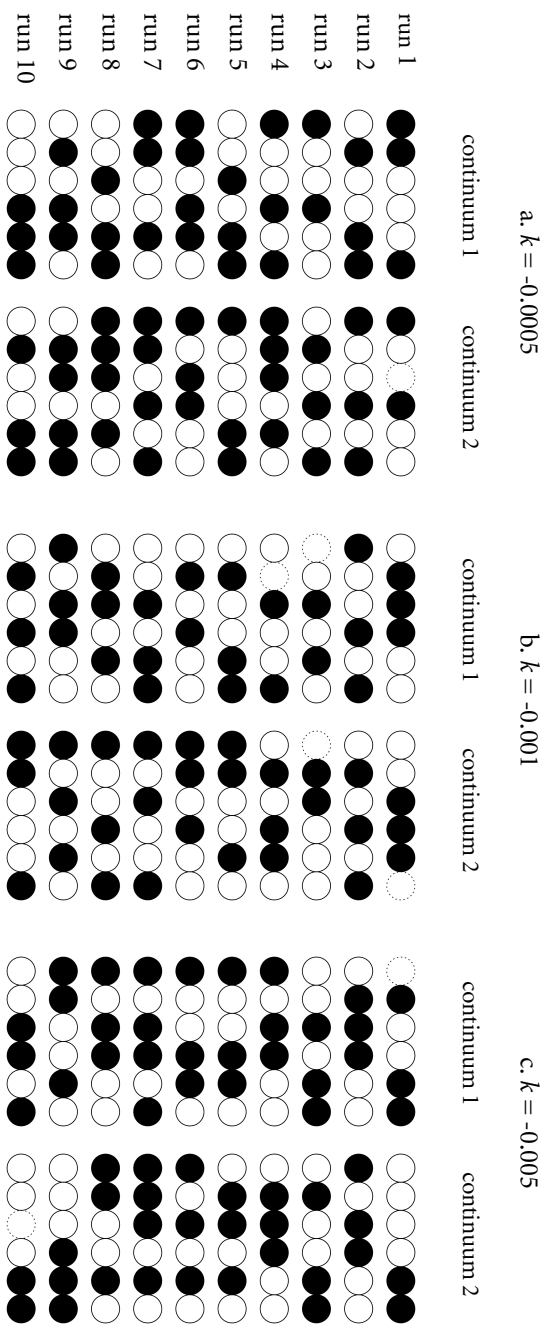


Figure 2.17. Distributions of phonological categories across SF nodes over ten runs, for different values of k .

2.6 Robustness against variation in word learning

Although the logistic curves from §2.5 are probably more realistic than the two-stage model from §2.3, it seems unlikely that word learning in reality proceeds as smoothly and incrementally as these logistic curves suggest. In this section, I test the robustness of the neural network against variation in the acquisition process, introducing a number of sources of variation. Although many such sources are conceivable, I investigate only four here: activation of an unintended meaning (§2.6.1); reduced activation in the lexicon (§2.6.2); delayed lexical development (§2.6.3); and mishearings (§2.6.4).

2.6.1 Mismatches between meaning and sound

In this subsection, I explore mismatches between intended meaning and sound, when the learner associates a sound with a lexical category that is not associated with that sound in the input language. I make the probability that such an error occurs in a learning step dependent on the activations at the lexical level in that learning step: this probability equals 1 minus the lexical activation, so it is large at the onset of learning and gradually decreases asymptotically to zero. For the three logistic curves in Figure 2.16 (p. 35), the probabilities that an error occurs look as follows:

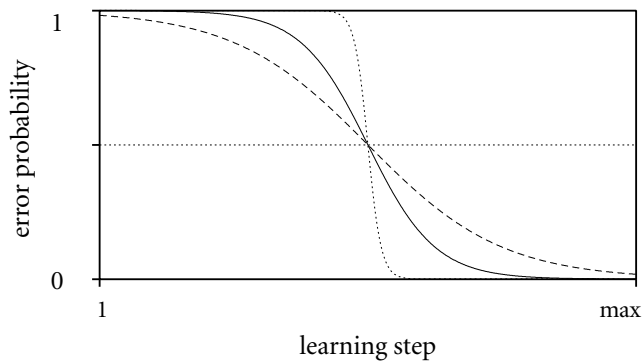


Figure 2.18. Error probabilities as a function of learning step, for different values of k . Dotted line: $k = -0.005$; solid line: $k = -0.001$; dashed line: $k = -0.0005$.

If an error indeed occurs in a learning step, a lexical category is selected at random. This can actually also be the intended category, so that the chance that a mismatch between sound and meaning indeed occurs in a language with n lexical categories is $(n-1)$ in n . The nodes of the selected lexical category are then activated by the amount defined by the logistic activity function.

While lexical errors are initially frequent, at the same time the activities at the meaning layer are still small, and as a result audition-based features emerge at SF.

The word learning errors will cause these features to become connected to all lexical categories to a certain extent. After all, the learner does not receive any feedback and therefore does not detect any errors; as a result, a learning step with an error strengthens connections between active nodes exactly the same way a learning step without an error does. Figure 2.19 shows a network after 10000 learning steps ($k = -0.0005$). For example, if we look at node 1 on the leftmost SF layer, we see that this node is most strongly connected to the right half of the auditory continuum, and to the “s” and “z” lexical categories; however, some weaker connections exist to the left half of the auditory continuum too, and to the first and second lexical categories, due to erroneous sound–meaning associations.

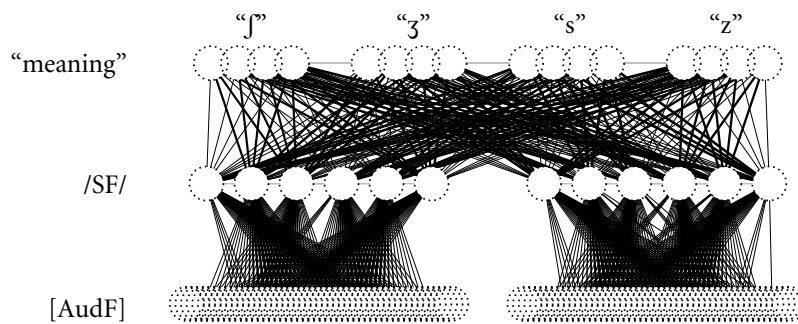


Figure 2.19. A network after 10000 learning steps (with sound–meaning mismatches).

As the error probability decreases and the activations at the meaning layer increase, the faulty connections between the meaning and SF layers, and between the SF and AudF layers, gradually disappear. Figure 2.20 shows the same network from Figure 2.19 after 6000 additional learning steps. The connections between the meaning and SF layers are much stronger than they were in the previous figure; the learner’s knowledge of the relation between sound and meaning has consolidated.

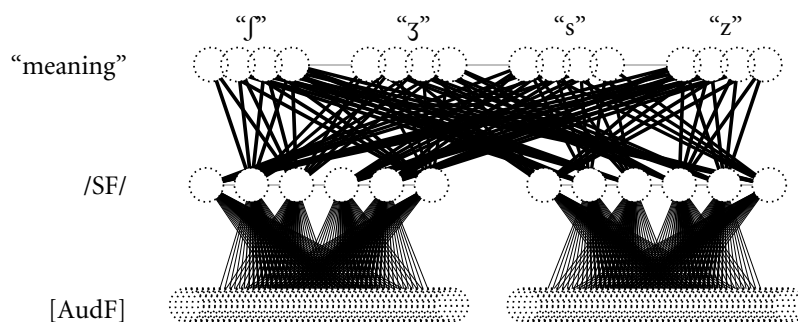


Figure 2.20. The same network after 16000 learning steps.

Eventually, again, categorical behaviour emerges at SF, even if word learning proceeds imperfectly: the nodes are divided across categories in the familiar 3–3, 3–2 and 4–2 distributions.

2.6.2 Reduced activities in the lexicon

In this subsection, I implement errors by subtracting a certain amount from the activities in the nodes of the intended lexical category, effectively hampering the associations between meaning and sound. The subtracted amount decreases with each learning step: it is drawn from a normal distribution whose mean and standard deviation decrease from 0.5 in the first learning step to 0 in the last learning step. The logistic curve for $k = -0.0005$ from Figure 2.16 is repeated here as Figure 2.21a. Figure 2.21b plots the probability distribution of the subtracted activities as a function of learning step: the average is drawn in black, the average ± 1 standard deviation is drawn in grey. The resulting activities are plotted in Figure 2.21c.

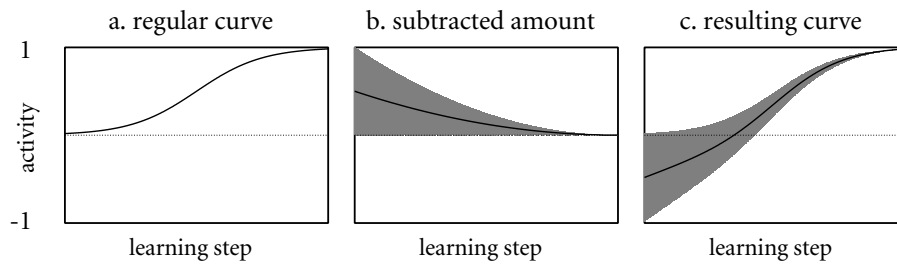


Figure 2.21. Probability distributions of the reduced activities in the meaning layer as a function of learning step.

Figure 2.22 shows a network ($k = -0.0005$) after 2000 learning steps with reduced activities at the meaning layer.

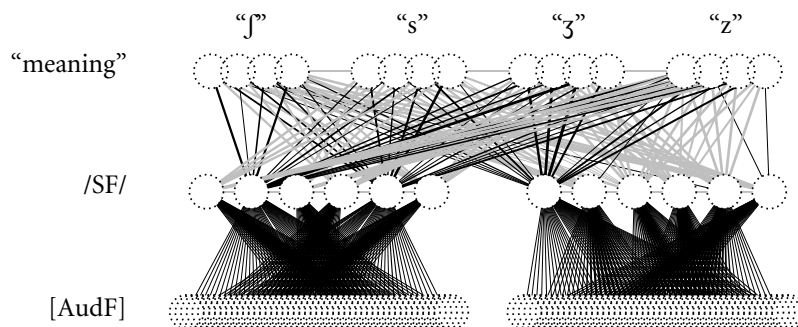


Figure 2.22. A network with reduced activities at the meaning layer after 2000 learning steps.

Since the average activities are negative for the first few thousand learning steps, inhibitory connections have emerged between the lexical and phonological layers, which could perhaps be interpreted as a strong discouragement of any association between meaning and sound. The only excitatory connections in this interface are connected to SF nodes that have not specialised in any part of the auditory continua.

Meanwhile, some categorical behaviour seems to emerge in the phonology–phonetics interface: on the left auditory continuum, SF nodes 1 and 3 seem to become connected to the rightmost peak, and SF nodes 4 and 6 to the leftmost peak. On the right auditory continuum, SF nodes 2 and 6 seem to specialise in the leftmost region of the continuum, and nodes 3, 4, and 5 in the rightmost region.

Figure 2.23 shows the same network after 16000 learning steps.

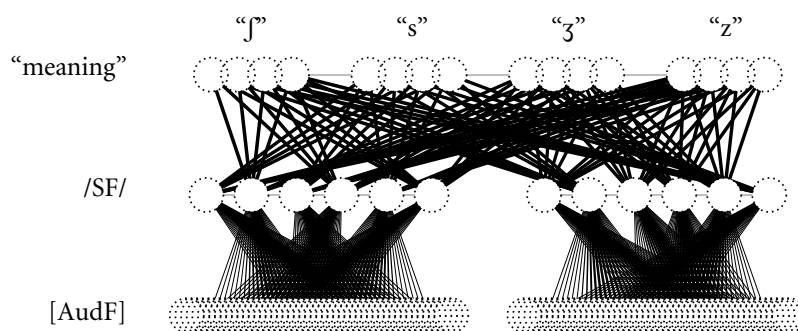


Figure 2.23. *The same network after 16,000 learning steps.*

The inhibitory connections between the lexical and phonological layers have disappeared; the network looks exactly like the other mature networks seen earlier in this chapter. Also, the incipient categories at the SF layers have remained fairly stable through the reorganisation of the semantics–phonology interface: the main differences are that SF nodes 2 and 5 on the leftmost layer have now been recruited by contrasting categories, and node 1 on the rightmost layer has been subsumed under the 3–4–5 category.

2.6.3 Delayed lexical development

So far, the inflection point of all logistic curves (where $a(s) = 0.5$) was exactly halfway between the first and last learning steps: in terms of formula (2.2), $m = 8000.5$ in all simulations so far. In this subsection, I model a learner whose lexicon starts developing only at a later point in time, implying a higher value of m in formula (2.2). The logistic function describes the activities at the lexical layer only, and it does not affect the activities at AudF; therefore, the auditory distributional learning still proceeds as usual. Let us investigate a network in which the lexical activations develop according to the function in Figure 2.24, where $k = -0.0005$ and $m = 16000$.

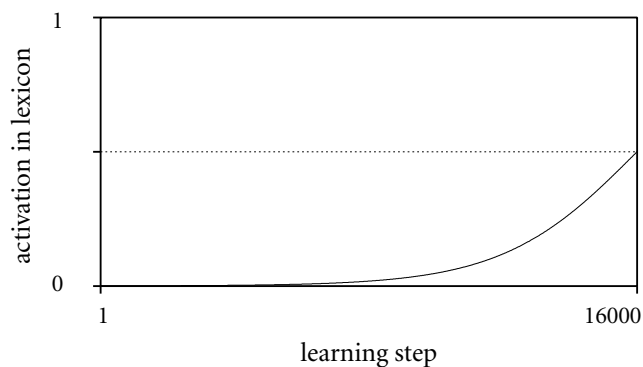


Figure 2.24. *Activations at the lexical layer for $m = 16000$ and $k = -0.0005$.*

In a simulation of 16000 learning steps using the two-stage model (§§2.3–4), the integral of the formula that expresses activations in the lexicon as a function of learning step is exactly 8000;⁵ in each of the three curves from Figure 2.16, this sum is 7999.5 (the area under the curves), in Figure 2.24, this area is only 1385.6. Nevertheless, after 16000 learning steps, the semantics–phonology interface seems to be fairly well-organised:

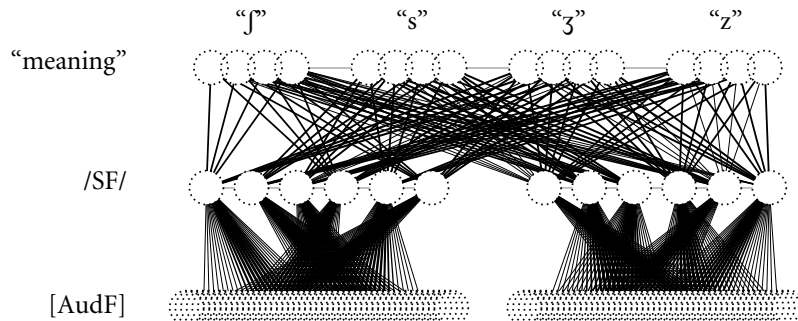


Figure 2.25. *A neural network in which lexical development was delayed.*

Although the connections between the lexical and phonological layers are less strong than in other mature networks, stable lexical knowledge still seems to have developed, in spite of the delay: comparing this model with the one with the sound–meaning mismatches from §2.6.1, the network in Figure 2.25 looks more like the mature network in Figure 2.20 (p. 38) than like the network in Figure 2.19.

⁵ These numbers may turn out a little bit higher or lower in practice; the lexical categories are selected at random, and if the transmission noise scatters an input token in a learning step to a value below node 1 or above node 48, the token is abolished and a new learning step begins.

2.6.4 Mishearings

Of course, errors need not occur at the lexical level exclusively; they may also occur at the auditory level. Imagine that a learner for some reason grows up in an extremely noisy environment, causing them to “mishear” their input. Remember from §2.3.1, more specifically Figure 2.5 (p. 27), that normally distributed transmission noise is added to the selected token in every learning step. In order to facilitate mishearings, I increase the standard deviation of the noise distribution by a factor 10, so that it equals 50% of the auditory continuum instead of 5%: in our network with 48 nodes per AudF continuum, the transmission noise will now be drawn from a normal distribution $\mathcal{N}(0, 24)$ instead of $\mathcal{N}(0, 2.4)$. As in §2.6.1, the probability that a token is misheard equals 1 minus the activation in the lexicon, and therefore mishearings become less likely over time.

The huge standard deviation of the transmission noise causes the auditory input to differ greatly between learners, and therefore their representations at SF differ as well. In many individual learners, it is in fact difficult to say exactly how many categories emerge there, because individual nodes tend to become connected to overlapping regions of the auditory continua. Figure 2.26 compares the total activity at SF as a function of AudF node between networks without mishearings (the left figure) and networks with mishearings (the right figure); these values are the summed activities in the six SF nodes belonging to an auditory continuum when the activity of an AudF node is set at 1 and the activity is spread upwards to SF. All activities were measured after 8000 steps and averaged over ten runs of simulations.

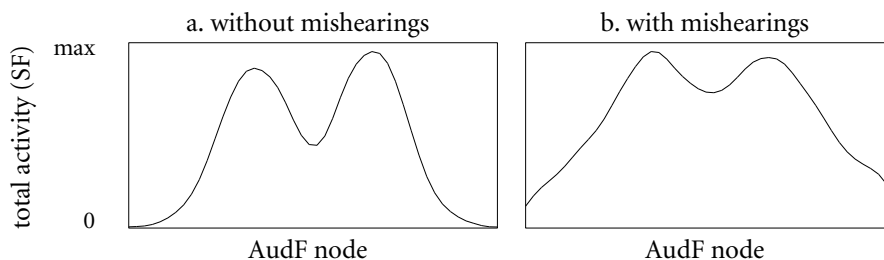


Figure 2.26. Averaged summed activities in networks with and without mishearings.

Two distinct peaks are seen in the left figure. The summed activities at SF are highest when AudF nodes at 35% and 65% of the continuum are active, where the peaks in the input probability distributions lie; the activities go towards zero at the edges of the continuum. The different heights of the peaks are due to unequal numbers of SF nodes associated with these peaks, resulting in different degrees of activation. In the right figure, in spite of the variability between individual networks, there are still two peaks in the averaged activities, also corresponding to the peaks in the input probability distribution. However, the peaks are much wider, and the valley between them

is much shallower than in the networks without mishearings: the valley depths are at 53 and 23% of the maximum activity, respectively. This suggests that the categories in the networks with mishearings are less clear-cut, displaying more overlap (cf. §5.9.1 in Boersma, Benders & Seinhorst 2020).

Figure 2.27 shows a network that combines mishearings with unintended meaning–sound correspondences, after 12000 learning steps. Although the semantics–phonology interface already seems to have become quite stable at this point, with only strong connections between both levels, the SF nodes are connected to unusually large regions at AudF, as a result of the scattering of auditory tokens.

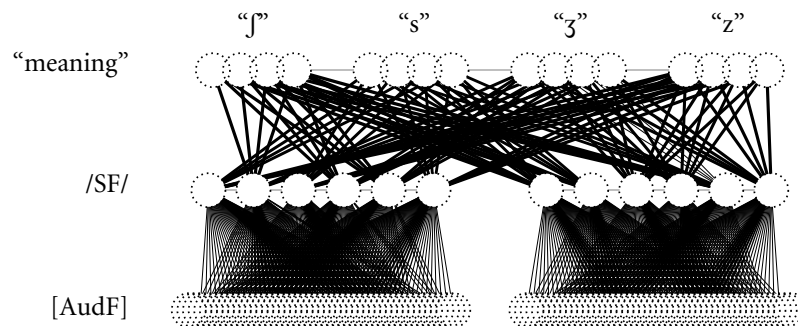


Figure 2.27. A network with mishearings as well as unintended meanings after 12000 learning steps.

After another 4000 learning steps, the regions in which the SF nodes have specialised have narrowed as the probability of mishearings has decreased. The network again resembles the mature networks seen earlier in this chapter, with the familiar 3–3 division of nodes across categories on both auditory continua; the network requires some more input tokens to get rid of the few remaining stray cue connections. These results suggest that stable categories emerge even with large variance in the auditory input, and that speaker normalisation is unproblematic when the lexicon is in place.

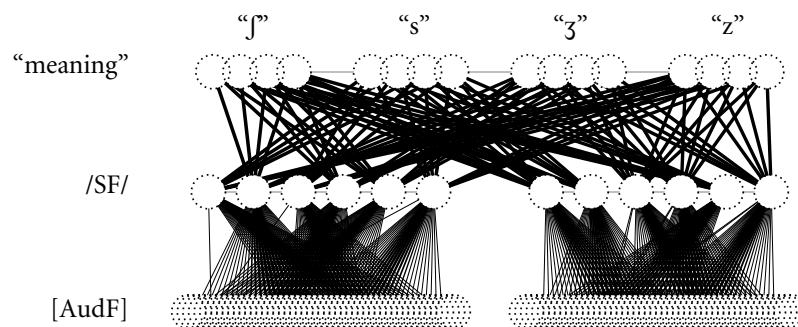


Figure 2.28. The same network after 16000 learning steps.

The results presented in this section suggest that the learning process is robust against variation in word learning, whether this variation occurs in the semantics or in the phonetics; the network even reliably induces phonological categories when errors occur simultaneously and at different levels of representation, as in the network from Figures 2.27–28, and when lexical acquisition is delayed, as in §2.6.3. These irregularities during the learning process are eventually resolved: as long as the probability that an error occurs decreases over time, the network eventually converges to a state with stable phonological categories at SF.

2.7 Phonemes and features (again)

Remember from §1.1 (p. 3) that a phoneme is traditionally considered a bundle of phonological features: phonemes and features are in a hierarchical relation. However, in Chládková’s (2014: 95) simulations of the acquisition of vowel systems, phonemes and features would emerge concurrently in a single surface representation: that is, nodes at SF could represent both feature values and specific phonemes. The neural networks in this chapter, on the other hand, induced only features. There are (at least) two possible reasons for this. Firstly, the architecture of the network maintained the separation of the auditory continua at SF, making the induction of phonemes in the phonology less likely; and secondly, all input languages were neatly structured in terms of features.

Earlier in this chapter (in §2.2.2, p. 24), I decided that the auditory continua could represent periodicity and spectral centre of gravity (CoG) in sibilants; an inventory that has a binary contrast on both continua, as used throughout this chapter, is likely something like {ʃ s ʒ z} (p. 25), and a language with a three-way contrast on the CoG dimension and binary voicing contrast is probably {ʃ ɛ s z z z} (§2.4, p. 32). Such systems are readily representable in terms of features, and these features are well-motivated, because there are always two or three segments within the inventory that share a feature value. By contrast, a system like {ʃ ɛ s z}, with three voiceless sibilants and only a single voiced category, displays two gaps, because it lacks {z z} even though their constituent feature values exist in the language. In this system, [z] is the only voiced phoneme, whereas the [voiceless] feature is shared by three segments: therefore, a feature-based representation does not seem to be much more parsimonious than a mixed representation, which might look something like “[voiceless] + [z]”.

In this section I test the behaviour of the network when it is trained on three inventories with different numbers of categories: {ʃ ɛ s z}, {ʃ ɛ s z z z}, and {ʃ ɛ s z z z z}. I compare two layouts of the model: the familiar layout with two phonological surface layers (§2.7.1), and a layout with a single phonological surface layer (§2.7.2) consisting of six nodes that are connected to both auditory continua.

2.7.1 First layout: two separate surface layers

This layout is the exact same one that I have been using so far, with two surface layers each connected to one auditory continuum. Figure 2.29 shows a network after 16000 learning steps with an { ξ e s z} inventory. On the left SF, a ternary CoG contrast has emerged: nodes 1 and 2 are connected to both retroflexes, nodes 3 and 6 to the alveopalatal category, and nodes 4 and 5 to the alveolar. The right SF encodes a binary voicing contrast: nodes 1, 4, 5 and 6 encode voicelessness, and nodes 2 and 3 represent voicing. As in all the previous simulations in this chapter, features have emerged.

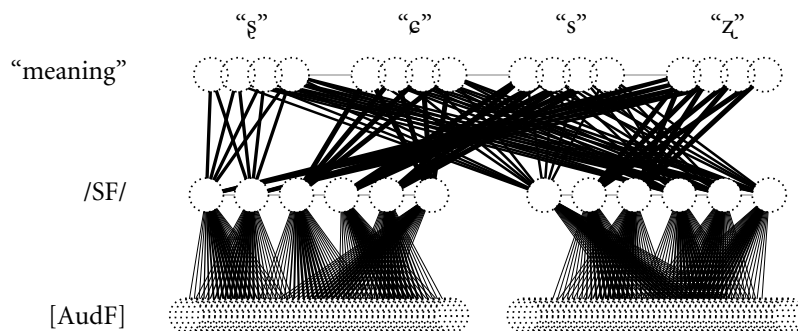


Figure 2.29. A network with two surface layers after 16000 learning steps.

The network shows the same behaviour for the two larger inventories; the categories that emerge at the SF layers are features only, never phonemes.

2.7.2 Second layout: a single surface layer

Figure 2.30 shows a neural network with a single SF layer that is connected to both auditory continua. The weights of the inhibitory connections between SF nodes is fixed at -0.1, instead of -0.05 in earlier simulations.

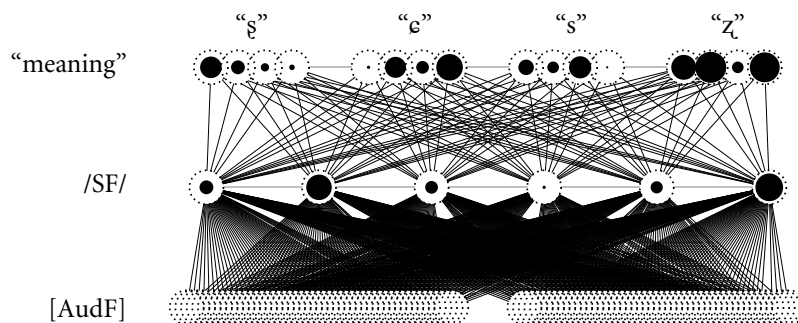


Figure 2.30. A neural network with a single surface layer.

Because this layout facilitates the integration of auditory information in the phonology, we might expect phonemes to emerge here. The network after 16000 input tokens from an $\{\text{ʂ ɛ s z}\}$ inventory looks as follows:

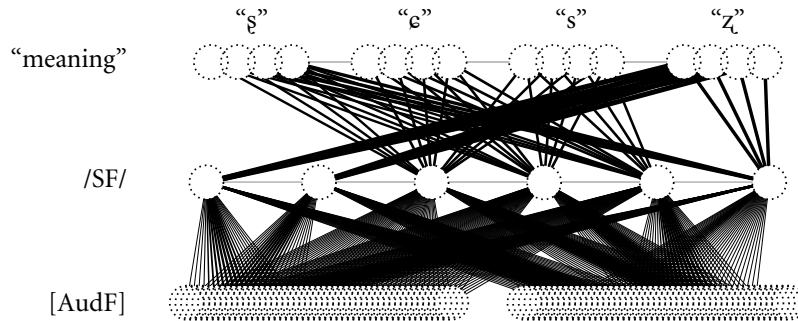


Figure 2.31. *The same network after 16000 learning steps.*

Indeed, the surface representation seems to contain both features and phonemes: nodes 1 and 2 are connected to the “z” category, nodes 3, 4 and 5 are connected to the voiceless segments, and node 6 represents both retroflex segments. Because of the layout of the network, all SF nodes are connected to both continua; so while SF node 6, for instance, represents only the retroflexes on the CoG continuum (and may therefore be interpreted as a feature like [retroflex]), it is also connected to both the voiced and voiceless areas on the voicing continuum at AudF. This happens because whenever the retroflex area on the CoG continuum at AudF is active in a learning step, either the voiced or voiceless area on the periodicity continuum at AudF is active simultaneously as well; both nodes will activate one or more nodes at SF, whose connections with all active nodes at AudF are subsequently strengthened. Similarly, the [voiceless] nodes at SF do not differentiate between the three possible CoG values: whenever the voiceless region at AudF is active, simultaneously one of the three CoG regions is active too, and therefore the [voiceless] nodes at SF will become connected to all three peaks on the CoG continuum at AudF.

Only the retroflexes are represented individually, but not the alveopalatal and the alveolar. This is probably due to the fact that retroflexes occur more often in the input than alveopalatals or alveolars: half of the input tokens are retroflexes.

In the two larger inventories, $\{\text{ʂ ɛ s z z}\}$, and $\{\text{ʂ ɛ s z z z}\}$, the network always represents the voicing contrast, and in a few cases a single CoG value as well. Again, this is probably due to frequency differences in the input between both features: after all, the voicing contrast is binary, so that an $\{\text{ʂ ɛ s z z z}\}$ learner is fed some 8,000 voiceless tokens, and some 8000 voiced ones. Because the CoG contrast is ternary, the same learner is only fed some 5333 tokens of each CoG value; the network apparently encodes the more frequent distinction. This issue is not resolved by adding more nodes to the SF layer.

The question begs itself whether SF nodes 1 and 6 in Figure 2.31 represent the “z” category, or perhaps just the feature [voiced]: after all, /z/ is the only segment with this property. Since SF nodes 1 and 2 are connected to both AudF continua, when we activate these nodes and spread the activities, both auditory properties of [z] become active too, suggesting that the SF nodes represent the individual phoneme; on the other hand, as explained earlier, an SF node that represents [voiced] cannot be active during learning without the retroflex areas on the CoG continuum being active at the same time, so this SF node inevitably becomes connected to these areas of the auditory continuum. The answer to the question, then, seems to be that we cannot always determine whether a node represents a feature value or a phoneme: it may do both at the same time. This seems to be a property of the inventory itself, rather than of the neural network.

The layout with the two individual SF layers, used throughout this chapter with the exception of the current subsection, best models the emergence of features, since it reliably represents each individual feature value in the inventory; I use this layout in the simulations presented in the next chapter.

2.8 Grounding, valency, and underspecification

The exact nature of phonological features has been a matter of discussion ever since their conception, and after almost a century this discussion has not abated. This section discusses three topics concerning features in relation to the neural network model: grounding (§2.8.1), valency (§2.8.2), and underspecification (§2.8.3).

2.8.1 Grounding

Many phonologists agree that features are phonetically GROUNDED, meaning that they relate to auditory or articulatory properties. Jakobson, Fant and Halle (1952) define an auditorily grounded feature set, with such features as “strident” versus “mellow”, and “continuant” versus “interrupted”; a theory of phonology that argues for auditorily grounded primitives exclusively is Element Theory (Harris & Lindsey 1995). Chomsky and Halle (1968) defined a feature set that is rooted in articulation (e.g. [±round], [±nasal]). In Direct Realist frameworks (Lieberman et al. 1967; Fowler et al. 1980; Lieberman & Mattingly 1985), such as Motor Theory, it is even argued that phonological features do not exist; instead, the articulatory movements themselves are the objects of speech perception.

The model presented in this chapter does not possess an articulatory layer or sensorimotor knowledge, so the emerging features cannot be articulation-based; in an extended version of the model, the learner should also be able to acquire knowledge of sound-to-gesture mappings (in the BiPhon model: the relation between the auditory and articulatory forms). With such an extension of the model,

articulation-based features can be induced, probably relating to articulatory gestures that require little effort and result in sounds that are perceptually distinct within the sound system. In the model as it is presented here, categorical behaviour emerges after auditory distributional training only (as it did in Boersma, Benders & Seinhorst 2020 as well), so these features are initially strictly auditorily based; they can grow to incorporate other sources of information too, such as lexical information (as I showed in §2.4) or morphological alternations (as modelled by Chládková 2014). As a result of this interaction, features may emerge that cannot strictly be defined in terms of phonetic properties, potentially giving rise to “unnatural” classes as found by Mielke (2008): such features are often called *SUBSTANCE-FREE* (Fudge 1967; Blaho 2008; Samuels 2011).

Because the neural network is accessible to visual inspection, we can directly observe the way in which phonological categories emerge; we are able to draw conclusions about the grounding of the categories because we can see how the SF layer becomes connected to the other levels of representation within the network. In other computational models of category learning (e.g. Feldman, Griffiths & Morgan 2009; McMurray, Aslin & Toscano 2009), their existence could only be inferred; Pajak, Bicknell and Levy (2013), for instance, train a Bayesian learner on a bimodal distribution, and the output of the learner is also bimodal, but it is unclear whether different phonological representations are involved.

2.8.2 Valency

A second topic in feature theory is *VALENCY*, the number of values that a feature can assume. I take *Bisa* as an example again (remember Table 1.1 on p. 3), which has the plosives {p t k b d g} and the fricatives {f s v z}. We need three contrastive properties to define this inventory: manner (plosive vs. fricative), voicing (voiceless vs. voiced), and place of articulation (labial vs. coronal vs. dorsal). The manner and voicing contrasts are *BINARY* (also “bivalent” or “equipollent”), meaning that they could be represented as taking one of two possible values: we can, for instance, postulate a feature [voice] and assign positive [+] or negative [-] values to the segments possessing and lacking this property, respectively. Features were assumed to be (universally) binary by e.g. Jakobson (1941), Jakobson, Fant and Halle (1952) and Chomsky and Halle (1968). Multiple binary features are needed to express a contrast between more than two feature values: in a language with three vowel heights, the high vowels can be described as [+high], the low vowels as [+low], and the mid vowels should be specified as [-high, -low]. The combination [+high, +low] is deemed articulatorily impossible, as the features are each other’s polar opposites.

Another view holds that features are *PRIVATIVE* (or “monovalent”), meaning that their possible values only describe those categories that possess that value (cf. Van der Hulst 2016 for discussion). In this view, high vowels could for instance be specified as [HIGH], mid vowels as [MID], and low vowels as [LOW].

Different assumptions regarding valency make different predictions with regard to potential natural classes: again considering the example of vowel height, a theory with monovalent features predicts that high and mid vowels could not form a natural class, and these vowels are not therefore expected to pattern together in phonological processes, whereas a binary feature value [–low] would subsume both vowel heights in a single natural class. Boersma (1998: 355) argues that the assumption of binarity arose because most phonetic continua cannot easily be divided into more than two categories.

In the neural network model presented in this chapter, more than two phonological categories may emerge on a single continuum, as they did on the (leftmost) place continuum of the network trained on the {ʂ ɛ s z} inventory in §2.7.1 (Figure 2.29, p. 45). Each of the three peaks in the CoG distribution is encoded by two SF nodes; each of these two nodes is connected only to that peak. These results are expected if we assume that features are privative, and unexpected if features are binary: after all, two binary features are required to define a three-way contrast (as shown above in the vowel height example), and therefore the surface representations of the different categories should display overlapping nodes. Chládková (2014: ch. 5) found privative representations in her simulations of feature induction in vowel systems; different values of the height and backness features were represented by different SF nodes. In the remainder of this dissertation, I assume that features are privative, and write them in [SMALL CAPS].

2.8.3 Underspecification

A third topic in feature theory is UNDERSPECIFICATION, a notion that entails that sometimes certain feature values may not be specified at all. This notion applies, for instance, in Turkish vowel harmony, where a high vowel occurs in the possessive morpheme (with the exception of the third person plural); this vowel agrees with the last vowel in the noun stem in terms of backness and roundedness. Since the vowel in the suffix copies these specifications from the vowel in the stem anyway, it could be argued that the vowel in this specific suffix does not need any backness and roundedness specifications of its own (Clements & Sezer 1982), only its height specification.

Empirical support for the psycholinguistic reality of underspecification comes from experiments suggesting that listeners seem to accept a labial where a coronal is expected (because they do not show a mismatch negativity response, for instance), but they do not accept a coronal where a labial is expected (a.o. Lahiri & Van Coillie 1999; Cummings, Madden & Hefta 2017). Dijkstra and Fikkert (2011) found that already at 6 months, Dutch infants accept /p/ as an instance of |t|, but not vice versa, and Tsuji et al. (2015) found the same effect in Dutch and Japanese infants between 4 and 6 months old. Lahiri and Reetz (2002, 2010) explain this asymmetry by positing that in the lexicon, coronals have no place feature value, so that any

perceived place feature value is acceptable to the listener, while labials have a [LABIAL] specification, so that any perceived non-labial elicits a mismatch response. This response, then, must result from a comparison between the features extracted from the speech signal and the features in the lexicon (Lahiri & Reetz 2010: 50), so between the surface and underlying forms. Over recent years, behavioural and neurophysiological evidence for many other underspecified features besides [CORONAL] has been found: for instance; Scharinger et al. (2016) found that German front vowels are underspecified for backness, and Schluter, Politzer-Ahles and Almeida (2016) found that English |h| is underspecified for place.

On the other hand, Ren, Cohen Priva and Morgan (2019) argue that in many studies that support underspecification, this conclusion is unwarranted because the lexicon is not involved in the experimental task: in the studies with infants, the participants are too young to have formed a lexicon, and in most studies with adults, no word recognition takes place because the stimuli are non-words or isolated segments. Instead, the authors ascribe any asymmetries to differences in frequency of occurrence, and to higher variability within coronals than within labials. Indeed, Boersma (1998: §9.5) computes confusion probabilities between feature values with different frequencies of occurrence, showing that apparent underspecification stems from a frequency bias: in a listener who has optimised their perception to make the fewest mistakes, an uncommon value is relatively more likely to be confused for a common value than vice versa, so the auditory correlate of the uncommon value is more distinctive, and will therefore be specified more strongly. The coronal feature value, being articulatorily least effortful and therefore more frequent, is then expected to have a weaker specification.

It is not straightforward to evaluate the prediction of underspecification in light of the neural network model, mainly because it is unclear how exactly it would look in the model. The most straightforward interpretation would probably be a lack of activity in the lexical layer in the neural network; in word recognition, then, fully specified features at SF need to be mapped to nothingness at UF, and in production, nothingness at UF needs to be mapped to fully specified features at SF. This situation is at odds with the fundamentals of neural networks, which are characterised by activation spreading between (and possibly within) all levels. Since a lack of activity in the nodes in one level cannot elicit activity in connected nodes in another level, this cannot be the correct solution.

In the networks in this chapter, the auditory probability distributions of all lexical categories were artificial and identical; it would be interesting to see how manipulations of these distributions, reflecting a more naturalistic situation, affect the other representations and the processes in which they are involved.

The cultural evolution of auditory contrast

The neural networks from the previous chapter served the purpose of phonological category creation only. The networks in this chapter also perform another task that is crucial to the average speaker–listener: they speak as well as listen. I create a number of diffusion chains in which a learner first acquires a perception grammar and then produces spoken output, which constitutes the input to a second learner, who acquires a perception grammar on the basis of that input, and so on. We see that, irrespective of the dispersion of the initial inventory, sufficient auditory contrast emerges over a number of generations of learners, as a result of the interaction of the prototype and articulatory effects.

3.1 Explanations of auditory dispersion

The effects of the forces of auditory distinctiveness and articulatory ease, discussed in §1.3.1, can be witnessed in the structure of sound systems. The AUDITORY DISPERSION in such systems, i.e. the distribution of the auditory correlates of phonological categories in an n -dimensional phonetic space, is such that sufficient auditory contrast is maintained, but not so much that any more articulatory effort is required than necessary. Sound systems with exaggerated auditory contrasts, such as Ohala’s (1980: 185) proposed inventory /d’k’ ʃ ɪ m r ʌ/, would come at an astronomical articulatory cost, and do not seem to exist in the languages of the world.

Liljencrants and Lindblom (1972) simulated the optimal dispersion of vowel categories by treating the vowels as magnets: they found that entropy in the vowel space was lowest when the magnets were at maximal distances from one another. Ten Bosch (1991) repeatedly increased the distance between the two closest categories only, which yields better results than finding the minimal entropy in the entire system; he also included an articulatory effort function. Optimality-Theoretic formalisations of the structure of sound systems (Flemming 1995/2002; Kirchner 1998/2001; Padgett 2001, 2003) often make use of constraints that explicitly aim to maintain contrast between categories.

Such goal-oriented elements were absent from De Boer’s (2000; 2001) computer simulations; he modelled dyads of agents in which the agents played a vowel imitation game, imitating each other and providing each other feedback whether the

This chapter is an extended version of §6 from Boersma, Benders and Seinhorst (2020) as well as §4 from Seinhorst, Boersma and Hamann (2019).

original and imitated tokens belonged to the same lexical category; he found that dispersion might result from such self-organisation, meaning that it might be a consequence of the negotiation process between agents, without the agents *intentionally* improving their vowel systems. However, some of the actions that the agents could perform in De Boer's model were unrealistic (e.g. adding a random vowel, removing a vowel), and it is unclear how the model would fare without these actions.

In Hall's (2011) explanation of dispersion, some features are underspecified, and only specified features are phonetically enhanced in production. He provides a grammar that predicts typological patterns without the use of any teleological elements, but does not explicitly formalise the phonetic enhancement process.

According to Boersma and Hamann's (2008) OT formalisation, embedded within the BiPhon framework (familiar from Figure 2.2, page 23), an inventory is optimally dispersed if the so-called PROTOTYPE and ARTICULATORY EFFECTS cancel each other out. The prototype effect emerges during perceptual learning, when the learner learns which realisations of a lexical category can be confused with another category; more peripheral tokens are less confusable, and therefore better suited as prototypes.⁶ Because of the bidirectionality of the model, the speaker will want to use these same prototypical tokens in production too; however, such tokens likely require more articulatory effort than more central tokens. Due to stronger articulatory restrictions on peripheral tokens than on central tokens, the most frequently produced token will be a compromise between prototypicality and articulatory ease: it will be more central than the prototype. Using computer simulations, Boersma and Hamann show that if one effect is larger than the other, they will come to strike a balance within a number of generations, yielding an optimally dispersed inventory without any need for dispersion constraints or other teleological mechanisms.

3.2 Speech production in the neural network

In any Optimality-Theoretic formalisation, the problem mentioned at the beginning of Chapter 2 persists: constraints refer to already existing phonological categories. In this chapter I investigate how the neural network from the previous chapter, which has proven capable of handling category creation, behaves in a diffusion chain: does a language change as it is passed on from one generation to the next, and if so, in what way?

Of course, to answer this question, the network needs to be able to speak: after all, it needs to produce an output that can serve as the input to a new learner. The model from the previous chapter was only used to listen, but because the network is symmetric and the learning algorithm is direction-insensitive, the network is suited

⁶ Boersma and Hamann reduced the model to two levels: a surface form, containing lexically contrastive categories, and an auditory form, evaluated by articulatory constraints.

for use in the production direction as well. It needs to be extended with an articulatory layer, which is implemented here as a single node, connected to both auditory layers through inhibitory connections whose weights follow a parabolic function: the weights are greater at the effortful peripheries of the auditory continua, militating against the production of such tokens more strongly than the production of more central tokens. For the simulations in this chapter, the weights of the articulatory connections follow the function in Figure 3.1, with a value of -0.2 at the extremes (nodes 1 and 48), and -0.08 at the centre (halfway between nodes 24 and 25).

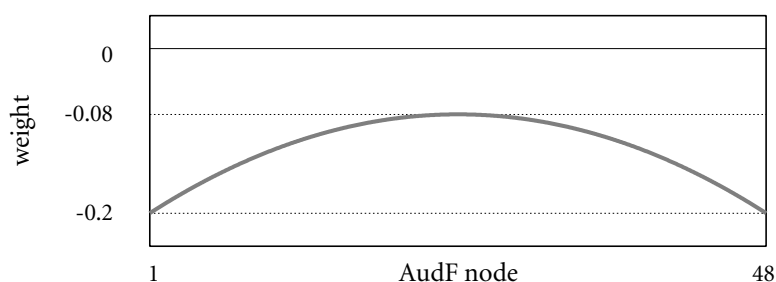


Figure 3.1. *The parabola defining the weights of the connections between the articulatory node and the auditory continua: the peripheries of the continua are inhibited more strongly than their centres.*

Remember from Figure 2.2 (p. 23) that in the BiPhon model, the auditory and articulatory representations are connected by sensorimotor connections; following Boersma and Hamann, I make the simplifying assumption that the learner's sensorimotor knowledge is perfect, so that the connections between the AudF and ArtF levels express articulatory effort instead.

In this extended version of the neural network, speaking entails the activation of the nodes that represent a certain meaning, and spreading the activities downwards to SF and AudF; at the same time, the articulatory node is activated and spreads activity upwards, inhibiting the nodes at AudF. The activities in the network are additive: to compute the activity of node j , we need to know for each node i that is connected to j its activity and the connection weight w_{ij} , multiply these two values, and add all these products (Lorente de N6 1938). Because the articulatory connection weights are negative, this addition may result in negative activities at AudF; I set such activities to 0 (cf. Usher & McClelland 2004 and Bogacz, Usher, Zhang & McClelland 2007 for the undesirable effects of negative activities).

Figure 3.2 shows the network from §2.10 (p. 30) producing the leftmost lexical category. The corresponding nodes in the lexical layer are activated and clamped, activity is spread to SF, after which the SF nodes are clamped, followed by activity spreading to (unclamped) AudF from both SF and ArtF.

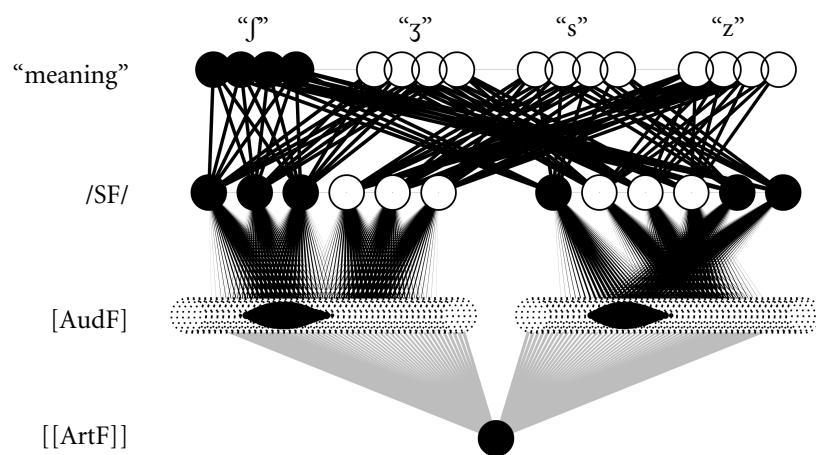


Figure 3.2. The network from Figure 2.10, now extended with an additional articulatory layer, producing the “ʃ” category.

The resulting activities at AudF for a given lexical category can be directly interpreted as a frequency distribution of auditory tokens: the activity of an AudF node equals its probability in an output distribution (see §2.5 in Boersma, Benders and Seinhorst 2020 for other possible interpretations). This probability distribution can be fed into a new network, thus creating a vertical transmission chain.

In earlier work (Seinhorst 2012), I modeled the emergence of auditory contrast with neural networks in which the phonological categories were present from the start, so the category creation stage was lacking; the simulations in this chapter include both stages. All simulations in this chapter were done with the two-stage learning strategy from §2.3. I simulate the cultural evolution of three types of initial distribution: a “standard” distribution (§3.3), a skewed and exaggerated distribution (§3.4), and a distribution with one bimodal category (§3.5).

3.3 A “standard” initial distribution

In §2.3 and §2.5–6, the network was trained on a language with four lexical categories, defined on two auditory continua: periodicity, and spectral centre of gravity (CoG). Treating the two-dimensional space defined by these continua as a Cartesian coordinate system with lower bounds of 0 and upper bounds of 100, the coordinates of the peaks of the auditory distributions of the four categories lie at (35, 35), (35, 65), (65, 35), and (65, 65). I refer to this language as “standard”. Figure 2.4 (p. 25) showed the probability distribution of tokens along a single auditory continuum; Figure 3.3 is a plot of the probability densities in two dimensions, discretised in a 48×48 grid.

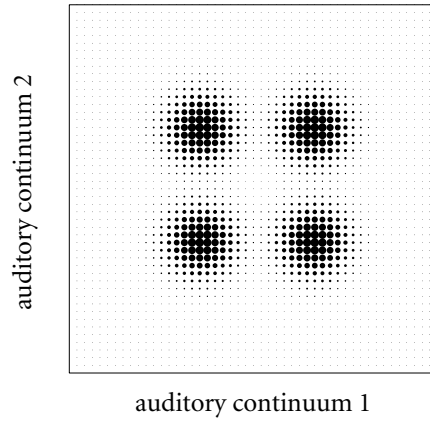


Figure 3.3. The probability densities of a “standard” inventory with four categories.

The evolution of this inventory is drawn in Figure 3.4. The figure shows the input probability distributions for each lexical category (on the horizontal axis), over ten generations (on the vertical axis), averaged over five runs of simulations. For each individual distribution, the thicker black curve connects the averages of each generation; the thinner grey curves connect the averages ± 1 standard deviation. Although there are four lexical categories, and therefore four probability distributions, only two distributions per continuum are visible in the figure: after all, as shown in Figure 3.3, on each continuum there are two categories that overlap entirely. These curves therefore coincide in Figure 3.4.

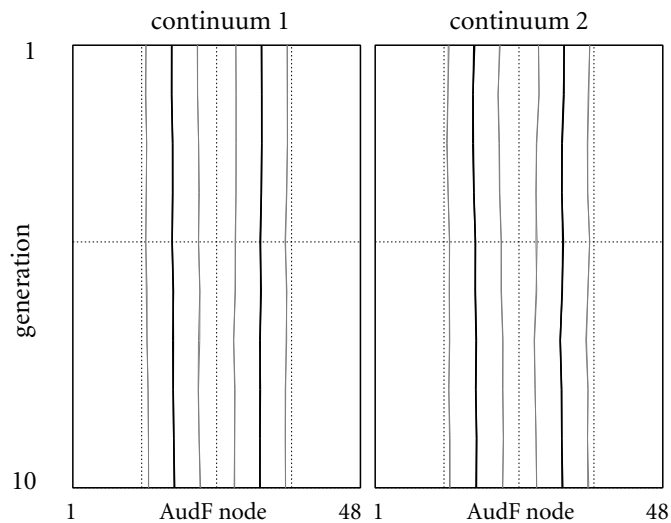


Figure 3.4. The evolution of a “standard” inventory over ten generations.

The curves in Figure 3.4 run pretty much vertically, which means that this language is transmitted faithfully between generations; there are some minor fluctuations due to the random transmission noise, but no noteworthy changes occur.

The feature induction process also occurs robustly in every generation. Figure 3.5 visualises the SF representations in the ten generations of learners in one run of simulations; this visualisation will be familiar from Chapter 2. Nodes belonging to contrasting feature values are drawn in contrasting colours; nodes that have remained unrecruited are drawn with dotted edges. In 15 out of 20 cases, the division between categories is 3–3; in 5 cases, it is 3–2.

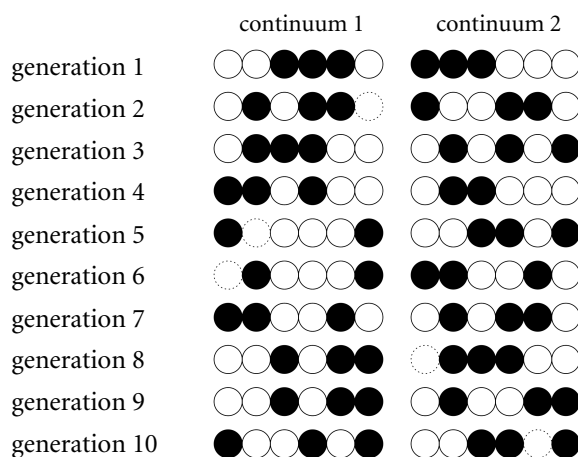


Figure 3.5. *The emergence of binary features across the six SF nodes of both continua over ten generations in one run of simulations.*

This “standard” inventory is already very neatly structured in terms of features: since every peak in the auditory probability distributions is shared by two lexical categories, it may be fairly unsurprising that the model indeed induces features. In the following two sections, I investigate the behaviour of the network under somewhat more adverse conditions. In the input language in §3.4, none of the peaks in the auditory distributions is shared by any lexical categories, and in §3.5, one of the lexical categories is distributed bimodally.

3.4 A skewed, exaggerated initial distribution

This section explores the evolution of another language with four lexical categories. The peaks of the auditory distributions of the four categories lie at (15, 65), (35, 15), (65, 85), and (85, 35), which means that their probability densities resemble a square that has been rotated counterclockwise by 23.6°; consequently, the pooled distri-

bution now has four peaks on each auditory continuum instead of two. Also, the peaks in the probability distributions are spaced 1.8 times further apart than they were in the ‘standard’ distribution (at a distance of 53.9 units in the grid, as opposed to 30 units for the “standard” inventory), so that the contrasts in this inventory are exaggerated:

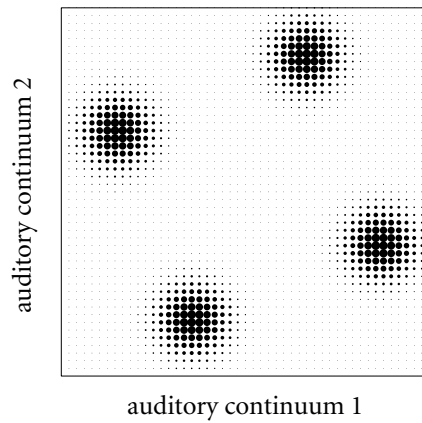


Figure 3.6. *The probability densities of a skewed, exaggerated inventory with four categories.*

Figure 3.7 shows the evolution of the input probability distributions for this inventory over forty generations, averaged over five runs of simulations. A number of interesting things can be seen in this figure.

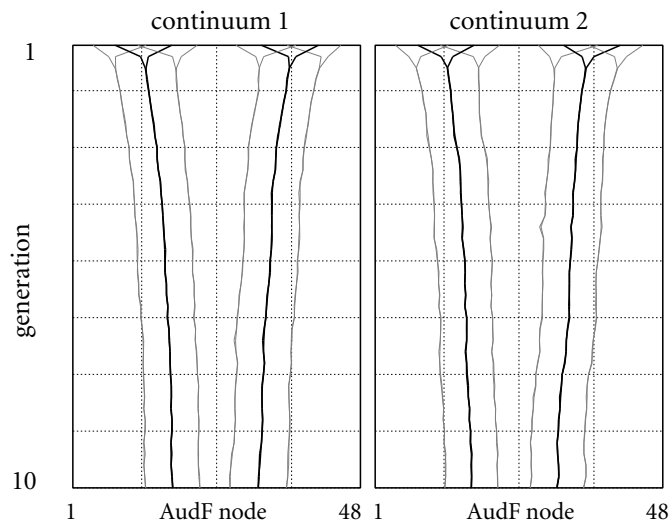


Figure 3.7. *The evolution of a skewed, exaggerated inventory over forty generations.*

Initially, we can still distinguish four individual probability distributions in the figure; however, after two generations, the categories that were at 20% distance from each other in the original language (i.e. those at 15 and 35% of both continua, and those at 65 and 85%) have merged. Apparently, in the original language, these two peaks were close enough that their pooled distribution caused the induction of a single category at SF in the distributional learning phase: the valley depths at 25% and 75% of the continuum are only circa 24% of the maximum probability, that at 50% is circa 80%. Because the lexical learning phase strengthened the connections between the lexical categories and their respective (unimodal) auditory input distributions, the output distributions of these categories in the first generation are also bimodal with a shallow valley, but they display a slight preference for their original peak. Figure 3.8 shows these output distributions of generations “0” (the original language) and 1. We can see strongly increased overlap in generation 1, but also remnants of the original peaks.

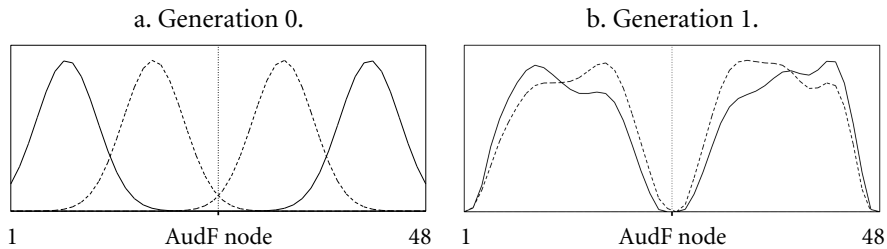


Figure 3.8. Output probability distributions of generations 0 (the original language) and 1. The horizontal axis shows auditory continuum 1, the vertical axis shows the probability densities of the tokens.

Figure 3.9 shows the division of categories across SF nodes in the first generation of learners for all five runs of the simulation. Nodes that represent contrasting feature values are drawn in black and white; unrecruited nodes are drawn with a dotted edge. Some nodes do not represent a feature value, but instead have been recruited by a single lexical category only; such nodes are drawn in grey.

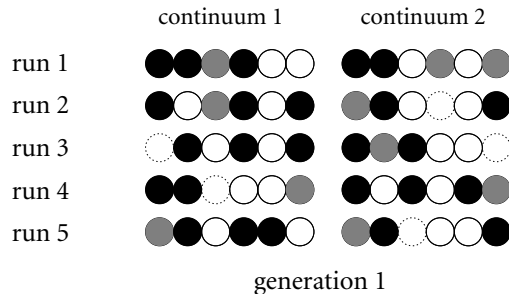


Figure 3.9. Division of categories across SF nodes in the first generation. Grey nodes belong to a single lexical category only.

Every learner in generation 1 induced two feature values (drawn in black and white in Figure 3.9), each consisting of two or three nodes; this must be due to overlap between the two pairs of closer peaks in the pooled distributions of the original distribution. In many cases, one or two lexical categories also have their own dedicated node, probably because the peaks in the pooled distribution did not yet overlap completely.

In the second generation, the overlap has increased further: the articulatory effect has pushed the outer peaks towards the centre of the continuum, so that the pooled distribution is now unimodal (albeit quite wide). Only a single feature value is induced for this peak during the distributional learning phase, and as a result, the lexical categories become connected to the phonological categories in the exact same way during the subsequent lexical learning phase; therefore, the categories at SF have identical output distributions at AudF, or in other words, these categories have merged. This is reflected in the representations at SF, shown in Figure 3.10, which have only two contrasting feature values left.

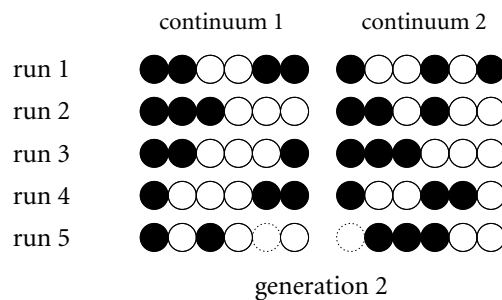


Figure 3.10. *Division of categories across SF nodes in the second generation. There are no longer any nodes that belong to a single lexical category only.*

Once this merger is complete, the contrast within the inventory is still very large: on both continua, the distance between the centres of the auditory distributions equals almost 50% of the continuum. This contrast is bigger than it was in the stable “standard” distribution, and this exaggeration is gradually resolved: each generation adds a small articulatory effect, causing the distributions to move closer towards each other. After some 35 generations, the locations of the distributions stabilise: although the articulatory effect wants to push the categories further towards the centre of the continuum, the prototype effect prevents them from approaching each other too closely. This final state is the stable distribution familiar from §3.3.

Figure 3.11a shows the output distributions of the second generation on auditory continuum 1; Figure 3.11b shows the output distributions of the 40th (and last) generation of learners. The output of the second generation is strictly bimodal, and by the 40th generation, due to the articulatory effect, the peaks are narrower, and the exaggerated contrast has been reduced.

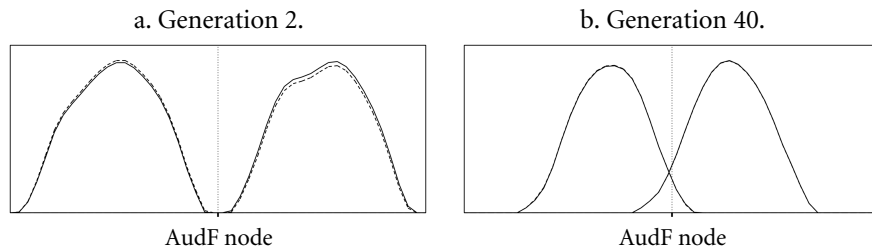


Figure 3.11. Output probability distributions for generations 2 and 40 (the horizontal axis shows auditory continuum 1, the vertical axis shows the probability densities of the tokens).

Three types of learning play a role in this diffusion chain: distributional learning, lexical learning, and iterated learning. All three types are crucial in the evolution of the inventory. Even though none of the peaks in its initial distribution were shared between lexical categories, features already emerged within the first generation, during distributional learning (the beginning of a merger); because the articulatory effect caused the pooled input distributions to the second generation to become unimodal, the lexical categories that share a feature value became connected to the SF layer in the exact same way during lexical learning, yielding identical output distributions of auditory tokens (the completion of the merger); and the iterated learning process gave the prototype and articulatory effects time to balance out.

3.5 A distribution with one bimodally distributed category

As I mentioned in §2.4.1, languages with a bimodally distributed category, that is, a category whose probability distribution has two peaks, do not seem to exist in natural languages. Boersma and Hamann (2008) modeled the evolution of such a category within OT; they showed that it can be represented in a grammar, but is diachronically unstable, becoming unimodal within a small number of generations. In this section, I explore the evolution of a language with two lexical categories of which one is bimodal. The initial probability distribution of the categories is shown in Figure 3.12. The probability densities of the unimodal category are drawn as black circles, and the densities of the bimodal category as white circles. The peak of the unimodal category lies at (45, 60); the bimodal category has a smaller peak at (15, 40), and a larger peak at (75, 40). The larger peak is twice as high as the smaller peak. On auditory continuum 1, the smaller peak is closer to the edge of the continuum than the larger peak; on auditory continuum 2, the distributions of both lexical categories are fairly close together, at the same distances as the closer categories from the skewed inventory from §3.4.

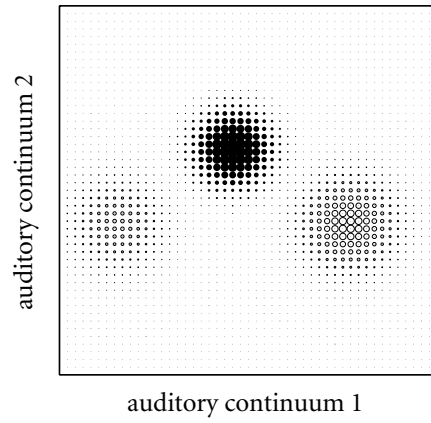


Figure 3.12. *The probability densities of an inventory with one bimodal category.*

The evolution of this inventory over eighty generations is shown in Figure 3.13.

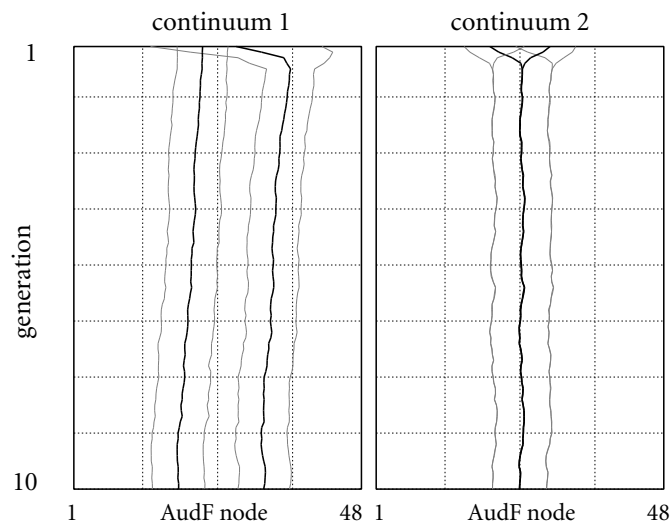


Figure 3.13. *The evolution of a language with one bimodally distributed category.*

In order to better understand the evolution of this inventory, it is useful to take a look at the output distributions on continuum 1 of the original language and the first four generations, as well as the 80th (and last) generation. These six plots are shown in Figure 3.14.

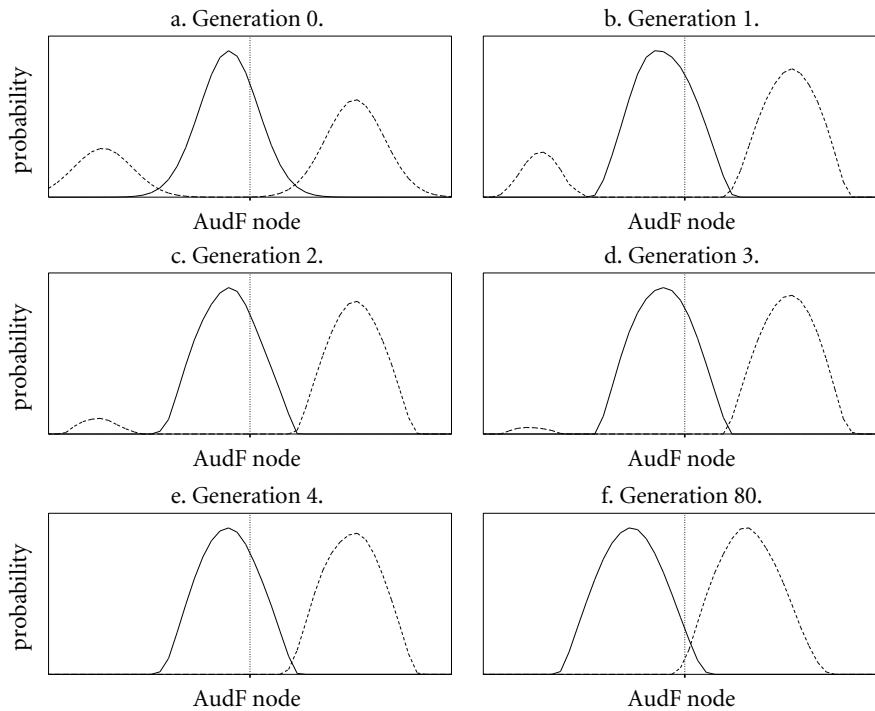


Figure 3.14. Output probability distributions on continuum 1 for different generations of learners, averaged over five runs of simulations.

In the first few generations, the articulatory effect reduces the size of the smaller peak of the bimodal category, thereby causing every following generation to prefer the larger peak of this category; as a result, the mean produced token shifts towards higher values on the continuum, as reflected in the rightward shift in Figure 3.13. Because the left peak becomes increasingly smaller, the standard deviation of the output probability distribution as a whole also decreases strongly; this too is clearly visible in Figure 3.13. By the fourth generation, the bimodal category has become unimodal, leaving room for both categories to shift towards more central values. They eventually stabilise around the familiar values of 35% and 65% of the continuum.

In the initial distribution, the peaks of the bimodal category were of unequal size; however, even when both peaks of the bimodal category are of equal size initially, learners will soon prefer one peak over the other, perhaps because one peak was randomly selected more often during perceptual learning, and/or because the random transmission noise scattered tokens in one direction more often than in the opposite direction (Seinhorst 2012: 37–38). This process is irreversible once one peak has outgrown the other.

On the second auditory continuum, the distance between the peaks in the initial distribution, which equals 20% of the continuum, results in a shallow valley in the auditory environment of the learner, causing some learners to induce two separate representations for the peaks while the categories merge in other learners. Figures 3.15a–f plot the output distributions of the network in generations 0 through 4 and generation 80. Note how Figure 3.15b resembles Figure 3.8b: in both cases learners seem to prefer the original peak, but nevertheless an irreversible process of merger has begun.

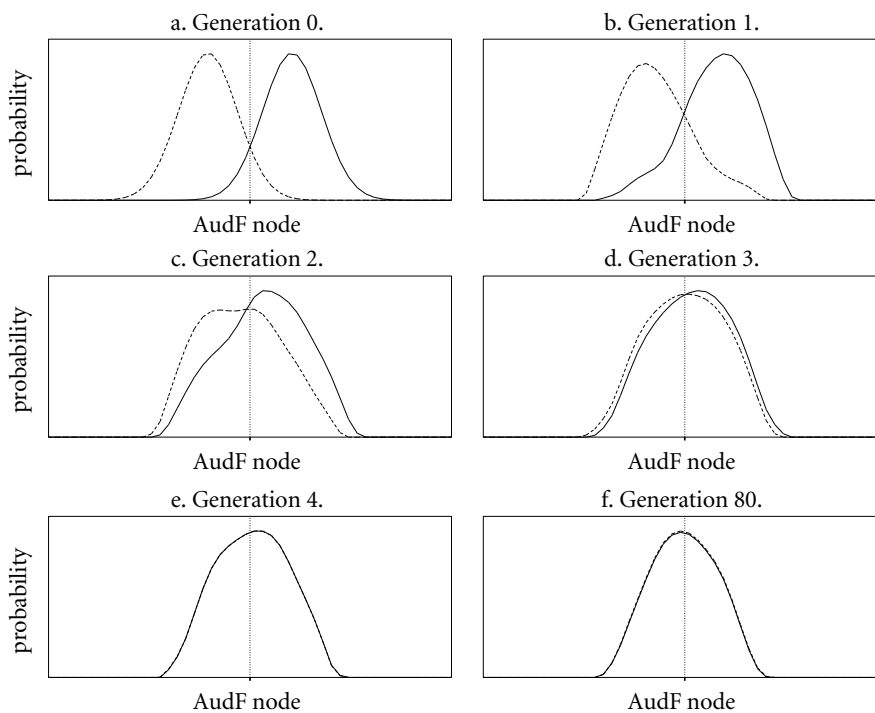


Figure 3.15. *Output probabilities on continuum 2 for generations 0–4 and 80.*

The evolution in Figure 3.15 is mirrored in learners' representations at SF, shown in Figure 3.16. The figure displays the representations at SF on auditory continuum 2, for all five runs, for the first four generations. Nodes that belong to contrasting lexical categories are drawn in black and white; nodes that are shared by the lexical categories are drawn in grey. Most learners in generation 1 induce representations in which the two lexical categories are represented both separately (white and black nodes) and together (grey nodes); only the representation of generation 1 in run 2 does not display any overlap between the two lexical categories. Once nodes have become shared by both categories in perceptual learning, the bidirectionality of the

network will cause their realisations to merge too, just as they did in the skewed distributions from §3.4. This process reinforces itself with every following generation, and Figure 3.16 shows that by the fourth generation all SFs consist only of shared nodes or unrecruited nodes.

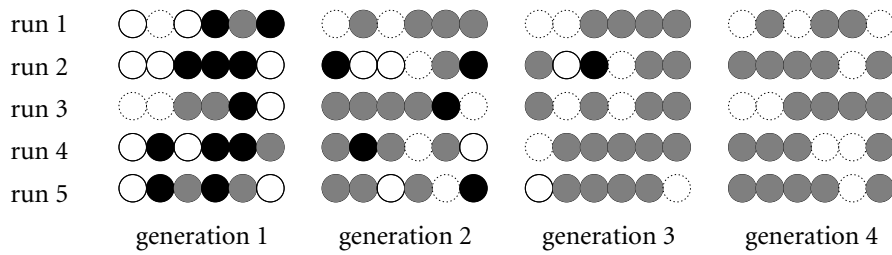


Figure 3.16. *Division of categories across SF nodes in the first four generations, for all five runs.*

After the two categories have merged, a smaller part of the auditory continuum is used (compare Figures 3.15a and f); consequently, the SF representations of all five learners in generation 4 contain one or more unrecruited nodes.

The simulations in this chapter predict that any initial distribution will reach a stable state over time, but some attested sound changes are known to be cyclic: because the perception of speech sounds usually involves the mapping of more than one auditory continuum to a phonological category (unlike the simulations presented here), a change towards optimal dispersion on one auditory continuum may result in suboptimal dispersion on another continuum, thus setting in motion a potentially never-ending chain of sound changes (see Boersma 2003 for a formalisation).

The big fish eat the little ones
The big fish eat the little ones

(Radiohead — Optimistic)

Part II:
EXPERIMENTS

4

Phonological pattern learning (1): a 3×2 parameter space

The iterated learning paradigm (§1.2.3, p. 7) explicitly links tendencies in language evolution and linguistic typology to individual language learners. Their biases in language acquisition, perception and production may be amplified with every subsequent generation; these biases determine how cross-linguistic distributions of typological traits develop diachronically. Suppose a morpheme exists with two allomorphs, A and B. Learners are biased towards B: with every generation, 1% of all instances of A is replaced by instances of B, but 0% of all Bs is replaced by A. This means that if we start with 1000 instances of A and 1000 instances of B, after one generation, ten instances of A have been replaced by B. This growth rate decreases over time, because the number of target forms on which it operates decreases; although the inductive bias still exists, the growth rate asymptotically approaches zero, and eventually allomorph A will be (nearly) extinct.

To investigate such inductive biases in the learning of patterns of phonological feature combinations, I conducted experiments with human learners, manipulating pattern complexity. In the first set of experiments, described in this chapter, learners were trained on phonological patterns that could be described as combinations of one maximally ternary feature (i.e. a feature with two or three values) and one maximally binary feature (i.e. a feature with one or two values); in the second set of experiments, presented in Chapter 5, learners were trained on phonological patterns that could be described in terms of two ternary features.

This chapter is structured as follows. After a discussion of studies investigating the learning of patterns of non-linguistic feature combinations (§4.1), a description of two relevant complexity measures (§4.2), and a discussion of earlier studies of phonological pattern learning (§4.3) I introduce the experimental task, stimuli, and analysis (§4.4), as well as results from two experiments: a pilot study with 48 participants using handshapes as stimuli (§4.5), and a larger study with 96 subjects, using handshapes as well as spoken language as stimuli (§4.6 and §4.7, respectively). The results are interpreted in terms of the two complexity measures in §4.8; the effect of modality is assessed in §4.9.

This chapter expands on Seinhorst (2017), in which the data from §4.5 were published. The data from §4.6 were published as Seinhorst (2016a).

4.1 Pattern learning of non-phonological feature combinations

In experimental psychology, the learning of classes of non-linguistic feature combinations has been investigated since at least the early 1960s (a.o. Shepard, Hovland and Jenkins 1961; Nosofsky et al. 1994; Feldman 2000). These sources use a set of eight stimuli whose properties are described in terms of three binary features. This data set can be divided into two mutually exclusive classes of four stimuli each in $\binom{8}{4} = 70$ different ways, all of which may be – through rotation and/or mirroring – reduced to one of six possible CATEGORY STRUCTURES or TYPES. These are types I–VI in Figure 4.1. The three features are represented in the three dimensions of the cubes; members of contrasting classes are drawn in contrasting colours.

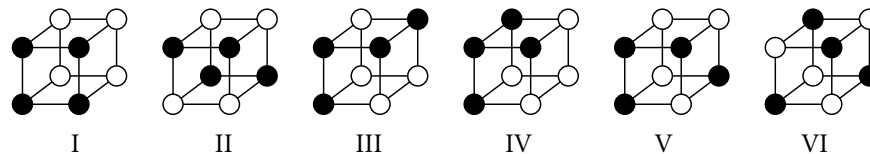


Figure 4.1. The six category structures or “types” from Shepard, Hovland and Jenkins (1961). Stimuli in contrasting classes are drawn in contrasting colours.

As noted above, for each type a number of permutations exists: different groups of stimuli that are nevertheless identical in terms of their relations. By way of example, let us look at type I. Its four stimuli lie on the same surface, and a cube has six surfaces, so this type has six permutations:

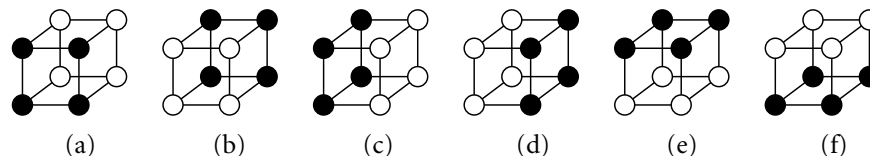


Figure 4.2. The six permutations of Shepard type I.

Suppose that the three features are shape (square vs. circle), size (small vs. large), and colour (black vs. white).⁷ Figure 4.3 shows the stimuli in all six possible permutations of type I, labeled (a)–(f). The figure shows how these permutations are related through rotation and mirroring: for instance, class A in permutation (a) is identical to class B in permutation (b), and vice versa; permutations (c) and (d) show this same kind of mirroring, as do (e) and (f). In order to classify a stimulus

⁷ These are the features that Shepard et al. used in the introduction of their study (with triangles instead of circles), and in their second experiment. In their first experiment, they used frames with three positions; in each position, one of two objects appeared (a nut or bolt in the top centre, a candle or light bulb in the bottom left, and a violin or trumpet in the bottom right).

from any type I division as belonging to class A or B, two features can be disregarded: for instance, in permutations (a) and (b) in Figure 4.3, this decision can be made on the basis of colour alone, disregarding shape and size; colour and size can be ignored in permutations (c) and (d); and colour and shape are redundant in permutations (e) and (f).

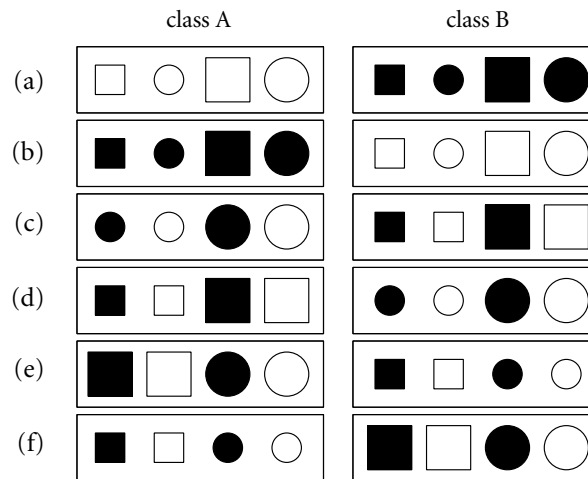


Figure 4.3. *The stimuli in each of the six permutations of type I.*

The internal structures of the types become more complicated for higher type numbers. In type II divisions, for instance, only one feature can be disregarded; for types III–VI, none of the features can be ignored. In type VI, the two members of each pair of stimuli within a class differ from each other on two features, and share only a single feature:

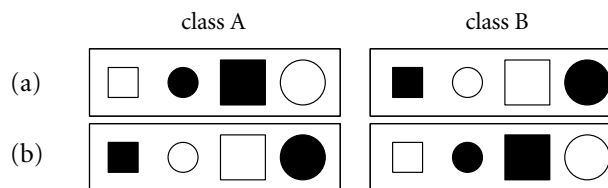


Figure 4.4. *The stimuli in both permutations of type VI.*

Shepard, Hovland and Jenkins carried out two experiments. In the first experiment, learners categorised stimuli as belonging to class A or B, and immediately received feedback on their response. The experiment was completed when participants had given 32 consecutive correct responses. In the second experiment, subjects were asked to formulate the rules they thought underlay the division, and two weeks later were instructed to recreate the division from memory. The results of both experi-

ments reflect the increasing difficulty of the six types: learners perform best on type I category structures, worse on divisions of type II, even more poorly on types III–V, and worst on type VI. For instance, many participants indicated that they had learned type VI divisions by rote. This order was replicated in a number of follow-up studies, both with human and computer learners (a.o. Kruschke 1992; Nosofsky et al. 1994; Edmunds & Wills 2016), although some studies found an advantage of type IV over type II, based on the kind of training and on properties of the stimuli. Nosofsky and Palmeri (1996), for instance, found such an advantage using stimuli in which the three relevant dimensions were perceptually integrated (hue, brightness and saturation in colours) rather than distinct; Love (2002) found the same advantage when learning was incidental (i.e. when it was a consequence of performing a different task) rather than intentional. This advantage may be explained because the manipulations lead to differences in selective attention and rule construction; these differences may have an effect on the learning of linear and non-linear separations, that is, those in which a line (or in this three-dimensional case, a plane) can be drawn separating the stimuli in both classes, as in type IV, and those in which no such lines divide the stimuli, such as type II. However, the advantage of linear separability is unclear: Medin and Schwanenflugel (1981) found no such advantage, and Levering, Conaway and Kurtz (2020) found that a non-linearly separable structure was acquired more easily than a linearly separable one.

Griffiths, Christian and Kalish (2008) conducted an iterated learning experiment using the Shepard types, each of which occurred as the input to the first learner in a diffusion chain. From a class of four stimuli, one stimulus was removed at random, and a learner was asked which of the five remaining stimuli (including the removed one) would complete the set. The new set of four stimuli formed the input to a following participant, with one stimulus being removed at random, and so on. Because the three stimuli are compatible with multiple types, the choice of the added stimulus reveals biases in the induction of a category structure. Indeed, regardless of the initial type, diffusion chains tended to converge to type I, the type on which Shepard, Hovland and Jenkins' learners performed best. These results suggest that learners induce exceptionless classes of stimuli if possible (see §4.2.3).

4.2 Complexity measures: feature economy and logical complexity

The title of this dissertation contains the word “complexity”. This word is ubiquitous in the linguistic literature, and it is used in countless different ways; often it is not even operationalised or quantified. In phonology, perhaps the most widely known use is found in the description of syllable structure, where positions are called complex if they contain more than one segment. However, many other interpretations are possible: see, for instance, Pellegrino et al. (2009) and the contributions therein.

In the last ten or fifteen years, much research has been done into the roles that complexity (in its various definitions) plays in language acquisition and in linguistic typology. Usually a fairly direct, inverse relation is assumed between the two: if a trait is less complex, it is probably easier to learn and more typologically frequent, and if a trait is more complex, it is probably more difficult to learn and less typologically frequent (Kirby and Hurford 2002; Christiansen and Chater 2008; Chater and Christiansen 2010; Kirby et al. 2015).

This dissertation focuses on two specific quantifications of complexity that can be applied to sets of feature combinations such as phoneme inventories, namely FEATURE ECONOMY and LOGICAL COMPLEXITY (or INCOMPRESSIBILITY). Both in the current chapter and in Chapter 5, I compare these measures as predictors for ease of learning in phonological acquisition tasks; in Chapters 6 and 7, I use them to assess the complexity of sound changes and sound systems.

4.2.1 Feature economy

The notion of FEATURE ECONOMY has been around since at least De Groot (1931) and Mathesius (1931). It states that phoneme inventories tend to maximally combine their phonological features: if a feature value is used distinctively in a sound system, it is likely used to define as many phonemes as possible. Clements (2003: 289) was probably the first to *quantify* the economy of an inventory: he proposed equation (4.1), in which C is the number of categories, and F the number of binary features that is needed to define the inventory.

$$(4.1) \quad \text{economy index} = \frac{C}{F}$$

For the consonant system of Hawaiian, which has eight categories and is defined using five binary features, this equation yields an economy index of 1.60; the French consonant system, with 18 consonants and seven binary features, has an index of 2.57 (Clements 2003: 290).

Clements' quantification depends on the number of features in the inventory, which makes comparison between languages intransparent: for instance, a language that uses two binary features and uses all four combinations has an economy index of 2.0, while a language with six binary features that uses all 64 combinations has an economy index of 10.7. This issue is largely resolved in Hall's (2007: 176) "Exploitation" measure, which computes an economy index by dividing the size of an inventory, that is, the number of phonemic categories in that inventory, by the size of the PARAMETER SPACE, that is, the number of *possible* contrasts that can be defined with the relevant features. Hall assumes binary features, but the principle can be extended to features of any valency. Equation (4.2), then, defines the exploitation index of an inventory with C categories, n_a contrasts on feature a , n_b contrasts on feature b , and so on:

$$(4.2) \quad \textit{exploitation index} = \frac{C}{n_a \cdot n_b \cdot n_c \cdot \dots}$$

The sibilant inventory $\{\text{ʃ } \text{ɛ } \text{s } \text{z}\}$ from §2.7, with its binary voicing contrast and ternary COG contrast, has an exploitation index of $\frac{4}{2 \cdot 3} = 0.67$. As opposed to equation (4.1), equation (4.2) makes no assumptions about valency (§2.8.2, p. 48), so it is compatible with binary as well as privative features.

Because the exploitation index is a proportion, in theory, its value lies between 0 and 1 for any inventory, allowing for cross-linguistic comparisons. However, in practice, its value can never be exactly 0, because the number of categories in a sound system is never zero; also, only a limited number of denominators exists for a given inventory. As a result, the lower bound of the range of possible exploitation indices is (much) greater than zero. For instance, to define an inventory with two phonemes, only a single binary distinctive feature is needed, so the only possible value of the exploitation index is 1. Similarly, a three-category inventory can only be defined either with one ternary distinctive feature or with two binary ones, so the lowest possible index for such an inventory is 0.75 (but see also §4.3.2, p. 80). The lower bound of an inventory with n categories is given by equation (4.3):

$$(4.3) \quad \textit{exploitation}_{\min} = \frac{n}{2^{n-1}} \text{ for } n \in \mathbb{Z}^{\geq 2}$$

The bars in Figure 4.5 indicate the ranges of possible values. The lower bound of this range approaches zero only in larger inventories, so an issue remains in comparing the exploitation measures of inventories of different sizes.

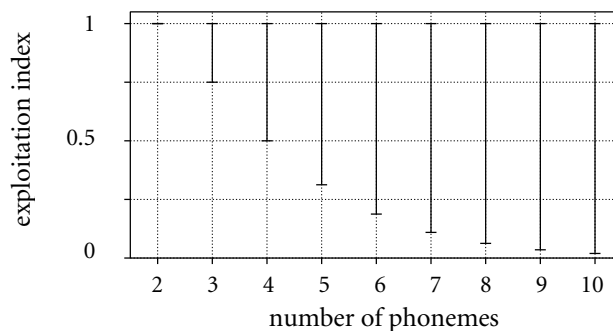


Figure 4.5. Possible values of exploitation index as a function of inventory size.

Hereafter, whenever I use the term “feature economy”, I refer to the exploitation measure; I use the abbreviation “ E ”.

Applying this measure to the six Shepard types, we divide the number of categories in a class (which is always four) by the product of the number of shape, size and colour distinctions within a class. For each type, Table 4.1 shows the number of categories in a class, the number of distinctions on each of the three features

(indicated with the generic labels a , b and c), and the economy indices E . Note that E does not necessarily correlate with the number of features that are relevant to learners in the tasks: in fact, this measure only distinguishes between type I on one hand, and all five remaining types on the other hand. If feature economy is related to ease of learning, higher values of E should correspond to better performance in a learning task. While this measure correctly predicts learners to do best on type I, it fails to make further distinctions in the learnability of the remaining types.

Table 4.1. *Feature economy indices of the six Shepard types.*

type	categories	distinctions on feature			E
		a	b	c	
I	4	2	2	1	1.0
II	4	2	2	2	0.5
III	4	2	2	2	0.5
IV	4	2	2	2	0.5
V	4	2	2	2	0.5
VI	4	2	2	2	0.5

4.2.2 Logical (Boolean) complexity

Feldman (2000) calls upon a different measure of complexity: he argues that participants' scores in Shepard, Hovland and Jenkins' experiments can be predicted by logical (Boolean) complexity or incompressibility, and that more compressible data sets are easier to learn. Table 1 from Feldman (2000: 631) is repeated here as Table 4.2 (next page). For each Shepard type, this table lists the disjunctive normal form (a summation of the feature values of all members of the class), the minimal formula (the shortest possible description of the class; a generalisation over the disjunctive normal form) and the logical complexity index lc of a class. This complexity index has been quantified as the number of literals in the minimal formula. More compressible sets can be represented with a shorter minimal formula, and are hence less complex, analogous to the Kolmogorov complexity in algorithmic information theory (Solomonoff 1960; Kolmogorov 1963). Again, a , b and c are the three binary features; I refer to their two possible values by using subscripts for the values (e.g. a_1 , b_2).⁸ The notation $a_1b_2c_1$ means $a_1 \wedge b_2$, describing a stimulus that simultaneously possesses properties a_1 and b_2 ; $a_1 + b_2$ means $a_1 \vee b_2$, so a_1 or b_2 ; $(a_1)'$ means $\neg a_1$, so any value of a that is not a_1 ; $(b_1c_1)'$ is a shorthand for the combinations of properties b and c that are not b_1c_1 , so b_1c_2 , b_2c_1 , and b_2c_2 if both features are binary.

The minimal formula can be seen as an abstract representation of a description in prose, and the length of the minimal formula roughly corresponds with the length

⁸ For binary features, it suffices to indicate whether a stimulus has that property or not; however, the notation with subscripts can be extended to non-binary contrasts, as I do later in this chapter.

of this description: for instance, we could describe type I succinctly as, for instance, “all squares” (two words), while type II requires a more elaborate description such as “the white squares plus grey circles” (six words), and type VI is “the small white and large grey squares plus the large white and small grey circles” (15 words).

Table 4.2. *Logical complexities of the six Shepard types.*

type	disjunctive normal form	minimal formula	lc
I	$a_2b_2c_2 + a_2b_2c_1 + a_2b_1c_2 + a_2b_1c_1$	a_2	1
II	$a_2b_2c_2 + a_2b_2c_1 + a_1b_1c_2 + a_1b_1c_1$	$a_1b_1 + a_2b_2$	4
III	$a_2b_2c_2 + a_2b_2c_1 + a_2b_1c_2 + a_1b_2c_1$	$a_2c_2 + b_2c_1$	4
IV	$a_2b_2c_2 + a_2b_2c_1 + a_2b_1c_2 + a_1b_2c_2$	$a_2(b_1c_1)' + a_1b_2c_2$	6
V	$a_2b_2c_2 + a_2b_2c_1 + a_2b_1c_2 + a_1b_1c_1$	$a_2(b_1c_1)' + a_1b_1c_1$	6
VI	$a_2b_2c_2 + a_2b_1c_1 + a_1b_2c_1 + a_1b_1c_1$	$a_1(b_2c_1 + b_1c_2) + a_2(b_2c_2 + b_1c_1)$	10

For type III, Feldman (2000) gives the minimal formula $a_2(b_2c_2)' + a_1b_2c_1$, so $lc = 6$: however, the shorter minimal formula in the table suffices to describe this type.⁹ The logical complexity indices reported by Feldman correctly predict the hierarchy found by Shepard, Hovland and Jenkins, as well as the results from his own learning experiments with smaller category structures. Interestingly, the corrected values in Table 4.2 do not explain the greater ease of type II over type III; these types differ in the number of properties to which the learner needs to pay attention – the minimal formulas show that feature c can be ignored in type II but not III – but the logical complexity values are by definition unaffected by the number of relevant features.

The issue with limited ranges of possible values for smaller inventories exists for the logical complexity measure too: for instance, the highest possible lc index of an inventory with three categories is 2. This makes between-language comparison difficult for smaller systems, as it is for feature economy as well.

4.2.3 Regularity

Feature economy indices are proportions with numerators that must be natural numbers \mathbb{Z}^+ (i.e. the set of positive integers), so their range is $(0, 1]$; logical complexity indices are also natural numbers \mathbb{Z}^+ , so their range is in principle $[1, \infty]$. Of all six Shepard types, only type I is REGULAR or SYMMETRICAL (these terms were also used in §1.1.2, p. 4): it maximally combines its constituent feature values, without exceptions or gaps. This is also the definition of our feature economy measure, so for any regular inventory, $E = 1.0$ by definition. Because a regular inventory has no exceptions or gaps, its minimal formula has minimal length, so its logical complexity index is 1 (cf. §4.3.2, p. 80). The two measures are inversely related: in a regular inventory, $E = lc = 1$, and gaps will lower E and raise lc .

⁹ Thanks to Silke Hamann, who pointed out Feldman’s mistake to me.

4.3 Pattern learning of phonological feature combinations

Many phonological pattern learning studies investigate the learning of phonotactic regularities that refer to combinations of phonological features (a.o. Pycha et al. 2003; Skoruppa & Peperkamp 2011; Baer-Henney 2015), sometimes creating data sets on the basis of the Shepard types. Moreton, Pater and Pertsova (2017) use all six Shepard types to investigate the learning of phonotactic dependencies. In their first experiment, they created a lexicon of $C_1V_1C_2V_2$ words in which both consonants and both vowels are defined using two binary features ($[\pm\text{voiced}]$ and $[\pm\text{coronal}]$ for the consonants, and $[\pm\text{high}]$ and $[\pm\text{back}]$ for the vowels), yielding a total of eight features per word: for each participant, three out of these eight features were selected at random, defining the subset of words to which the participant would be exposed. A type I language, for instance, might consist of words whose C_1 was voiced; a possible type II language would consist of words whose C_1 was voiced if V_2 was back, and so on. After exposure, learners would hear pairs of words, one conforming to the input and one not conforming to the input, and were asked to select the conforming word. Contrary to Shepard, Hovland and Jenkins' participants, type II learners performed worse than learners of types III, IV and V (but still better than type VI learners); in a second experiment using visual stimuli, Moreton, Pater and Pertsova found type II to be more difficult than type III.

Kuo (2009) used two of the six Shepard types in a phonological learning task. She trained her participants on two types of words with complex onsets. The first segment was one of the set $\{p\ t\ p^h\ t^h\}$, the second was one of the set $\{j\ w\}$; the second segment could depend on the first in two different ways. In one type, the choice of the glide was dependent only on the place feature of the preceding plosive, disregarding the aspiration feature (so $\{p\ p^h\}$ versus $\{t\ t^h\}$; Shepard's type II), while in the other type, both features needed to be taken into account (so $\{p\ t^h\}$ versus $\{p^h\ t\}$; Shepard's type IV). Learners were better able to generalise the first rule to novel stimuli than the second.

Saffran and Thiessen (2003) trained 9-months old learners of English on bisyllabic words. In one task, the possible onsets and codas form natural classes (e.g. $\{p\ t\ k\}$ as onsets and $\{b\ d\ g\}$ as codas), and in another, they do not (e.g. $\{p\ k\ d\}$ as onsets and $\{t\ b\ g\}$ as codas, respectively); infants have a stronger novelty preference in the former task than in the latter. Similarly, Cristià and Seidl (2008) found that 7-month old infants showed a larger novelty preference when they were trained on words with non-continuant onsets (e.g. $\{k\ b\ m\ n\}$) and tested on words with continuant onsets than when they were trained on words whose onsets could not be described as a natural class. These results suggest that infants are able to detect patterns in perceptually similar stimuli, and that they can generalise these patterns to novel words; that is, they seem to induce a phonological feature.

Skoruppa and Peperkamp (2011) investigated the role that phonological features play in rule learning by creating two variants of French, one with and one without vowel rounding harmony (“Harmonic French” and “Disharmonic French”, respectively). They found that adult learners of both variants are able to discriminate between congruent and incongruent stimuli; learners of a mixed variant, in which the rounding assimilation rule cannot be expressed in terms of phonological features, are significantly worse at the same task than learners of Harmonic French, and than learners of Disharmonic French.

4.3.1 Category structures of plosive inventories

The learning experiments that I present in this dissertation make use of category structures similar to the Shepard types, but in the experiments described in this chapter, they mirror the structure of cross-linguistically frequent plosive inventories. I chose plosive inventories because all spoken languages described so far make use of this type of speech sound (Maddieson 1984; Mielke 2008), an observation that allows me to compare experimental results with typological data in Chapter 7. Many languages have (at least) a three-way place contrast, usually [LABIAL] versus [CORONAL] versus [DORSAL], and most implement an additional laryngeal contrast, such as voicing or aspiration. Table 4.3 lists the combinations of these cross-linguistically frequent features, using a voicing contrast. I assume that all features are privative (cf. §2.8.2, p. 48).

Table 4.3. *Cross-linguistically frequent feature combinations in plosive inventories.*

	[LABIAL]	[CORONAL]	[DORSAL]
[VOICELESS]	p	t	k
[VOICED]	b	d	g

The translation of this inventory into a Shepard type-like structure yields a grid representation in two perpendicular dimensions: in one dimension (e.g. voicing), there are two values, in the other (place) there are three.¹⁰ Figure 4.6 shows some examples of inventories that can be represented within this grid. Suppose that in the figure, the bottom row contains all voiceless segments, the top row all voiced ones; of the three vertical edges, the left one connects the labials, the central one the coronals, and the right one the dorsals. The top left vertex thus indicates |b|, the bottom right vertex corresponds to |k|. Filled circles indicate categories that are part of the inventory, empty circles correspond to categories that are not part of the inventory.

¹⁰ A three-dimensional representation, more specifically a equilateral triangular prism, would better reflect the assumption that, at least phonologically, all feature values are in a sense equidistant from one another; however, a visualisation in two dimensions can more easily be extended to two ternary features, as happens in Chapter 5.

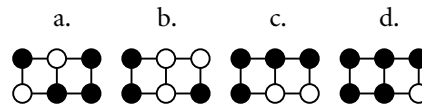


Figure 4.6. Four examples of inventories, represented as category structures: a. {t k b g}, b. {p k b}, c. {p b d g}, d. {p t b d g}.

Languages generally employ at least three of these segments, which means that a total of $\binom{6}{3} + \binom{6}{4} + \binom{6}{5} + \binom{6}{6} = 20 + 15 + 6 + 1 = 42$ different plosive inventories can be drawn from this set. Each of these can be reduced to one of eight different category structures; these are depicted in Figure 4.7. The names of the types consist of their parameter space size, that is, the number of theoretically possible feature combinations given the relevant features, followed by a capital letter which designates the category structures alphabetically.¹¹

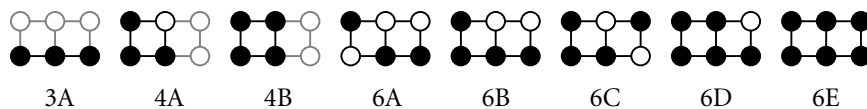


Figure 4.7. The eight different plosive category structures.

Each type, except for type 6E, has multiple permutations: type 3A has 2, type 4A 12, type 4B 3, type 6A 6, type 6B 6, type 6C 6, and type 6D 6. These permutations are found through rotation and mirroring. As an example, Figure 4.8 shows all six permutations (a–f) of type 6A:

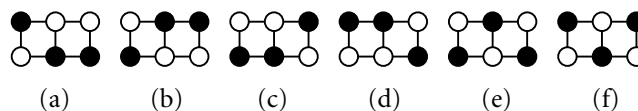


Figure 4.8. The six permutations of type 6A.

Pursuing the same assignment of feature values to the grid from Figure 4.6, these permutations correspond to {t k b}, {p d g}, {p t g}, {k b d}, {p k d}, and {t b g}, respectively. While these are six different inventories in terms of the segments employed, they are identical in terms of their category structure.

Comparing Figures 4.6 and 4.7, we see that the inventory in Figure 4.6a is a type 6C language; that in 4.6b a type 4A language; that in 4.6c a type 6B language; and that in 4.6d a type 6D language. Maori, having the plosive phonemes {p t k}, and Yidiny, with {b d g}, are both type 3A languages; Finnish, with {p t k d}, is a type 6B

¹¹ In Seinhorst (2016a, 2017), I used Roman numerals to designate the category structures, as Shepard, Hovland and Jenkins did; however, since more category structures are introduced in Chapter 5, a combination of a number and a letter is more distinctive.

language; Dutch, having {p t k b d}, is a type 6D language. Polish, with {p t k b d g}, and Mandarin Chinese, with {p t k p^h t^h k^h}, are both examples of type 6E languages, with different laryngeal features: Polish has a binary voicing contrast, Mandarin Chinese has a binary aspiration contrast. Maranungku has the plosives {p t k p: t: k:}, which could be another example of a type 6E language, with a binary length contrast instead of a binary laryngeal contrast. The {ʃ ɛ s z} inventory from §2.7 (p. 44) is another type 6B system, with sibilants instead of plosives.

Of course, many languages use more features than the category structures from Figure 4.7 can capture, and/or more elaborate contrasts within those features, and/or additional features that stand in a hierarchical relation to those features. For the sake of simplicity, I disregard those in this part of the dissertation, and explore only two-dimensional inventories.

4.3.2 Feature economy, logical complexity, and gaps

For each of the eight types, we can establish a feature economy index and a logical complexity index. In order to establish these values, we need to consider the relevant features in each type. If a type lacks a feature value, this value and the categories it might form were drawn in grey in Figure 4.7. For instance, only two place feature values exist in types 4A and 4B, so a learner of this type has no evidence to induce more than these two values: the complexity measures are established only on the basis of feature values for which positive evidence is found.

For each of the eight types in Figure 4.7, Table 4.4 lists the number of categories C , the number of laryngeal contrasts L , the number of place contrasts P .

Table 4.4. *Contrasts for all eight category structures.*

	type							
	3A	4A	4B	6A	6B	6C	6D	6E
categories	3	3	4	3	4	4	5	6
laryngeal contrasts	1	2	2	2	2	2	2	2
place contrasts	3	2	2	3	3	3	3	3

Note that a type 6A language, according to the table, has a binary laryngeal contrast, even though the place contrast suffices to distinguish between the phonemes of such a language. However, ignoring the laryngeal contrast might result, for instance, in |p| erroneously being realised as /b/, or vice versa, thus obscuring the difference between types 3A and 6A; therefore, this contrast is taken into account, even though it is redundant at the phonemic level. I return to this question in §8.3.1 (p. 165).

The feature economy indices of all eight types, computed according to formula (4.3), are listed in Table 4.5.

Table 4.5. *Feature economy indices for all eight category structures.*

	type							
	3A	4A	4B	6A	6B	6C	6D	6E
<i>E</i>	1.0	0.75	1	0.5	0.67	0.67	0.83	1

Table 4.6 presents the disjunctive normal forms, minimal formulas and logical complexity indices of the eight types. The laryngeal feature is represented as *a* and can have one of two values a_1 and a_2 ; the place feature is represented as *b* and can have one of three values b_1 , b_2 and b_3 . All minimal formulas and logical complexity indices are defined relative to the parameter space. Compare, for instance, types 4A and 6B in the table, which have identical minimal formulas and complexities, but different parameter space sizes: in type 4A, the place feature is binary, and in type 6B it is ternary.

There are three regular inventories, making full use of all possible feature combinations; these inventories necessarily differ in their parameter space size. They are indicated with the minimal formula *A* [all], and I have assigned them an *lc* index of 1.

Table 4.6. *Disjunctive normal forms, minimal formulas given the parameter space, and logical complexity indices for all eight category structures. The letters a and b indicate the laryngeal feature and the place feature, respectively.*

type	disjunctive normal form	minimal formula	<i>lc</i>
3A	$a_1b_1 + a_1b_2 + a_1b_3$	<i>A</i> [all]	1
4A	$a_1b_1 + a_1b_2 + a_2b_1$	$a_1 + b_1$	2
4B	$a_1b_1 + a_1b_2 + a_2b_1 + a_2b_2$	<i>A</i> [all]	1
6A	$a_2b_1 + a_1b_2 + a_1b_3$	$a_1(b_2 + b_3) + a_2b_1$	5
6B ¹²	$a_1b_1 + a_1b_2 + a_1b_3 + a_2b_1$	$a_1 + b_1$	2
6C	$a_1b_1 + a_1b_2 + a_2b_1 + a_2b_3$	$a_1b_2 + a_2b_3 + b_1$	5
6D	$a_1b_1 + a_1b_2 + a_1b_3 + a_2b_1 + a_2b_2$	$a_1 + b_1 + b_2$	3
6E	$a_1b_1 + a_1b_2 + a_1b_3 + a_2b_1 + a_2b_2 + a_2b_3$	<i>A</i> [all]	1

Contrary to Feldman (2000), I do not use negation in the minimal formulas. The complexity values from the previous table are also used in Chapter 7, where plosive inventories from spoken languages are analysed. Natural language acquisition is an implicit process, in which the learner usually has no awareness of the structures underlying their language, and has trouble making their knowledge explicit; therefore, I consider it unlikely that, for instance, a type 6D learner will represent this system as $(a_2b_3)'$. In order to compare the results from this chapter with the

¹² In §2.7, I suggested two possible representations of the sibilant inventory $\{\zeta \epsilon s z\}$, namely “[VOICELESS] + $|z|$ ” (p. 42) and “[VOICELESS] + [RETROFLEX]”; note that the logical complexity indices of these representations are the same, namely 2.

typological data, I do not use negation anywhere. It is an open question what degree of awareness participants have of their input in short experiments like the ones reported here, and the answer to this question bears directly on the logical complexity values. I compare minimal formulas with and without negation as predictors of learnability in §8.2.1 (p. 159); in those comparisons, the presence or absence of negation does not seem to influence the fit of the model.

Table 4.7 summarises the values of both complexity measures for all types:

Table 4.7. *Feature economy and logical complexity indices for all eight types.*

	type							
	3A	4A	4B	6A	6B	6C	6D	6E
<i>E</i>	1.0	0.75	1.0	0.5	0.67	0.67	0.83	1.0
<i>lc</i>	1	2	1	5	2	5	3	1

As I noted in §4.2.3, we see in Table 4.7 that feature economy tends to increase as logical complexity decreases: this negative correlation between the two measures is strong (Spearman’s rho: -0.866). (I do not report confidence intervals of this correlation because it was not computed from a random sample.) Perhaps the most notable difference between the measures is that the *E* indices of types 6B and 6C are the same, while the *lc* indices differ by three: both types use four out of six possible categories, but in type 6B these categories are arranged in a more compressible manner, because one of the laryngeal feature values is used exhaustively. Using a description in prose again as a proxy of compressibility, a type 6B language such as {p t k d} can be summarised as “all voiceless and coronal categories”, while a type 6C language such as {p t d g} requires a description like “the voiceless labial, both coronals, and the voiced dorsal”. The correlation between the number of categories in a type and its *E* index is negative but weak ($\rho = -0.237$); the correlation between the number of categories in a type and its *lc* index is positive but weak ($\rho = 0.413$).

The remainder of this dissertation focuses on these two measures of complexity, although of course more measures exist. Feature economy is a well established concept in the analysis of phoneme inventories, and logical complexity seems to play a role in the learning of feature combinations. Also, both measures are fairly independent of inventory size: by definition, a regular inventory has a feature economy index of 1 and a logical complexity index of 1, whether it contains three or thirty categories. This property does not hold for all complexity measures: for instance, the average number of minimal pairs in which a category participates might be an interesting measure of complexity, but this number is strongly correlated with the number of categories itself (see §8.2.2, p. 162). With “high complexity”, I hereafter refer to low values of *E* and high values of *lc*; an increase in *E* and a decrease in *lc* both qualify as a “reduction in complexity”.

4.3.3 Regularity (again)

Of the eight types from Figure 4.7, three are regular (3A, 4B and 6E); the other five have one or more of gaps, that is, categories that are absent from the language but that could have existed given the defining features of the inventory. Both experimental data and observations about natural language acquisition suggest that learners' errors tend to be regularising (a.o. Singleton & Newport 2004; Hudson Kam & Newport 2005, 2009; Reali & Griffiths 2009; Ferdinand, Kirby & Smith 2019). This tendency may be attributed to inductive biases: in the learning process, hypotheses favouring regular systems may have higher a priori probabilities and are thus more likely to be selected in acquisition, as Griffiths, Christian and Kalish' (2008) results also suggested. Such errors are readily explained in terms of a reduction of complexity, although the two measures may make different predictions: the feature economy of an inventory always increases when a gap is filled, but in order to evaluate a change in logical complexity, we need to know which gap exactly has been filled. For instance, if we add a category to a type 6A system, this may result in a type 6B system ($lc = 2$), or in a type 6C system ($lc = 5$), depending on the gap. In the former case, the logical complexity has been reduced; in the latter, it stays the same. In both cases, E increases from 0.5 to 0.67.

4.4 Learning experiments: stimuli, method, and analysis

The remainder of this chapter presents two experiments, nearly identical in terms of design, that aim to answer the question what effect the complexity of a phonological pattern has on its learnability. Such studies, probing the learning of linguistic structures and/or the mappings between them, are usually **IMPLICIT LEARNING** experiments, in which participants are not instructed about relevant properties of the input. In implicit learning experiments, participants are often expected to keep track of statistical properties of the input, such as the simultaneous or sequential occurrence of certain traits.

An issue in the experimental investigation of the acquisition of sound systems, at least if the participants have acquired a certain amount of phonological knowledge, is the risk of interference from participants' language background with the learning task. For instance, if one of the categories to be learnt is absent from the participant's language, they are likely to map tokens from that (foreign) category to a different, native category (a.o. Lisker 2001 on the perception of the Polish three-sibilant system by English listeners with a two-sibilant system; for Dutch listeners' perception of [g], which is lacking in their language, see Schuttenhelm 2013). If the segments in the stimulus set are a subset of the participant's segment inventory, they could simply draw upon part of their knowledge, in which case any gaps in the subset do not correspond to gaps in the learner's inventory. This makes it difficult to probe the role that learning biases play in the emergence of a new feature system.

Also, there is some evidence that structurally similar rules with different phonological features are learnt differently: for instance, 5-year old learners of German apply a final devoicing rule to novel words a disproportionately high amount, and a (historically motivated but now unproductive) vowel assimilation rule a disproportionately low amount (Van de Vijver & Baer-Henney 2012), supposedly because the former rule is phonetically “natural” and the latter is not.

A possible solution to these problems is the use of linguistic stimuli in a different modality, namely signs.¹³ This strategy avoids influences from the extant language system: Smith, Abramova and Kirby (2012), for instance, use sign language to investigate how the encoding of semantic features emerges in a preset meaning space (viz. manner and path in descriptions of movement). In their experiment, however, an influence of the participant’s language background still seems likely, as a result of the semantic component of the task; for the investigation of phoneme inventories no transfer is expected, because the substance of the features is defined in different modalities.

4.4.1 Stimuli

The two learning experiments described in this chapter both used handshapes as stimuli. 48 learners participated in the first experiment; the second experiment was a larger-scale replication, with 96 participants. In order to explore the role of modality, the same 96 participants also executed the same task with speech stimuli. I hereafter refer to the smaller-scale experiment as “experiment A”, and to the larger experiment as “experiment B”; the handshape learning tasks are tasks “A1” and “B1”, respectively, and the speech learning task “task B2”. In terms of their design, the pilot study and the follow-up study were nearly identical; in the follow-up study, the method was identical for both tasks, but the stimuli were slightly different, as I specify below. Experiment A was conducted in the Netherlands; Experiment B was conducted in Düsseldorf, Germany.

For tasks A1 and B1, a data set was created that consists of six simple handshapes. In analogy with the ‘basic’ plosive set {p t k b d g}, these handshapes can be described as a combination of two features: a binary thumb opposition feature, and a ternary handshape feature. The thumb could be either opposed or unopposed; the three handshapes are (1) zero fingers pointing up (a clenched fist); (2) the index finger pointing up; (3) four fingers, i.e. all except the thumb, pointing up. Tokens of the six handshapes, converted to greyscale, are shown in Figure 4.9.

¹³ I’m greatly indebted to Roland Pfau and Anne Baker for their suggestion to use the visual modality.

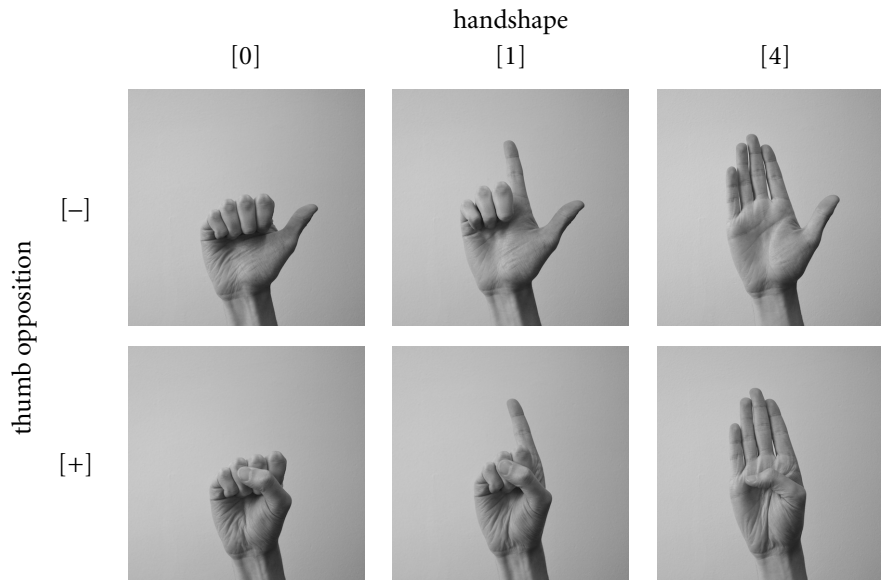


Figure 4.9. *The six phonemes of the artificial sign language.*

Handshape is one of the three cardinal components of sign language phonology (Stokoe 1960; Brentari 1990; Crasborn 2001; Van der Kooij 2002); natural sign languages tend to make many more distinctions than the three presented here, and referring to these three values as [0], [1] and [4] is a gross oversimplification of the possible contrasts in natural sign language as well as their hierarchical organisation (Sandler 1989; Van der Hulst 1993; Van der Kooij 2002). Thumb opposition is not a prevalent distinction in natural sign languages (Anne Baker, p.c.), and the six handshapes in Figure 4.9 lack the other two cardinal sign phonological components, namely movement and location; for this reason, I refer to the categories as mere “handshapes”, not full-fledged “signs”.

The features and feature values defining the six handshapes in Figure 4.9 were chosen because they seem perceptually sufficiently distinct, while occupying a small amount of physical space, making them well suited to be photographed for the experiment. Additionally, movement and orientation contrasts seem to be more difficult to perceive and produce than handshape contrasts, in so far as handshape and orientation are distinguished (regarding production errors in late learners, cf. Ortega & Morgan 2010 for British Sign Language, Willoughby et al. 2015 for Auslan, and Ferrara & Nilson 2017 for Norwegian Sign Language; for perception errors in late learners of American Sign Language, see Bochner et al. 2011).

For task A1, a native signer of Sign Language of the Netherlands was photographed producing each handshape ten times; for task B1, a non-native signer was photographed producing eight tokens of each handshape (this signer is seen in

Figure 4.9). Multiple tokens were photographed of each category, so that the data set contained within-category variability. This was done in order to replicate the lack of invariance with which every learner of a phonological system is faced, and from which she has to induce the intended discrete categories that are relevant in her language.

In task B2, spoken Standard German stimuli were used. Standard German has six plosive categories, usually transcribed as {p t k b d g}; the laryngeal contrast is commonly described as [\pm spread glottis] (Iverson & Salmons 1999; Jessen & Ringen 2002; Honeybone 2005), meaning that word-initial {p t k} tend to be implemented with long VOTs and aspiration, while word-initial {b d g} have short VOTs and no aspiration. For the long-lag categories {p t k} in word-initial position, Kuzla and Ernestus (2011: 153) report an average VOT of 53 ms, and 17 ms for the short-lag categories {b d g} in that same position; Jessen and Ringen's (2002: 202–203) speakers' average VOT in word-initial [d] was 13.1 ms; Lisker and Abramson (1964), Keating (1984), and Ouddeken (2018) place the boundary between the short- and long-lag categories around 20–30 ms. There is, however, a lot of between-speaker variation in VOT: Hullebus, Tobin and Gafos (2018) found that in a sample of 25 native speakers of German, average VOTs for word-initial [t] ranged from 42 to 116 ms, and from 39 to 131 ms for word-initial [k].

The native speaker who recorded the stimuli in task B2 pronounced each of the six syllables {pa, ta, ka, ba, da, ga} eight times. The recording was made in a sound-proof booth, using a Sennheiser MKH 105 T microphone. The average VOTs of the syllables are given in Table 4.8:

Table 4.8. Average VOTs per category of the stimuli in task B2.

	[p]	[t]	[k]	[b]	[d]	[g]
VOT (ms)	95.47	93.89	94.08	-108.89	-56.84	-21.39

These averages are fairly extreme, compared to earlier sources on VOT in German plosives; this may be due to the recording setting, as the speaker read aloud a list containing only the six syllables of interest, which perhaps caused her to hyper-articulate. Nevertheless, this enhanced contrast is likely to minimise perceptual confusion during the learning phase of the experiment. Before task B2 started, the spoken stimuli were assessed in a listening experiment in the computer programme Praat (Boersma & Weenink 2018), in which four natives of German had to categorise all stimuli in a forced-choice experiments with the orthographic labels <pa>, <ta>, <ka>, <ba>, <da>, <ga>, <ma> and <sa> as response categories. Stimuli were presented in random order; each individual stimulus was presented three times, so the participants heard 144 tokens. All four listeners categorised all tokens correctly.

4.4.2 Method

Experiments A and B were run in ED, a freeware alternative to E-Prime (Vet 2013). In each individual task, participants were trained on one of the eight category structures from Figure 4.7: they were exposed to tokens of the categories that appeared in their input type.

In tasks A1 and B1, a photo of a handshake token was shown on the computer screen for 2000 ms; after 2000 ms, a “Next” button appeared under the photo, which the participant had to click to proceed to the next stimulus. The stimuli were presented in random order. Each individual token appeared in the input three times, so in experiment A, a learner saw 30 tokens of each category, and 24 tokens in task B1. A type 4B learner thus saw a total of 120 photos in task A1, and 96 photos in task B1. The number of photos shown per category, rather than the total number of photos, was kept constant between participants in an attempt to reduce the risk that experiments with smaller types would be too monotonous while learners of larger types would receive too little input to perform the task accurately.

After exposure, a test screen appeared showing eight categories; in tasks A1 and B1 these were photos of the six possible handshakes plus two control handshakes. The controls were not possible handshakes of the artificial sign language: in one, the thumb and index fingers created a circular shape while the remaining fingers pointed up (used in English-speaking countries to convey the meaning ‘okay’), in the other the little finger and thumb pointed up, the remaining fingers down. In task B2, the categories were orthographic representations of the six possible syllables, with <ma> and <sa> serving as controls. Juxtaposed to each of the eight categories in the test screen were sliders, whose leftmost position was labelled “not at all” (or in German: “gar nicht”), and whose rightmost position was labelled “very often” (in German: “sehr oft”). Participants were asked to carry out a frequency estimation task by adjusting the positions of the eight sliders to indicate the relative frequency with which each of the categories had appeared in the input. There were no ticks along the sliders, in order to avoid a preference for the values associated with these ticks. The initial position of the slider was random. Participants had to adjust the positions of all eight sliders; only then could their output be registered. Participants who responded to a control category with a non-zero frequency were excluded from analysis and replaced with a new participant.

In both experiments, participants were first familiarised with the task in a training phase, using pictures of geometric shapes as stimuli. An average experiment, including the training phase and a short debriefing, lasted between 15 and 30 minutes (depending on type size). Both experiments were approved by the Ethics Committee of the Faculty of Humanities of the University of Amsterdam. Participants in experiment A were recruited among colleagues, family and friends; participants in experiment B were recruited through a mailing list of the psycholinguistics laboratory of the Heinrich Heine University in Düsseldorf, Germany, and

through the distribution of flyers at that university. This experiment was administered in a sound-proof booth. In tasks A1 and B1, only adults without any prior knowledge of any natural sign language were allowed to participate; in task B2, only native speakers of German participated. The same participant pool performed tasks B1 and B2; to avoid a learning effect, task B1 was presented before task B2 in 48 participants, and in the reverse order in the remaining 48 participants. In experiment B, tasks were counterbalanced for inventory size and complexity.

4.4.3 Average misestimation as a measure of learnability

To assess the learnability of a type, I assess its error score, or the average misestimation in the frequency estimation task. The scale on which participants indicated estimated frequency was discretised in 100 steps of equal size. The left-hand end of the slider was assigned the value 0, the right-hand end was assigned the value 100; participants did not see these values, only the labels “not at all” (or in German: “gar nicht”) and “very often” (or in German: “sehr oft”). The highest indicated frequency of any category, if lower than 100, was scaled up to 100, and the other frequencies were adjusted accordingly. The lowest indicated frequency, if higher than zero, was not set to be zero, as this value has an absolute interpretation: zero means that the participant has indicated not to have seen or heard a category at all. For all six categories, the difference between rounded estimated frequency and input frequency was computed (the latter being either 0 or 100, for categories that are absent or present in the input, respectively).

Table 4.9 shows an example of frequency scaling and the computation of an error score. These are the responses of a fictional type 6D learner. The highest raw estimated frequency is 94; this value is scaled up to 100, and the other indicated frequencies are multiplied by $\frac{100}{94}$ as well (rounded off to two decimals in the table; only unrounded scores were used in the analysis of the experimental results). The difference between the input frequency and the scaled estimated frequency is the misestimation. In the table, I refer to the handshapes as feature combinations composed of two numbers: firstly the handshape feature, given as the number of fingers pointing up (0, 1 or 4), and secondly the thumb opposition feature, expressed as a Boolean variable (0 = thumb unopposed, 1 = thumb opposed). The error score is the average misestimation per input category, that is, the sum of all six numbers in the fifth column (in Table 4.9: 46.81) divided by the number of categories in the input (for a type 6D learner: five). The error score of our fictional learner, then, is 9.36.¹⁴ Error scores are expected to be correlated with complexity: learners are likely

¹⁴ This calculation differs slightly from the one in Seinhorst (2016a, 2017), where the sum of the misestimations was divided by six. I adjusted this because the size of the parameter space in a type is not always six, and because it is easier to compare the average misestimation per input category, as used in this chapter, between types.

to obtain higher error scores on more complex types (i.e. those with low E values or high lc values).

Table 4.9. *An example of frequency scaling and the computation of the error score.*

handshape ¹⁵	estimated frequency		input	misestimation
	raw	scaled	frequency	
/0 0/	94	100.00	100	0
/0 1/	80	85.11	100	14.89
/1 0/	83	88.30	100	11.70
/1 1/	91	96.81	100	3.19
/4 0/	12	12.77	0	12.77
/4 1/	90	95.74	100	4.26

4.4.4 Match as a measure of learnability

Although the error scores will provide insight into the learnability of a type, they do not tell us what *kind* of errors learners make. For instance, do learners indicate having seen only the categories in their input, or do they add or eliminate categories? It is not straightforward to answer these questions, because the response variable is continuous, not categorical. This property of continuity can be very informative, because variation in the estimated frequencies may be indicative of uncertainty in the learner, which we could not gauge if the response variable were categorical: for instance, if a learner is unsure whether she has seen a certain category, she might respond with a “not seen” in a categorical task while she might have responded with a low but non-zero value in the current task. However, if we want to answer the question whether categories have been added or removed, this uncertainty needs to be dichotomised somehow. I consider a participant to have added a category if that category did not appear in their input, yet they assigned it a scaled estimated frequency of 25 or higher; I consider a participant to have removed a category if that category appeared in their input, yet they assigned it an estimated frequency of 0. These criteria allow me to determine (mis)matches between learners’ input and output types in a categorical way, and hence determine observed transition probabilities between types as well as any complexity differences between learners’ in- and output. Applying this criterion to the example learner from Table 4.9, we conclude that this learner replicated their input type correctly: they indicated having seen a handshape that was not in their input, but the scaled estimated frequency of this category, namely 12.77, does not exceed the threshold value of 25.

¹⁵ The notation in the table suggests that these categories exist at the level of the Surface Form rather than the Underlying Form, that is, they are allophones rather than phonemes; without a lexicon, this distinction cannot be made, but considering that the thumb opposition feature is not contrastive in all types, I err on the side of caution.

4.4.5 Analysis

In order to analyse the association between the error scores and the complexity measures, I carried out linear regressions with error score as the outcome variable, and type, number of categories, feature economy, and logical complexity as predictors. In order to analyse the association between the number of correct replications and the complexity measures, I carried out logistic regressions with error score as the outcome variable, and type, number of categories, feature economy, and logical complexity as predictors. All statistical analyses in this dissertation were performed with the software R (R Core Team 2019); the “lmerTest” package (Kuznetsova et al. 2017) was used for the mixed-effects models in Chapters 5 and 7.

I centered the range of predictors that are interval variables, such as E and lc , around 0; for binary categorical variables, I coded one value as -0.5, and the other as 0.5, depending on prior expectations about the direction of the effect. For instance, the range of lc values in the eight category structures is [1, 5], which is shifted to [-2, 2] in the models; and in the analysis of the effect of modality on error score (§4.8.1), the “modality” predictor has two values and error scores are expected to be higher in the speech task, so “handshape” is coded as -0.5 and “speech” as +0.5. When testing for the effect of one of the eight category structures, I coded this type as +7/8 and the other seven types as -1/8. By coding predictors in this manner, all reported effect sizes, confidence intervals and p values pertain to comparisons with the average of all groups, not with one subgroup.

4.5 Experiment A: task A1 (implicit learning of handshapes)

Forty-eight adults participated in experiment A, each of whom was exposed to one of the eight types from Figure 4.7.

4.5.1 Error scores

Table 4.10 reports the average error scores per type.

Table 4.10. *Average error scores per type in task A1.*

	type							
	3A	4A	4B	6A	6B	6C	6D	6E
average error	7.8	4.6	12.7	23.4	8.6	59.9	18.0	8.1

A significant effect of type on error score was found for type 6C: on average, the error scores of learners of this type were 47.98 higher than the error scores of the other seven types (95% CI 35.91...60.05, $p = 2.9 \cdot 10^{-10}$). No significant effect of any of the other types was found.

The number of categories in a type might be a viable predictor of error score, because it determines the length of an individual experiment: it is well possible that in longer experiments, participants' attention waned. However, the effect of inventory size turned out to be non-significant (point estimate of the effect -0.190 , 95% CI $-6.36\dots5.98$, $p = 0.95$).

Figure 4.10 plots participants' error scores as functions of feature economy and logical complexity. For the purpose of visualisation, the error scores are divided into bins of size 1: the lowest possible bar, overlapping with the dashed line indicating a zero error score, indicates scores in the $[-0.5, 0.5)$ range (so $[0, 0.5)$ effectively), the next bar indicates the $[0.5, 1.5)$ range, and so on. The width of a bar is proportional to the number of participants whose error scores lie in that bin.

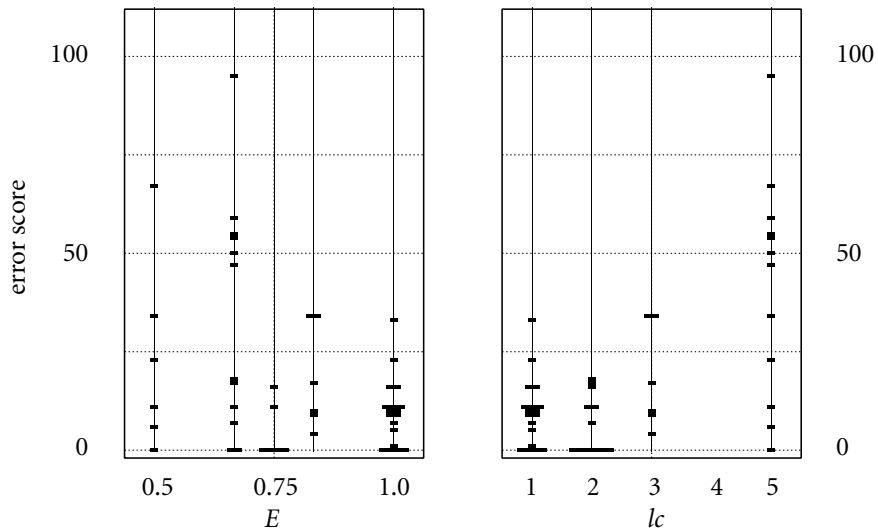


Figure 4.10. Plots of error scores in task A1 as a function of feature economy (left) and logical complexity (right). The width of a bar is proportional to the relative probability of that error score.

The learners in task A1 showed a significant effect of feature economy on error score: for every increase in E of 1, the error score decreased by 41.63 on average (95% CI $8.91\dots74.36$, $p = 0.014$; remember, however, that the range of E in the eight types was only 0.5. As expected, the direction of this effect is negative, with higher values of E being associated with lower error scores. These same subjects also showed a significant effect of logical complexity on error score: every increase in lc of 1 raised the error score by 8.39 on average (95% CI $5.39\dots11.40$, $p = 1.1 \cdot 10^{-6}$); as expected, the direction of this effect is positive.

Strikingly, none of the type 6C learners replicated their input correctly; they unanimously selected type 6E as output, assigning an average scaled frequency of 87 to the added categories, much higher than the threshold value of 25. The column “other” is needed because one type 6A learner indicated having seen only two signs; therefore, their output did not correspond to one of the eight types under investigation. Another type 6A learner removed one seen category and replaced it with an unseen one, effectively eliminating the thumb opposition contrast in their input. One type 6D learner filled the only gap in the system by producing a type 6E output. In total, two categories were eliminated, and 13 were added.

Perhaps a more insightful way to present these results is by considering all types as vertices in a directed graph. Such a graph might look like Figure 4.11, in which the three regular types are in the top row, and the two types with the highest logical complexity index are in the bottom row. Each type i is contained within a circle whose thickness is proportional to the observed transition probability p_{ii} , that is, the probability of correct replication (or “self-selection”); therefore, type 6C, in the bottom row of the graph, is not in a circle. Arrows between vertices i and j connect types with non-zero transition probabilities p_{ij} . The widths of the arrows are proportional to these probabilities.

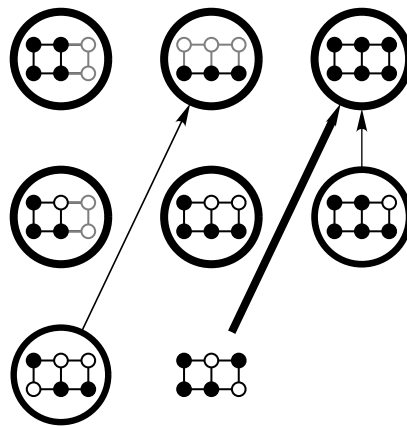


Figure 4.11. A graph showing observed transition probabilities between types in task A1. The thickness of the circle around a type is proportional to the probability that it is replicated correctly (or “self-selected”); the thickness of an arrow connecting two types is proportional to the observed transition probability between those types.

The three arrows in Figure 4.11 all point to regular types; most mismatches, except for the type 6A learner who reported two categories, favour types 3A and 6E. Since the regular types have high economy indices and low logical complexity indices, this regularising behaviour entails a considerable reduction of the complexity in the

entire data set; the sum of the lc indices of the output types will be lower than that of the input types, and the sum of the E indices of the output types will have increased. Table 4.13 compares the complexity measures of the in- and outputs: it lists the complexity indices of the eight types, the number of participants who learnt them (n_{before} ; the participant from the “other” column in Table 4.12 was not taken into consideration, so type 6A has only five learners); the number of participants who selected them after learning (n_{after}); and the contribution per type to the complexity measures before and after learning (lc_{before} , E_{before} , lc_{after} , E_{after}). Indeed, the total feature economy increased by 2.67 (7.0%) from 38 to 40.67: the point estimate of the average increase per participant equals $2.67/47 = 0.057$, which differs significantly from zero (95% CI 0.018...0.095, $p = 0.0048$). The total logical complexity decreased by 30 (26.1%) from 115 to 85: the point estimate of the average decrease per participant equals $30/47 = 0.638$, which differs significantly from zero (95% CI 0.212...1.064, $p = 0.0042$).

Table 4.13. *Summed complexity indices in participants’ in- and output in experiment A.* n_{before} = number of participants who received this type as input, n_{after} = number of participants who chose this type as output, E = feature economy, lc = logical complexity.

type	n_{before}	n_{after}	E_{before}	E_{after}	lc_{before}	lc_{after}
3A	6	7	6	7	6	7
4A	6	6	4.5	4.5	12	12
4B	6	6	6	6	6	6
6A	5	4	2.5	2	25	20
6B	6	6	4	4	12	12
6C	6	0	4	0	30	0
6D	6	5	5	4.17	18	15
6E	6	13	6	13	6	13
total	47	47	38	40.67	115	85

Participants reduced complexity unintentionally: they were not instructed to regularise, nor was there anything about the task itself that might lead learners to regularise. In fact, some participants remarked that the task was easy, while it turned out that they had unwittingly selected a regular type instead of their input type.

4.6 Experiment B: task B1 (implicit learning of handshapes)

The previous experiment was replicated with a larger sample of participants, consisting of 96 participants instead of 48.

4.6.1 Error scores

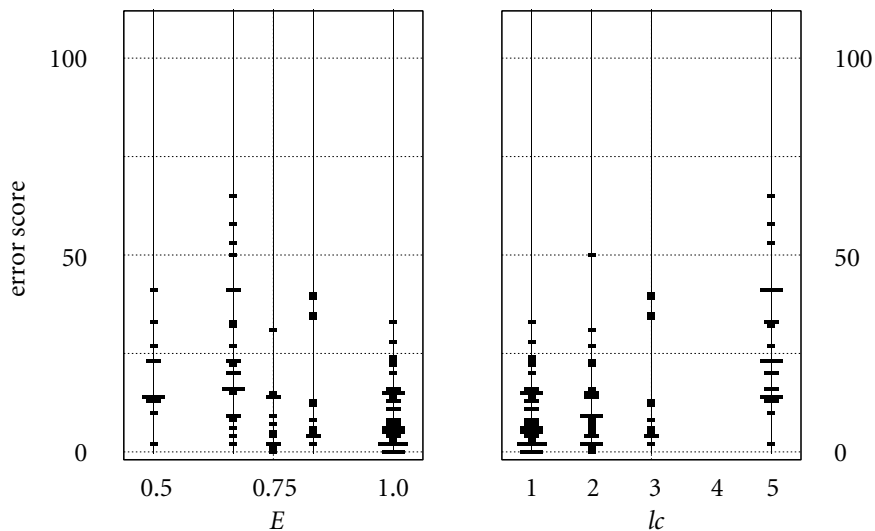
The average error scores per type in task B1 are reported in Table 4.14.

Table 4.14. Average error scores per type in task B1.

	type							
	3A	4A	4B	6A	6B	6C	6D	6E
average error	7.4	8.7	14.0	18.9	15.8	34.8	16.8	9.4

Two significant effects of type on average error score are found: the average error score of type 3A learners is significantly lower than the average of the other seven types (by 9.50, 95% CI 1.15...17.85, $p = 0.026$), and the average error score of type 6C learners is significantly higher than the average of the other types (by 21.75, 95% CI 14.43...29.07, $p = 5.8 \cdot 10^{-8}$). None of the other effects reached significance.

Figure 4.12 plots participants' error scores in task B1 as a function of the two complexity measures, in the same way as Figure 4.10.

**Figure 4.12.** Plots of error scores in task B1 as a function of feature economy (left) and logical complexity (right).

No significant effect of number of categories on error score was found (point estimate 0.0627, 95% CI -2.77...2.90, $p = 0.97$). Learners showed a significant negative effect of feature economy on error score (point estimate -25.74, 95% CI -40.92...-10.57, $p = 0.0011$), as well as a significant positive effect of logical complexity on error score (point estimate 4.21, 95% CI 2.64...5.78, $p = 7.0 \cdot 10^{-7}$).

4.6.2 Matches

Table 4.15 reports the proportions of matches in task B1. Of the 96 participants, ten produced a different type than they had learnt. In total, 14 categories were added, and none were eliminated.

Table 4.15. *Proportions of correct replications per type in task B1.*

	type							
	3A	4A	4B	6A	6B	6C	6D	6E
matches	1.0	1.0	1.0	0.92	0.83	0.5	0.92	1.0

In logistic regressions with match as an outcome variable, no significant effect of number of categories was found (average log odds $3.15 \cdot 10^{-16}$, 95% CI -0.63...0.71, $p = 1$). Feature economy does have a significant effect on match: an increase of E by 1 makes participants on average 187.54 times more likely to produce a match (95% CI 3.40...22,453 times, $p = 0.017$). Logical complexity also has a significant effect on match: for each increase in lc by 1, participants were on average 2.14 times more likely to respond with their input type (95% CI 1.38...3.77 times, $p = 0.0022$).

4.6.3 Complexity differences and regularisation

Table 4.16 lists the observed transition probabilities between types in task B1. All ten mismatches resulted in a type with a higher feature economy index than the input; eight reduced the logical complexity index, and two increased it.

Table 4.16. *Observed transition probabilities between types in task B1.*

	response type							
	3A	4A	4B	6A	6B	6C	6D	6E
input type 3A	1.0							
4A		1.0						
4B			1.0					
6A				0.92	0.08			
6B					0.83		0.17	
6C						0.5	0.25	0.25
6D							0.92	0.08
6E								1.0

Table 4.16 is visualised as a graph in Figure 4.13, again with arrows whose widths are proportional to the observed transition probability between the two connected types, and circles around types whose width is proportional to the probability of correct replication of that type.

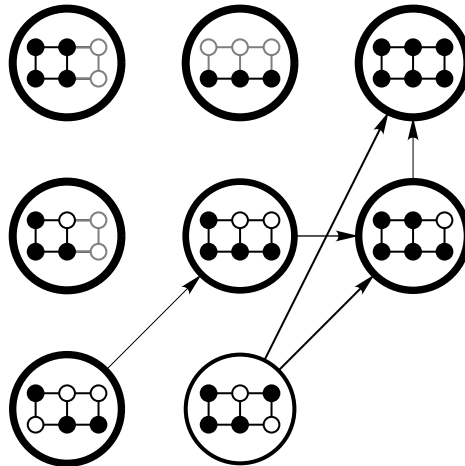


Figure 4.13. A graph showing observed transition probabilities in task B1.

The horizontal arrow between types 6B and 6D corresponds to the two mismatches that increased logical complexity; all other arrows point towards types with lower *lc* indices.

The total feature economy increased by 2.17 (2.8%) from 77 to 79.17: the average increase per participant is $2.17/96 = 0.0023$ (95% CI 0.0082...0.037), which differs significantly from zero ($p = 0.0024$). The total logical complexity decreased by 21 (8.8%) from 240 to 219: the average decrease per participant is $21/96 = 0.22$ (95% CI 0.044...0.393), which differs significantly from zero ($p = 0.015$).

4.7 Experiment B: task B2 (implicit learning of speech)

In this task, the stimuli were spoken syllables in Standard German. The same 96 participants from task B1 performed this task.

4.7.1 Error scores

Table 4.17 reports the average error scores per type in task B2.

Table 4.17. Average error scores per type in task B2.

	type							
	3A	4A	4B	6A	6B	6C	6D	6E
average error	32.6	27.1	22.1	28.0	30.7	23.0	29.7	22.2

The range of these averages is quite small, only just exceeding 10 (for comparison: in task A1 it was 55.3; in task A2 it was 27.4). None of the types had a significant effect on error score.

Figure 4.14 plots participants' error scores as a function of the two complexity measures, in the same way as Figures 4.10 and 4.12. Again, no significant effect of number of categories on error score was found: the point estimate of the effect size is -1.681 (95% CI $-5.346 \dots 2.075$, $p = 0.38$). No significant effect of feature economy on error score was found either: the point estimate of the effect size is -3.926 (95% CI $-25.281 \dots 17.429$, $p = 0.72$). Finally, no significant effect of logical complexity on error score was found: the point estimate of the effect size is -0.095 (95% CI $-2.480 \dots 2.290$, $p = 0.94$).

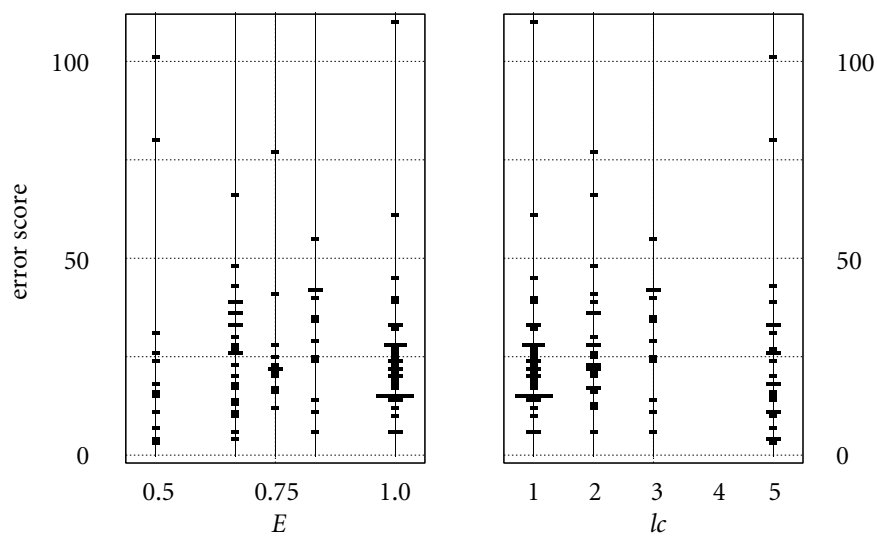


Figure 4.14. Plots of error scores in task B2 as a function of feature economy (left) and logical complexity (right).

4.7.2 Matches

Table 4.18 reports the proportions of matches and mismatches in task B2. Of the 96 participants, 25 produced a mismatch, introducing 27 categories and eliminating 5.

Table 4.18. Proportions of correct replications per type in task B2.

	type							
	3A	4A	4B	6A	6B	6C	6D	6E
matches	0.67	0.83	0.83	0.75	0.58	0.67	0.67	0.92

In logistic regressions with match as an outcome variable, no significant effect of number of categories is found (average log odds 0.1681 , 95% CI $-0.2904 \dots 0.6680$, $p = 0.49$). Feature economy does not have a significant effect on match either

(average log odds 1.2442, 95% CI -1.3400...3.8749, $p = 0.35$), nor does logical complexity (average log odds -0.1162, 95% CI -0.3985...0.1714, $p = 0.42$).

4.7.3 Complexity differences and regularisation

The observed transition probabilities between types in task B2 are given in Table 4.19. This table contains far more non-zero values than the comparable tables from tasks A1 and B1; note, for instance, that at least one learner of each type selected type 6D as their output. Many errors increased the logical complexity, and some learners even introduced feature values that did not appear in their input at all, for instance selecting type 6E as the output to a type 3A input.

Table 4.19. *Observed transition probabilities between types in task B2.*

		response type							
		3A	4A	4B	6A	6B	6C	6D	6E
input type	3A	0.67			0.08	0.08		0.08	0.08
	4A		0.83				0.08	0.08	
	4B		0.08	0.83				0.08	
	6A				0.75		0.08	0.08	0.08
	6B				0.08	0.58		0.33	
	6C		0.08				0.67	0.25	
	6D						0.08	0.67	0.25
	6E							0.08	0.92

Indeed, the corresponding graph reveals little structure in the errors:

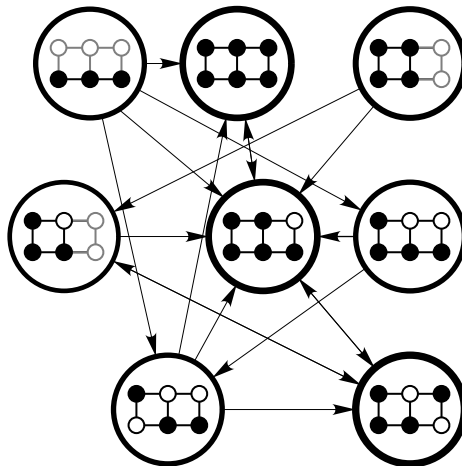


Figure 4.15. *A graph showing transition probabilities between types in task B2.*

The learners increased the total feature economy by 0.83 (1.1%) from 77 to 77.83; the average increase per participant is 0.0087 (95% CI -0.014...0.032), a difference that is not significantly different from zero ($p = 0.46$). The total logical complexity increased as well, contrary to expectation, by 4 (1.7%), from 240 to 244; the average increase per participant, whose point estimate is 0.042 (95% CI -0.176...0.260), is not significantly different from zero ($p = 0.71$).

4.8 Feature economy versus logical complexity

In this section, I compare the roles of feature economy and logical complexity as predictors of the results from the learning experiments.

4.8.1 Error scores

Table 4.20 repeats the effect sizes, 95% confidence intervals, and p values of number of categories, feature economy, and logical complexity on error score in all three tasks. In order to compare models, and thus determine which of the two complexity measures is the better predictor, the table also lists the AIC values (Akaike Information Criterion; Akaike 1974) of each model. In a comparison of AICs, the model with the lowest AIC should be considered the model that best fits the data; the difference in AIC between two models (Δ AIC) indicates the degree of information loss between them. A major advantage of AICs is that they can be computed for all sorts of regression models; this entails that I can use the same measure for all comparisons in this dissertation, and therefore the same criteria to determine which model should be considered the best. I follow Burnham and Anderson's (2004: 271) criteria for model selection: if Δ AIC ≤ 2 , the model with the higher AIC has "substantial support"; if $4 \leq \Delta$ AIC ≤ 7 , this model has "considerably less support"; and this model has "essentially no support" if Δ AIC > 10 .

Table 4.20. Summary of the results of tasks A1, B1, and B2: the effect of three predictors on error score.

task	predictor	effect size (95% CI)	p	AIC
A1	no. categories	-0.190 (-6.36...5.98)	0.95	433.57
	E	-41.63 (-74.36...-8.91)	0.014	427.18
	lc	8.39 (5.39...11.40)	$1.1 \cdot 10^{-6}$	408.50
B1	no. categories	0.0627 (-2.77...2.90)	0.97	782.96
	E	-25.74 (-40.92...-10.57)	0.0011	772.02
	lc	4.21 (2.64...5.78)	$7.0 \cdot 10^{-7}$	757.70
B2	no. categories	-1.68 (-5.35...2.08)	0.38	836.96
	E	-3.93 (-25.28...17.43)	0.72	837.63
	lc	-0.095 (-2.48...2.29)	0.94	837.76

The number of categories did not have a significant effect on error score in any of the experiments. In tasks A1 and B1, both feature economy and logical complexity had a significant effect on error score; in both tasks, the AICs of the logical complexity models are lowest, and Δ AIC exceeds 10 by far, making logical complexity the better predictor of error score in both tasks. The importance of *lc* over *E* as a predictor also becomes apparent in another way: in tasks A1 and B1, the addition of *lc* as a predictor to a model with *E* as a predictor yields a more significant improvement than adding *E* as a predictor to a model that already has *lc* as a predictor (in task A1: $p = 1.2 \cdot 10^{-6}$ and $p = 6.0 \cdot 10^{-3}$, respectively; in task B1: $p = 1.4 \cdot 10^{-4}$ and $p = 0.13$, respectively).

In task B2, neither feature economy nor logical complexity had a significant effect on error score; also, the small Δ AIC reveals that the goodness of fit of both models is comparable. Here, neither the addition of *E* as a predictor to the logical complexity model nor the reverse yields a significant improvement ($p = 0.70$ and $p = 0.74$, respectively).

4.8.2 Matches

Table 4.21 summarises the effects of number of categories, feature economy, and logical complexity on correct replication in all three tasks.

Table 4.21. Summary of the results of tasks A1, B1, and B2: the effect of three predictors on correct replication.

task	predictor	odds (95% CI)	<i>p</i>	AIC
A1	no. categories	1.152 (0.563...2.657)	0.71	50.19
	<i>E</i>	1335 (9.92...8.74·10 ⁶)	0.011	41.33
	<i>lc</i>	4.80 (2.23...19.88)	0.0018	25.71
B1	no. categories	1.00 (0.53...2.04)	1	68.15
	<i>E</i>	187.54 (3.40...22453)	0.017	61.42
	<i>lc</i>	2.14 (1.38...3.77)	0.0022	55.90
B2	no. categories	1.183 (0.748...1.951)	0.49	113.61
	<i>E</i>	3.470 (0.262...48.18)	0.35	113.22
	<i>lc</i>	0.890 (0.671...1.187)	0.42	113.47

Inventory size did not have a significant effect on match in any of the three tasks, and none of the predictors had a significant effect in task B2; in the tasks involving the learning of handshapes, both complexity measures yielded significant effects on correct replication. The AICs reveal that in tasks A1 and B1, the models with logical complexity as a predictor again have better fits than the models with feature economy, although Δ AIC for the models in task B1 was relatively small. In task B2, it is impossible to make a choice between the *E* and *lc* models.

4.9 The effect of modality

The comparison of predictors in the previous section seems to divide the three tasks by modality: in most models regarding the handshape learning tasks, logical complexity is the better predictor, while usually no difference can be found in the speech learning tasks. Indeed, the error scores in the speech-based task (B2) turn out to be significantly higher than those in the handshape-based tasks (A1 and B1), by 10.474 on average (95% CI 5.963...14.984, $p = 7.68 \cdot 10^{-6}$). In task B2, some learners added feature values that did not appear in the input, even though the spoken stimuli were categorised without errors by four native listeners (as described in §4.4.1); this did not happen in tasks A1 and B1, probably because learners had no previous experience with any of the handshape features and their values, and were therefore not aware of all possibilities.

In task A1, 9 learners produced a mismatch, that is, an output type different from their input type; 8 out of these 9 errors selected one of the three regular types. In task B1, 4 out of 10 errors were regularising; in task B2, 4 out of 25 errors were regularising. A Fisher's exact test reveals a significant effect of modality on the number of regularising errors ($p = 0.0018$), so learners regularise more often in the handshape tasks than in the speech task: the odds that a handshape learner regularises their input is 4768 times greater than the odds that a speech learner regularises theirs (95% CI 6.25... $1.82 \cdot 10^{21}$ times). This difference is also reflected in the difference in logical complexity between in- and output, which is significantly less negative in task B2 than in tasks A1 and B1 together (by 0.398, 95% CI 0.114...0.682, $p = 0.0062$), meaning that handshape learners reduce the complexity of their input to a larger degree than speech learners. Although handshape learners increase the feature economy of their input slightly more than speech learners, this effect is not significant (point estimate 0.0251, 95% CI -0.0019...0.0521, $p = 0.068$).

Any differences between modalities may be due to intrinsic modality-specific properties of the stimuli, but may also be ascribed to the fundamentally different tasks: in tasks A1 and B1, learners had to create categories before carrying out the experimental task, while in task B2, learners could perform the task relying on extant categories. The experimental design makes it impossible to decide between these explanations, but the latter hypothesis might be compatible with experimental evidence suggesting that children regularise more strongly than adult learners (Hudson Kam & Newport 2005, 2009; Culbertson & Newport 2015): perhaps this difference is not about age, but about the stage of language acquisition, where the added cognitive burden of category induction might lead early learners to regularise more. This effect has also been attested in late L2 learners (a.o. Meisel 2011; cf. also §§8.4.3–4).

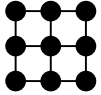
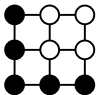
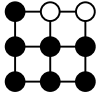
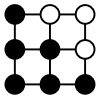
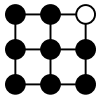
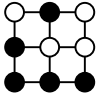
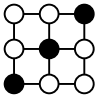
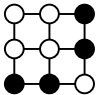
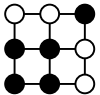
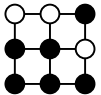
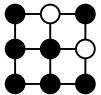
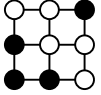
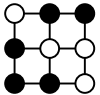
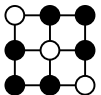
Phonological pattern learning (2): a 3×3 parameter space

The eight category structures from the previous chapter were informed by typological considerations. Each type represents one or more possible plosive inventories, and we can determine the frequency of occurrence of these inventories in a typological database (as I do in Chapter 7). However, if we want to tease apart the possible contributions of the two complexity measures to the types' learnability, the eight types are not ideally suited, considering the high degree to which these measures are correlated in the eight types. In the current chapter, I explore a larger parameter space, defined on two ternary features. This expansion allows for the definition of a larger number of category structures; in a subset of these, feature economy and logical complexity are barely correlated, which makes them well suited to explore questions of learnability. The new types are introduced in §5.1, and I present results from four tasks in a learning experiment: an implicit learning task, similar to the one from Chapter 4 (§5.2); a regularisation task with an iterated learning component (§5.3); a recognition task based on the card game SET® (§5.4); and a production task based on that same game (§5.5). Section 5.6 compares the three implicit-learning tasks from Chapters 4 and 5.

5.1 Expanding the parameter space: two ternary features

The category structures in Chapter 4 could be described with two features, one of which had a maximum of two contrasts, the other a maximum of three contrasts. An expansion of the parameter space to two ternary features allows for the definition of many more category structures: there are fourteen additional types with three or more categories. Visual representations of these types are introduced in Table 5.1 (next page), which also shows the types' feature economy and logical complexity indices: this visualisation is possible because each type has a unique combination of complexity indices, as opposed to the eight types in Chapter 4. Because each of the types uses all three contrasts on each feature, the denominator in the formula for the feature economy index is always 9, which entails that the number of categories in a type is perfectly correlated with its feature economy.

Table 5.1. *The fourteen new category structures, and their complexity indices.*

		feature economy								
		0.333	0.444	0.556	0.667	0.778	0.889	1.000		
1									1	9A
2									2	9B
3									3	9C
4									4	9D 9E
5									5	9F
6									6	9G 9H 9J 9K 9L
7									7	9M
8									8	9N
9									9	9P
		0.333	0.444	0.556	0.667	0.778	0.889	1.000		

The minimal formulas are listed in Table 5.2. The letters a and b represent the two features, as in Table 4.6 (p. 81); the possible feature values are $a_1, a_2, a_3, b_1, b_2,$ and b_3 . As in Chapter 4, I do not use negation to establish the complexity values.

Table 5.2. *Complexity indices of the additional category structures.*

type	minimal formula	lc
9A	A [all]	1
9B	$a_1 + b_1$	2
9C	$a_1 + b_1 + b_2$	3
9D	$a_1 + b_1 + a_2b_2$	4
9E	$a_1 + a_2 + b_1 + b_2$	4
9F	$a_1b_2 + a_2b_3 + b_1$	5
9G	$a_1b_1 + a_2b_2 + a_3b_3$	6
9H	$a_3(b_2 + b_3) + b_1(a_1 + a_2)$	6
9J	$(a_1 + a_2)(b_1 + b_2) + a_3b_3$	6
9K	$b_1 + b_2(a_1 + a_2) + a_3b_3$	6
9L	$a_1 + b_1 + a_2b_2 + a_3b_3$	6
9M	$a_1(b_1 + b_2) + a_2b_1 + a_3b_3$	7
9N	$a_1(b_1 + b_2) + a_2(b_1 + b_3) + a_3b_3$	8
9P	$a_1(b_1 + b_2) + a_2(b_1 + b_3) + a_3(b_2 + b_3)$	9

The expansion of the parameter space has been implemented in the set of handshapes by the addition of a “neutral” thumb opposition feature value, intermediate between the two values of the feature used in Chapter 4; in this “neutral” position, the thumb is positioned parallel to the index finger. A female signer was photographed producing each of the nine handshapes four times; tokens of the categories are shown in Figure 5.1 (next page). The neutral thumb opposition value is indicated as an empty feature value [\emptyset].

To further explore the roles that feature economy and logical complexity play in the learning of feature combinations, I conducted another experiment, to which I refer as “experiment C”. This experiment consisted of four tasks, which were administered in the order in which they are presented here; I call them tasks C1 through C4. A total of 84 subjects participated in experiment C, although not all of them performed all four tasks; the number of participants is specified in the description of each individual task. The experiment was designed and run in the computer programme Praat (Boersma & Weenink 2018). Approval from the Ethics Committee of the Faculty of Humanities of the University of Amsterdam was obtained prior to the experiment. Subjects were recruited through the distribution of flyers in two buildings of the University of Amsterdam; because the stimuli in the experiment were handshapes, only subjects who were unfamiliar with any sign language were allowed to participate, the same criterion as in tasks A1 and B1.

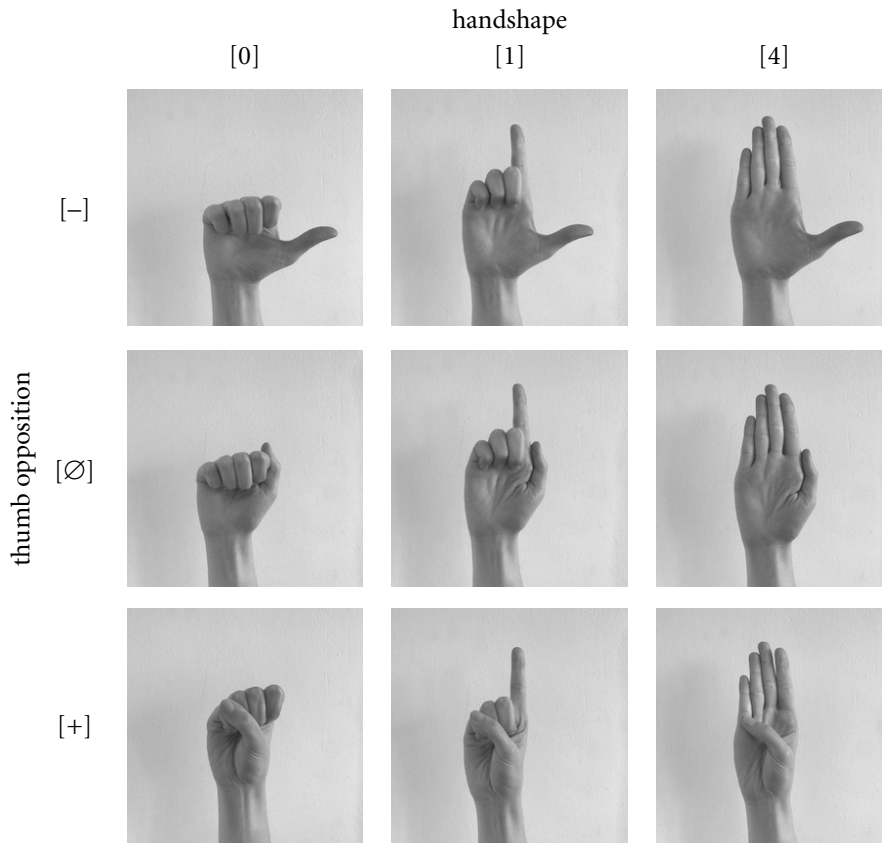


Figure 5.1. The expanded set of nine handshaped.

5.2 Experiment C: task C1 (implicit learning of handshaped)

The first task in experiment C is an implicit learning task, similar to the task in experiments A and B. Learners were exposed to handshape categories from one of the seven category structures in Figure 5.2. Types 9B, 9F, 9J and 9N have an *E* index of 0.556; types 9H, 9J, 9K and 9L have an *lc* index of 6. As a result of this plus-shaped selection, the correlation between feature economy and logical complexity in this set is close to zero ($\rho = 0.083$).

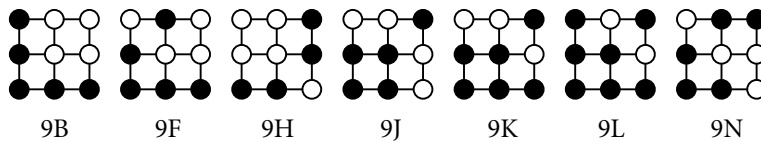


Figure 5.2. The seven category structures in task C1.

Each type was learnt by twelve participants. Multiple permutations exist of each type, of which one was selected at random in each individual experiment. The photos of the handshapes were presented in random order, and each handshape appeared in the input four times. After a photo had been shown for 1000 ms, a “Next” button appeared on the screen, which the participant could click in order to proceed to the next stimulus. After the exposure phase, a test screen appeared, in which learners reported for all nine possible categories, plus a control category, whether they had seen the category or not. The control category was a handshape in which the back of the hand faces the signer, with the index and middle fingers forming a V shape, sometimes referred to as the “victory sign”; none of the participants indicated having seen this sign. Before the experiment started, participants were familiarised with the procedure in a short practice phase, using geometric shapes as stimuli.

For each participant, two responses were collected: their response to the nine handshapes, and their reaction time. The response to the handshapes defines an output type, on the basis of which the values of three other variables can be determined, namely (mis)match between the in- and output types, the difference in feature economy between the in- and output, and the difference in logical complexity between the in- and output. Because participants’ responses to the handshapes were categorical rather than gradient (as they were in experiments A and B), the values of these variables could be determined straightforwardly. The reaction time is the time that elapsed between the moment the test screen appeared and the moment the learner clicked “Done” in that same window. Higher task complexity is expected to be correlated with longer reaction times.

5.2.1 Matches and regularisation

Out of the 84 participants, 39 produced the same output as their input. Table 5.3 shows the proportions of matches per type.

Table 5.3. *Proportions of correct replications per type in task C1.*

	type						
	9B	9F	9H	9J	9K	9L	9N
matches	0.67	0.5	0.42	0.5	0.25	0.5	0.42

In a logistic regression with match as outcome variable and feature economy as predictor, no significant effect was found (log odds: -0.636, 95% CI -5.121...3.763, $p = 0.78$); using logical complexity as a predictor instead, no significant effect was found either (log odds: -0.196, 95% CI -0.475...0.063, $p = 0.15$). Also, no significant effect of type on match was found.

None of the 45 mismatches were regularising; contrary to the learners in tasks A1 and B1, none of the learners in this task selected a regular type (in this task only

9A) as output. Nevertheless, learners did reduce the logical complexity of their inputs, by 0.512 on average (95% CI 0.153...0.871); this reduction differs statistically significantly from zero ($p = 0.0057$). Feature economy increased by 0.046 on average (95% CI 0.027...0.066); this increase differs statistically significant from zero too ($p = 9.6 \cdot 10^{-6}$).

5.2.2 Reaction times

Reaction times (RTs) usually do not follow a normal distribution; the right tail, with slower RTs, tends to be longer than the left tail (Luce 1986; Balota & Spieler 1999; Whelan 2008). To correct for this, researchers often transform the raw data, for instance by taking the reciprocal or a logarithm. In order to obtain a normally distributed RT measure, I transformed all RTs in this chapter in the way used by Lammertink et al. (2019) and Van Witteloostuijn et al. (2019): firstly, the n raw RTs were ranked from fastest (rank 1) to slowest (rank n); secondly, the ranks were scaled by first subtracting 0.5, and then dividing the outcome by n , yielding a set of values within the range (0, 1); thirdly, from these values, the quantiles of the normal distribution were computed. This transformation of raw RTs yields n boundary values of z for which the surface of the area under the bell curve of a standard normal distribution (with mean 0 and standard deviation 1) between $-\infty$ and z equals the scaled rank value. By way of example, Table 5.4 shows this transformation using some of the 84 RTs from task C1. The resulting RT quantiles range from -2.515 to 2.515.

Table 5.4. Transformation of the reaction times in task C1.

participant	reaction time (s)	rank	scaled rank	quantile
P01	19.157	19	0.2202	-0.7714
P02	14.226	8	0.0893	-1.3452
P03	34.737	60	0.7083	0.5485
...
P82	33.184	57	0.6726	0.4472
P83	27.975	39	0.4583	-0.1046
P84	28.320	40	0.4702	-0.0746

I refer to these n boundary values as “RT quantiles”; any reported effects reflect differences in these z values.

Figure 5.3 plots the RT quantiles, divided into bins of size 0.05 for purposes of visualisation, as functions of the two complexity measures. As in Figures 4.10, 4.12 and 4.14, the width of each bar is proportional to the number of learners in that bin. While the participants were slightly faster on inventories with higher E values (average RT quantile difference: 0.215, 95% CI -2.012...2.442), this effect was not significant ($p = 0.85$). Logical complexity, however, did have a significant effect on

RT quantile: for every increase of 1 in lc , the RT quantile went up by 0.136 on average (95% CI 0.0099...0.263, $p = 0.035$).

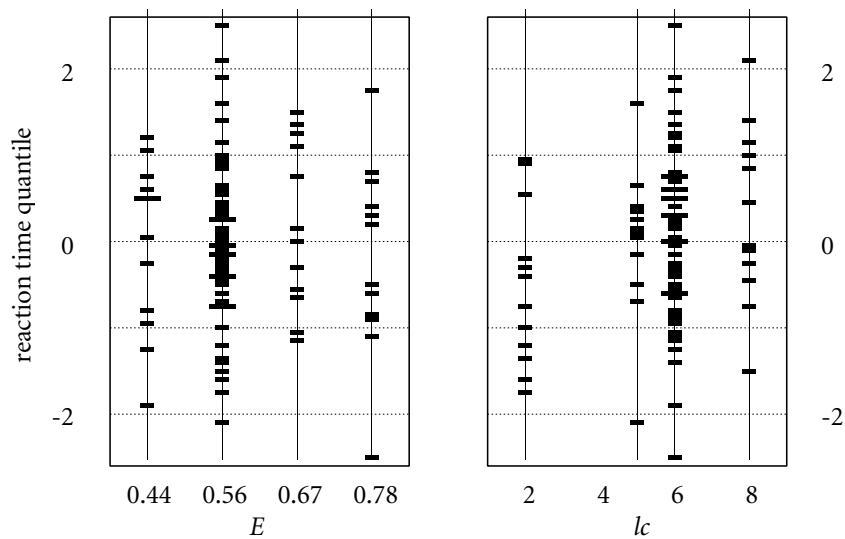


Figure 5.3. Plots of reaction times quantiles in task C1 as functions of feature economy (left) and logical complexity (right).

None of the seven types had a significant effect on RT quantile.

5.2.3 Logical complexity versus feature economy (again)

Table 5.5 lists the log odds, p values, and AIC values of feature economy and logical complexity as predictors of correct replication in task C1.

Table 5.5. Summary of the results of task C1: correct replication.

predictor	log odds (95% CI)	p	AIC
E	-0.636 (-5.121...3.763)	0.78	119.94
lc	-0.196 (-0.475...0.063)	0.15	117.84

According to Burnham and Anderson's (2004: 271) criteria for model comparison, cited in §4.8.1 (p. 100), the difference between the models is too small to consider one to be preferable over the other. This means that, even though the two predictors were nearly uncorrelated in this task, it cannot be determined which of the two better predicts the probability of a match in this task.

Table 5.6 lists the log odds, p values, and AIC values of feature economy and logical complexity as predictors of RT quantile in task C1.

Table 5.6. Summary of the results of task C1: RT quantile.

predictor	difference (95% CI)	<i>p</i>	AIC
<i>E</i>	-0.215 (-2.442...2.012)	0.85	243.07
<i>lc</i>	0.136 (0.0099...0.263)	0.035	238.52

Δ AIC is such that the *E* model has “considerably less” support than the *lc* model (Burnham & Anderson 2004: 271). Logical complexity, then, predicts RT quantile better than feature economy. This difference also becomes apparent because adding of *lc* to a model of RT quantiles with *E* as a predictor yields a significant improvement of the model ($p = 0.034$), while the reverse is not true ($p = 0.71$).

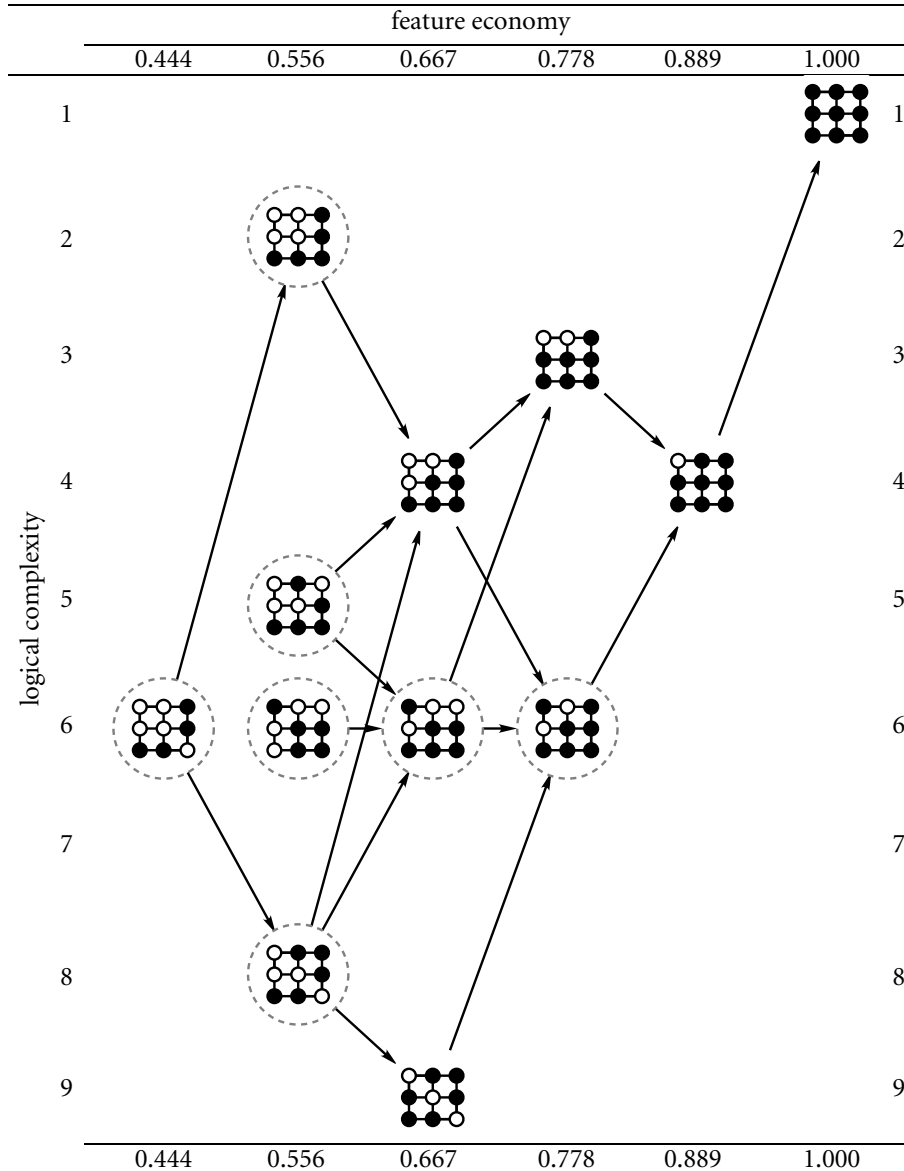
5.3 Experiment C: task C2 (regularisation in a diffusion chain)

The second task in experiment C was a regularisation task, intended to investigate how the regularisation process proceeds: when a new category is added, does the resulting inventory have the lowest complexity of all options? In this task, learners were shown the handshape categories in a type, and were asked which other category would fit those categories. Three kinds of responses were possible: participants could choose not to select a category at all, select the control category (the “V sign”), or select one of the remaining handshapes. For instance, if the input type contained five handshapes, the learner could choose one of four remaining categories. The number of categories that could be added was capped, because in a small-scale pilot study without such a maximum, learners would often regularise their input at once.

Because regularisation is a diachronic process and a maximum of only one category could be added, learners were grouped into diffusion chains, that is, the output of a learner served as the input to a following participant. The input types to the first generations of these chains were the same seven types from the previous task (cf. Figure 5.2). Table 5.7 shows the possible paths between types in this task. If an arrow points from type *i* to type *j*, then type *j* can be formed from type *i* by adding one category. The seven initial types are drawn in dashed circles. The figure shows that sometimes multiple types can be reached from one single type; sometimes a single type can be reached from multiple types; and sometimes the addition of a category necessarily increases the logical complexity of a type.

Each diffusion chain in task C2 consisted of seven generations, and each participant saw seven chains. Learners were assigned to these chains in a stepwise fashion: for instance, participant P08 formed the second generation of the second diffusion chain that started with type 9H, as well as the last generation in the first chain starting with type 9B, and so on. The assignment of learners to diffusion chains is shown in Table 5.8 (p. 112). The numbers “2; 3”, for instance, mean that a learner is the third generation in the second complete chain. If the first number in such a designation is zero, that chain does not reach its seventh generation.

Table 5.7. Paths towards regularity: an arrow from type i to type j means that j can be derived from i by adding one category. The initial category structures are in circles.



This stepwise assignment of learners to diffusion chains was chosen to ensure that a participant learned different stages of different chains; if a participant were the last generation in all their seven chains, all inventories in all chains had probably become regular. 63 subjects participated in this task, creating nine complete chains of type 9H, and eight complete chains of the other types.

Table 5.8. *Distribution of participants across diffusion chains in task C2.*¹⁶

participant	original type						
	9H	9B	9F	9J	9N	9K	9L
P01	1; 1	0; 1	0; 1	0; 1	0; 1	0; 1	0; 1
P02	1; 2	1; 1	0; 2	0; 2	0; 2	0; 2	0; 2
P03	1; 3	1; 2	1; 1	0; 3	0; 3	0; 3	0; 3
P04	1; 4	1; 3	1; 2	1; 1	0; 4	0; 4	0; 4
P05	1; 5	1; 4	1; 3	1; 2	1; 1	0; 5	0; 5
P06	1; 6	1; 5	1; 4	1; 3	1; 2	1; 1	0; 6
P07	1; 7	1; 6	1; 5	1; 4	1; 3	1; 2	1; 1
P08	2; 1	1; 7	1; 6	1; 5	1; 4	1; 3	1; 2
...
P55	8; 6	8; 5	8; 4	8; 3	8; 2	8; 1	7; 7
P56	8; 7	8; 6	8; 5	8; 4	8; 3	8; 2	8; 1
P57	9; 1	8; 7	8; 6	8; 5	8; 4	8; 3	8; 2
P58	9; 2	0; 1	8; 7	8; 6	8; 5	8; 4	8; 3
P59	9; 3	0; 2	0; 1	8; 7	8; 6	8; 5	8; 4
P60	9; 4	0; 3	0; 2	0; 1	8; 7	8; 6	8; 5
P61	9; 5	0; 4	0; 3	0; 2	0; 1	8; 7	8; 6
P62	9; 6	0; 5	0; 4	0; 3	0; 2	0; 1	8; 7
P63	9; 7	0; 6	0; 5	0; 4	0; 3	0; 2	0; 1

For each participant, the output type and reaction time were recorded in each chain.

5.3.1 Regularisation

All complete chains became regular in or within seven generations; for each type, Table 5.9 lists the first possible generation in which type 9A could be selected as output average generation, and the generation in which this happened on average. Once a chain had reached this state, learners never chose to add the control handshape, instead selecting the option not to add a category anymore. On average, chains became regular more slowly (i.e. in a later generation) than the minimum, because learners sometimes self-selected an irregular type.

Table 5.9. *The first possible generation in which a learner could select the regular output type (type 9A), and the average generation in which this happened.*

	original type						
	9B	9F	9H	9J	9K	9L	9N
minimum	4	4	5	4	3	2	4
average	4.75	4.63	6.11	5.00	3.75	2.50	4.88

¹⁶ The types are not in alphabetical order, because they were relabeled after the experiment had been designed.

Figures 5.4a–g (continues on page 114) show the paths that learners chose in the complete chains of each initial type, averaged over all complete chains of that type. The widths of the arrows are proportional to the observed transition probabilities between types. If a type is drawn inside a circle, then one or more learners of this type self-selected, that is, they chose not to add another category; the thickness of the circle is proportional to the probability of self-selection. The absence of a circle, then, means that all learners of that type selected a different output type. On the other hand, none of the learners of type 9A, the only regular type, chose to add the control category, so it is always circumscribed by a circle of maximum width. The width of an arrow between two types is proportional to the observed transition probability between those types. Dashed arrows indicate paths that are possible but unattested in the data. When multiple types can be reached from one type, these options have been ranked in the figure by their logical complexity values, the topmost one has the lowest lc index, and the bottom one has the highest. The figure reveals that learners did not always reduce complexity when they could.

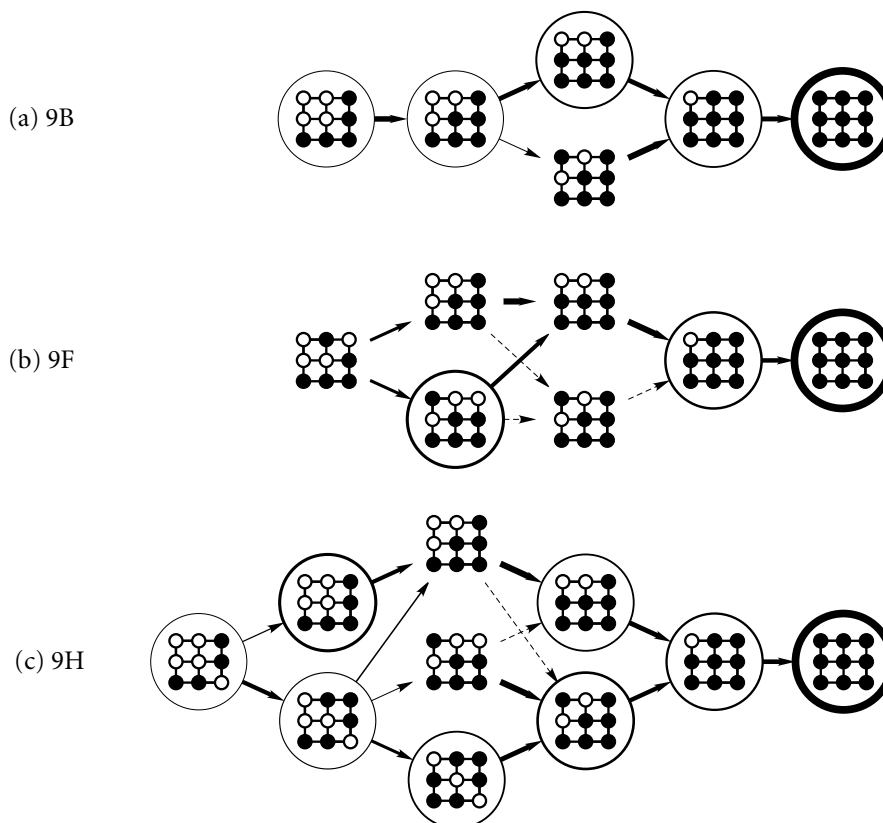


Figure 5.4. Attested paths towards regularity in task C2.

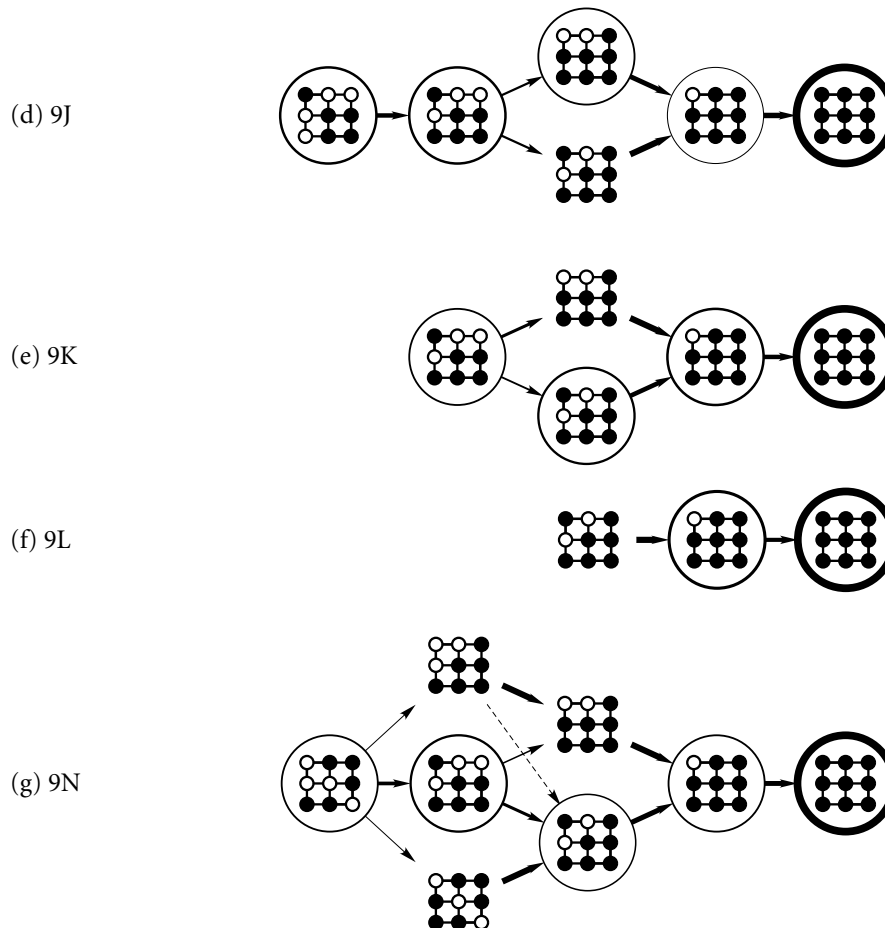


Figure 5.4 (continued). *Attested paths towards regularity in task C2.*

There are five types where the addition of a category can lead to different category structures, namely 9D, 9F, 9H, 9K, and 9N. To explore whether learners' choices between the possible category structures scores are motivated by considerations of complexity, I scored learners' output types as either optimal (i.e. having the largest reduction of lc) or non-optimal (i.e. not having the largest reduction of lc), and used these scores as a (binary) outcome variable in two mixed-effects logistic regressions, one with E as a between-subjects predictor, and one with lc as a between-subjects predictor. Participant served as a random within-subjects effect. There were 104 observations from 57 participants.

A significant effect of feature economy was found: for every increase of 1 in E , learners were on average 11381 times more likely to select the type with the lowest

complexity (95% CI 14.46...5.25·10⁷ times, $p = 0.012$). This effect is huge, as it its confidence interval; however, in these five types, the E values only range from 0.444 to 0.667. A significant effect of logical complexity was found as well, in the opposite direction: for every increase of 1 in lc , learners were on average 2.775 times *less* likely to select the type with the lowest complexity (95% CI 1.656...7.392 times, $p = 0.0023$). Therefore, learners are more likely to select the outcome type with the lowest lc for inputs with lower complexities (i.e. higher E values and lower lc values).

5.3.2 Reaction times

For each participant, the reaction time (RT) of each of their seven responses was recorded, defined as the time that elapsed between the moment the input type appeared in the window, and the moment the learner clicked “Done” in that same window. RTs were transformed as in §5.2.2 and served as the outcome variable in two mixed-effects linear regressions, both with participant as a random within-subjects effect, one with E as a between-subjects predictor, and the other with lc as a between-subjects predictor.

Significant effects were found for both complexity measures: for an increase of E by 1, RT quantile reduced by 2.281 (95% CI 1.846...2.717, $p = 4.3 \cdot 10^{-22}$), that is, learners responded faster on more economical inventories; for every increase of lc by 1, RT quantile increased by 0.129 (95% RT 0.0949...0.163, $p = 9.6 \cdot 10^{-13}$), that is, learners responded faster on more compressible inventories.

I do not assess the AIC values of models here, since the task simultaneously required the expansion of the inventory (and hence an increase of E) and the addition of the best fitting category (and hence a decrease of lc); the complexity measures both played a central role in this task, as opposed to the implicit learning tasks, and as a result their comparison would not be very insightful.

5.4 Experiment C: task C3 (classification of Sets)

The third and fourth tasks in experiment C are inspired by the card game SET®, described as a “game of visual perception”. In this game, each card contains a stimulus that can be described in terms of four ternary properties: colour (green, purple, or red), number (one, two, or three), shading (none, solid, or striped), and shape (lozenge, rounded rectangle, or squiggle): these properties define $3^4 = 81$ unique cards. A Set is defined as any triplet of cards on which the stimuli, for each of the four properties individually, either all have the same value, or all have a different value. For instance, a solid green lozenge, a solid green rounded rectangle, and a solid red squiggle do not form a Set: the numbers and shading agree across cards, and all the shapes differ, but two of the three stimuli have the same colour. A card with two striped purple lozenges, a card with two striped purple rounded rectangles, and a card with two striped purple squiggles do form a Set; a card with one striped

purple lozenge, a card with two unshaded green squiggles, and a card with three shaded red rounded rectangles form a Set as well. In the standard version of the game, twelve cards are laid out, and players need to find Sets as quickly as possible.

The $3 \times 3 \times 3 \times 3$ parameter space of the original game can be simplified to the 3×3 parameter space introduced in this chapter. I hereafter call two-dimensional Sets that differ in only one feature “type S1A” (with the 1 signifying “Setness”); consequently, I call two-dimensional Sets that differ in both features “type S1B”. Within this 3×3 parameter space, six different S1A types can be defined, and six different S1B types. Examples of both types are shown in Figure 5.5:

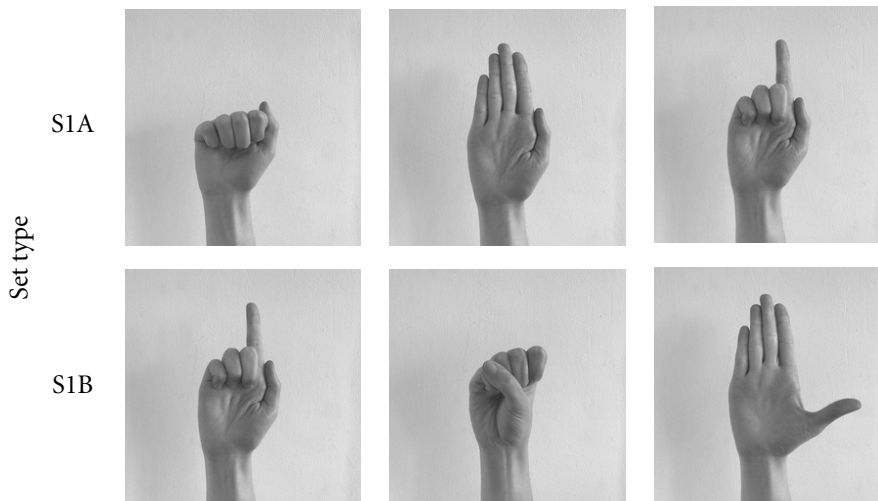


Figure 5.5. *Examples of Sets: triplets differing on one feature (type S1A), and triplets differing on two features (type S1B).*

Since there are 12 Sets, $\binom{9}{3} - 12 = 72$ triplets of categories exist that do not form a Set. Previous research has focused solely on the complexity of Sets but not non-Sets; however, non-Sets can also be further divided into two types. One type uses two values of each feature, and the other uses all three values of one feature, and two of the other. I call these types “S0A” and “S0B”, respectively. Of the 72 two-dimensional non-Sets, 36 are of type S0A, and 36 of type S0B. I refer to all four possible types together – S0A, S0B, S1A, and S1B – as “Set types”, even though types S0A and S0B are technically not Sets. Examples of types S0A and S0B are shown in Figure 5.6.

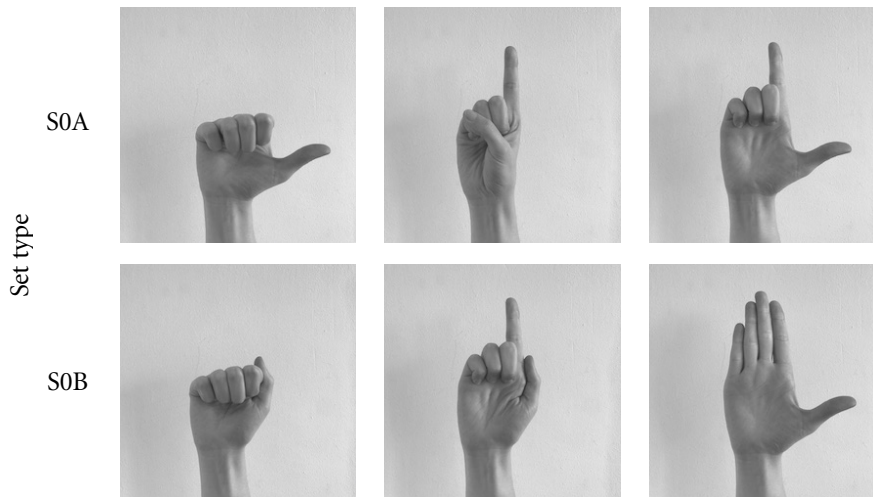


Figure 5.6. Examples of non-Sets: triplets using only two values of each feature (type S0A), and triplets using all three values of one feature and only two of the other (type S0B).

Figure 5.7 shows grid representations of all four Set types. One permutation of types S0A and S0B is shown: the three categories in an S0A triplet are connected through a solid line, those in an S0B triplet through a dashed line. All six permutations of type S1A and all six permutations of type S1B are shown. Because all four Set types fit within the 3×3 parameter space, they are also familiar category structures: type S0A corresponds to type 4A (see Figure 4.7, p. 79), S0B to 6A, S1A to 3A, and S1B to 9G (see Table 5.1, p. 104).

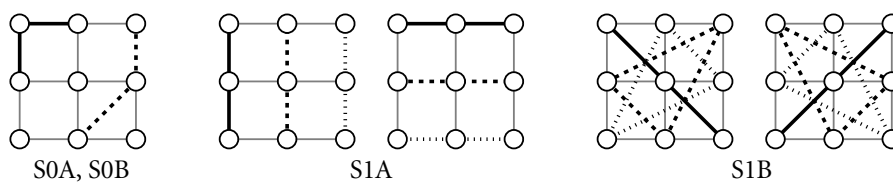


Figure 5.7. Grid representations of the four different Set types: non-Sets (one permutation of S0A is drawn with a solid line, and one of S0B with a dashed line) and Sets (all permutations of S1A and S1B, drawn with different lines).

Task C3 was a forced-choice classification task, in which participants had to classify triplets of handshapes as either a Set or not a Set. 71 learners participated. Subjects were asked if they were familiar with the game; before the task started, they were familiarised with the game in a practice phase, and they could repeat this phase until

they felt confident enough to start the task. Learners saw a total of 24 triplets of handshapes: all six S1A triplets, all six S1B triplets, and 12 randomly drawn S0 triplets. Sets and non-Sets were presented in random order. The order of presentation of the handshapes within the triplet was random as well. The displayed token of each category was also drawn randomly from the four available tokens.

Two responses were collected for every triplet: match, that is, whether the classification was correct, and reaction time. Earlier experiments investigating the effect of Set type on reaction time found that players recognised Sets with higher numbers of contrasts more slowly, and that players who were unfamiliar with the game recognised Sets more slowly than players who were already familiar with the game (Taatgen et al. 2003; Jacob & Hochstein 2008). I investigate these effects, plus the effect of Setness (is the input a Set or not? so S0AB vs. S1AB), and the difference between S0A and S0B non-Sets.

5.4.1 Matches

I carried out a mixed-effects logistic regression using correct classification as the (binary) outcome variable; predictors in this model were familiarity with the game (yes or no) and Set type, which was coded in such a way that the effects of Setness (S0AB vs. S1AB), of the difference between non-Sets (S0A vs. S0B), and of the number of relevant features within Sets (S1A vs. S1B) could be established. Participant served as a random effect; the model includes random slopes for Set type by participant, because different learners may not be equally sensitive to the effect of Set type.

Participants who were familiar with the game were on average 1.49 times more likely to categorise a triplet correctly than unfamiliar players; this difference is not significant (95% CI 0.631...3.53 times, $p = 0.36$). On average, participants were 3.34 times more likely to correctly categorise Sets than they were non-Sets; this effect is not significant (95% CI 0.522...21.33 times, $p = 0.20$) either. Learners were on average 3.28 times more likely to categorise S0A triplets correctly than they were S0B triplets; this effect is significant (95% CI 1.15...9.30 times, $p = 0.026$). Contrary to expectation, the odds of S1B triplets being categorised correctly was on average 1.58 times greater than the odds that S1A triplets are categorised correctly, but this difference is not significant (95% CI 0.35...7.03, $p = 0.55$). None of the interactions of familiarity and the four values of the Set type factor were significant.

5.4.2 Reaction times

Reaction times were transformed using the strategy from §5.2.2. The resulting RT quantiles were used as the outcome variable in a mixed-effects linear regression with the same predictors as the model from §5.4.1. On average, participants who were not yet familiar with the game were significantly slower than participants who were fami-

liar with it (average RT quantile difference: 0.42, 95% CI 0.17...0.66, $p = 0.0014$). No significant effect of Setness on RT quantile was found (average: 0.021, 95% CI -0.20...0.24, $p = 0.85$). As expected, learners were significantly slower in categorising S0B triplets than S0A triplets: the average RT quantile was 0.31 higher for S0B triplets (95% CI 0.18...0.44, $p = 7.3 \cdot 10^{-6}$). Within Sets, an effect of number of features on RT quantile was found: learners were slower on S1B types than on S1A types (average RT quantile difference: 0.41, 95% CI 0.28...0.53, $p = 1.15 \cdot 10^{-8}$). None of the interactions of familiarity and Set type were significant. The advantage of type S1A over S1B replicates earlier findings by Taatgen et al. (2003) and Jacob and Hochstein (2008); the advantage of type S0A over S0B can perhaps be explained by the lower complexity of the former type.

5.5 Experiment C: task C4 (production of Sets)

For every pair of handshapes, there is exactly one handshape with which that pair forms a Set. This can easily be verified in Figure 5.7, where any two categories are connected by only one line; the third category is found by following this line. The resulting Set may be of type S1A or S1B, depending on the relation between the original two categories. In task C4, the same 71 subjects from C3 participated; they were shown two handshapes, and were asked to select a third, thereby completing the Set. For each participant, two responses were recorded: match, that is, whether the resulting triplet did indeed form a Set, and reaction time.

5.5.1 Matches

A significant effect of familiarity on match was found: on average, familiar players were 2.66 times more likely to correctly complete a Set (95% CI 1.111...6.367 times, $p = 0.023$). Players were not significantly more likely to correctly complete S1B triplets than S1A triplets (average log odds difference: -0.0614, 95% CI -1.328...0.9821, $p = 0.91$); the interaction of familiarity and Set type was not significant either (average log odds difference: -0.5842, 95% CI -2.094...0.8744, $p = 0.42$).

5.5.2 Reaction times

Unfamiliar players were not significantly slower or faster than familiar players (average different in RT quantile: 0.2869, 95% CI -0.0194...0.5929, $p = 0.071$). Participants took significantly longer to complete S1B Sets than S1A sets (average RT quantile difference: 0.6303, 95% CI 0.5246...0.7360, $p = 6.7 \cdot 10^{-18}$), extending the results by Taatgen et al. (2003) and Jacob and Hochstein (2008) to the production direction as well. The interaction of familiarity and Set type was not significant (average RT quantile difference: 0.0529, 95% CI -0.1585...0.2650, $p = 0.49$).

5.6 Comparison of the implicit-learning tasks (A1/B1 and C1)

The tasks in experiment C explored a bigger parameter space than the tasks in experiments A and B. This space difference also came with complexity differences: the average logical complexity of the seven category structures in tasks C1 and C2 was higher than the average of the eight category structures in experiments A and B (by 3.071), and the average feature economy of the types in tasks C1 and C2 was lower than the average in experiments A and B (by 0.2148).

The three tasks A1, B1 and C1 all involve implicit learning, and for all three tasks, we have a “match” response (albeit a derived one in tasks A1 and B1). Learners in task C1 produced fewer matches (39 out of 84 participants) than learners in tasks A1 and B1 together (125 out of 144 participants); is this difference best explained by the lower E values in task C1, or the higher lc values? Or by differences in parameter space? Table 5.9 shows results from logistic regressions with match as outcome variable and the three measures as predictors. Because parameter space increases exponentially with each added feature, I take the binary logarithm of the parameter space size.

Table 5.9. Summary of the results of tasks A1, B1, and C1 together: the effect of three predictors on correct replication.

predictor	log odds (95% CI)	p	AIC
E	0.932 (0.638...1.226)	$2.4 \cdot 10^{-9}$	252.24
lc	-0.104 (-0.127...-0.081)	$1.7 \cdot 10^{-16}$	219.61
${}^2\log$ space	-0.401 (-0.504...-0.298)	$6.8 \cdot 10^{-13}$	236.07

The effects of all three predictors are significant. For an increase in 1 on the feature economy measure, learners are 2.54 times more likely to replicate their input type correctly (95% CI 1.89...3.41 times). The effect of logical complexity goes in the opposite direction: for every increase in 1 on the lc measure, participants are 1.11 times *less* likely to produce a match (95% CI 1.08...1.14). Increasing the ${}^2\log$ value of the parameter space by 1 makes a correct replication 1.49 times less likely (95% CI 1.35...1.66). Although all three predictors have significant effects, the magnitude of the differences in AIC between the models is such that we can assume lower numbers of correct replications to be explained best by higher logical complexity indices. The AIC of a logistic regression model with all three predictors is 217.82, so not low enough to prefer it over the lc -only model; this last comparison suggests that E and parameter space seem to play limited roles. I further discuss this comparison in §8.2.3 (p. 163).

Throughout these last two chapters, the effects of complexity on learnability varied, but learners reliably reduced the complexity of their input. In the following two chapters, I explore complexity in sound changes and sound systems.

*Let me take you on a trip
Around the world and back*

(Depeche Mode — World in my eyes)

Part III:
TYPOLOGY

Complexity in sound changes

The two chapters in this third part of the dissertation are dedicated to typological data: they compare the results from the experiments in the previous part to attested sound changes (this chapter) and sound systems (Chapter 7). This chapter is laid out as follows. I first discuss previous literature about the role of complexity in language change (§6.1), then flesh out the predictions from the experiments (§6.2), and finally assess complexity differences in four attested sound changes: §6.3 treats the consonants of Old English, Middle English and Modern English; §6.4 treats the vowels of Old English, Middle English and Modern English; §6.5 investigates the First Germanic Consonant Shift; §6.6 focuses on the (im)plosives of Zulu. In §6.7, I evaluate the results from §§6.3–6.6 in light of the research question.

6.1 Complexity in language change

In this section, I discuss competing factors in language change, including sound change. Does language change favour more complex systems, or less complex ones? This question may be impossible to answer, if only because a complexity-reducing change in one subsystem of a language can result in higher complexity in other subsystems, or vice versa. For instance, in early Old English the plural form of |mu:s| ‘mouse’ was /my:si/, with /y:/ existing only as a positional variant of |u:|, conditioned by the occurrence of the front vowel /i/ in the following syllable; at this time, no phoneme |y:| existed. Subsequently, the plural ending /i/ was dropped, and the singular and plural forms could only be distinguished by the vowel in the stem, raising /y:/ to phonemic status: that is, while /y:/ previously occurred only as a positional variant of /u:/, it became lexically contrastive after the loss of the plural suffix — it distinguished |mu:s| ‘mouse’ from |my:s| ‘mice’. While the loss of the plural suffix may have reduced the complexity of Old English plural morphology, it simultaneously added a new phoneme to the sound system, arguably making that system more complex.

The results from the Markov chain from §6.2.1 were presented earlier in Seinhorst (2016a); the sound change data from §§6.3–5 were published as Seinhorst (2016b). However, this chapter presents slightly different analyses than those sources.

6.1.1 Complexity-reducing linguistic change: compressibility

Kirby et al. (2015) argue that the evolution of language is driven by two competing forces: a tendency towards compressibility, and a pressure to communicate successfully. They define compression as “optimisation of a repertoire of signals such that the energetic cost of unambiguously conveying any meaning is minimised” (p. 88), which is a very broad definition; they take it to encompass various concepts such as combinatoriality (the property that smaller units can be combined into larger ones, obviating the need to store all possible combinations), regularity (a situation where there are no exceptions to the rules that generate a system, cf. §4.2.3), and logical complexity, but also articulatory ease. However, these various properties may be optimised in opposite directions, as Kirby et al. indeed concede. Defining compressibility more narrowly as minimal logical complexity, and expressivity as a unique relation between form (a structure) and meaning (a referent), they argue that compressible languages are easier to learn, but also less expressive, and therefore less suited for effective communication; communicatively effective languages, at least according to their definition, are holistic and incompressible. Using computer simulations and experiments with human learners, Kirby et al. show that cultural evolution results in languages that strike a balance between compressibility and expressivity. This kind of interaction between opposing forces of efficiency and clarity is familiar from §1.3.1 (p. 11), and from Chapter 3.

Lieberman et al. (2007) investigated 177 verbs that were irregular in Old English, and found that of those 177, 145 were still irregular in Middle English, and 98 are still irregular in Modern English (the authors treat regularity as a dichotomous property, but do not specify the decision criterion). They found a negative effect of token frequency on regularisation rate: on average, a verb that is n times more frequent in the CELEX database regularises \sqrt{n} times more slowly. A negative correlation between lexical frequency and regularisation rate was also reported by Bybee (2001: 181). An effect of register on the regularity of English verbs seems to exist: Gray et al. (2018) found significantly more regular variants of verbs (e.g. *wedded* vs. *wed*) in a sample of Tweets (between 2008–2017) than in a sample of scanned books on Google (2003–2008). However, the data sets from Lieberman et al. and Gray et al. suffer from a selection bias: in order to accurately establish the effect size of regularisation, one should not only look at irregular verbs becoming regular, but also at the reverse, namely regular verbs becoming irregular. This is what Augst (1975) did in his survey of strong verbs in Old High German, Middle High German and Modern German. He found that in the evolution from Old to Middle High German, eight verbs regularised, but also 19 verbs deregularised (p. 255). Berg (1998: 238) found 45 instances of verb regularisation in Jespersen’s grammar of English, and 21 instances of deregularisation. He reports that a chi-squared test with these numbers yielded a statistically significant result (indeed: $\chi^2 = 4.51$, $df = 1$, $p = 0.034$), concluding that regularisation is more frequent than deregularisation, but this result

teeters on the brink of significance: in a Fisher's exact test on the same numbers, the p value is just on the other side of the significance criterion (odds ratio 2.13, 95% CI 0.997...4.633, $p = 0.051$). A bigger problem is that the observations in these changes are probably not independent, in which case neither a chi-squared test nor a Fisher's exact test is applicable.

One possible cause of regularisation in natural language is the diachronic amplification of learning biases; language contact is another. Lupyán and Dale (2010) found that languages spoken in larger communities tend to have less complex inflectional morphology, probably because such communities experience more language contact than isolated communities, and contact situations involve late acquisition: remember from §4.9 that late learners tend to regularise more than early learners (but see also the criticism in §1.3.3). In a similar vein, Bentz and Winter (2013) established that languages spoken in communities with larger proportions of non-native speakers tend to have simpler nominal case systems. Using agent-based computer simulations, Dale and Lupyán (2012) found that morphological complexity is lower in societies with only adult learners than in societies with adult as well as infant learners; also, complexity is lower in larger communities, probably because of the larger number of interactions with adult learners. In language contact settings, not only late learners, but also native speakers may be involved in this simplification process: in foreigner-directed speech, native speakers tend to use simpler syntactic constructions, more high-frequency words, and more repetitions, in addition to exaggerated auditory cues such as vowel formants, vowel duration, and pitch (Long 1983; Rodríguez-Cuadrado, Baus & Costa 2018; Al Kendi & Kattab 2019; Rothermich et al. 2019). Complexity reduction, then, may be attributed to several factors.

6.1.2 Complexity-increasing linguistic change: communication

While Kirby et al. (2015) focus solely on the minimisation of polysemy as a measure of communicative success, natural languages tend to display many other traits that are probably beneficial in communication. One example would be agreement marking on verbs in languages with explicit subjects, such as English: in the clause *she reads* both the subject *she* and the ending *-s* of the verb signal the presence of a third-person referent. This doubling of information increases the chances that an intended meaning by the speaker is conveyed successfully to the listener, but while such redundancy is advantageous in terms of communicative success, it makes the relation between meaning and form less transparent (for a definition of the term “transparency”, as well as a distribution of redundant traits in a sample of thirty languages, see Hengeveld & Leufkens 2018). Dahl (2004) argues that long-distance (“non-linear”) phenomena, such as agreement marking, are usually absent from newly emerged languages such as creoles: they may enter a language as it matures, suggesting that languages may become more complex over time. In Nichols' (1992: 88)

analysis of possession marking in noun phrases and sentences in a sample of 184 languages, the morphological complexity values follow a more or less normal distribution, shying away from very low or very high values, rather than a power distribution in which many languages have a low complexity value and high values are rare (for an example of this latter type of distribution, see Figure 1.1 on p. 11). An example of a non-linear phonological property of a mature language, according to Dahl (2004), is autosegmental tone. Phonemic tone is also mentioned by Comrie (1992) as a property that was likely absent from early stages of human language, as well as phonemic nasalisation; McWhorter (1998) too asserts that prototypical creoles have no tone.

6.1.3 Sound change

The term “regularity” is often used in the context of sound change too, but there it means that the result of the change is predictable from its environment, without being restricted to certain lexical items. For instance, segment A > segment B (with “>” meaning ‘becomes’), but only before all instances of segment C. Dahl (2004: 157) argues that such regular sound change, often referred to as “Neogrammarian”, tends to reduce “phonetic weight”; this seems to be an umbrella term for both phonetic and phonological content, in the sense that a reduction of phonetic weight may entail, among other things, vowel centralisation as well as the loss of segmental material. In a similar vein, Langacker (1977: 102–103) argues that sound change tends to increase “signal simplicity”, which involves a tendency towards articulatory ease. However, Passy (1890) already stressed the importance of auditory contrast in addition to articulatory ease, as did many after him. For instance, Ohala (2009: 48) argues that sound change “weeds out similar sounding elements”, providing French vowels as an example: the inventory of nasal vowels in French was once as large as the set of oral vowels, but it has since reduced in size for reasons of auditory distinctiveness. Martinet (1955) and Boersma (1998) illustrate the interaction between articulatory and auditory considerations with many examples; both pressures, rather than minimisation of effort alone, were crucial in the simulations in Chapter 3.

Sound changes are usually described in terms of individual segments or natural classes; much less is known about the role that complexity may play in the change of entire sound systems. Regularity, or rather the maximisation of feature economy, is sometimes mentioned as a driving force in the typology of sound systems (De Groot 1948; Martinet 1955; Clements 2003), and if this is true it should be a driving force in sound change as well. Remember, for instance, Trask’s example from §1.1.2 (p. 4); a similar argument is made by Martinet (1955: 58). I test this hypothesis in this chapter (for sound changes), and in the following chapter (for plosive inventories).

6.2 Phonological change: inductive biases only

The previous chapters have supplied us with a little bit of evidence regarding the direction of language change. In all three implicit-learning experiments with novel stimuli in Chapters 4 and 5 (task A1: §4.5; task B1: §4.6; task C1: §5.2), learners significantly lowered the complexity of their input, providing evidence for an inductive bias towards regular systems. For tasks A1 and B1, we also have the observed transition probabilities available between all eight category structures from Chapter 4, repeated here as Figure 6.1.

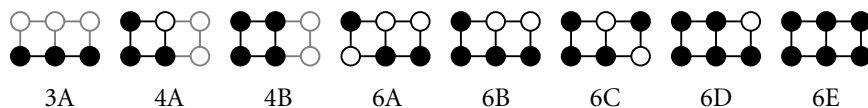


Figure 6.1. The eight plosive category structures from Chapter 4.

In this section, I explore the predictions that these results make concerning the direction of sound change.

6.2.1 Inductive biases only: cohort B1

The observed transition probabilities in task B1, presented in Table 4.16 (p. 96), are repeated here as Table 6.1 (only non-zero values are printed):

Table 6.1. Observed transition probabilities between types in task B1.

	response type							
	3A	4A	4B	6A	6B	6C	6D	6E
input type 3A	1.0							
4A		1.0						
4B			1.0					
6A				0.92	0.08			
6B					0.83		0.17	
6C						0.5	0.25	0.25
6D							0.92	0.08
6E								1.0

A table like this one, containing observed transition probabilities, can be considered a TRANSITION MATRIX. The rows of such a matrix constitute probability vectors, with non-negative entries that add up to 1. Each of the eight types can be seen as a possible state of a language. Languages can transition from one state to another; if such a change occurs, it is indicated by a non-zero outside the main diagonal. The transition probabilities in task B1 are shown as a transition matrix M_{B1} in (6.1):

$$(6.1) \quad \mathbf{M}_{B1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.92 & 0.08 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.83 & 0 & 0.17 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.92 & 0.08 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

A single cohort of learners, for instance each sample of participants in experiments A and B, can be considered as a generation in a vertical transmission chain, potentially producing a different language than the one they were exposed to. We can imagine a second generation, a new cohort of learners, who learn from the output of the first generation, and so on. Such a chain of transitions is memoryless: a type 6D inventory in generation t that derives from a type 6C inventory in generation $t-1$ is not expected to have any different properties than a type 6D inventory in generation t that derives from a type 6B inventory in generation $t-1$. This property is called the MARKOV PROPERTY, and a sequence of transitions like the one sketched here is a MARKOV CHAIN. Simplifyingly assuming that each generation of learners shows the exact same categorisation and production behaviour, and that the transition probabilities are the only determinant of language change, we can obtain the transition probabilities after t generations of learners by raising the transition matrix to the t^{th} power. Matrix (6.2) gives the transition probabilities between types after ten generations of B1 learners. Note that the rows still sum up to 1, rounding errors aside.

$$(6.2) \quad \mathbf{M}_{B1}^{10} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.42 & 0.26 & 0 & 0.25 & 0.08 \\ 0 & 0 & 0 & 0 & 0.16 & 0 & 0.51 & 0.32 \\ 0 & 0 & 0 & 0 & 0 & 9.8 \cdot 10^{-4} & 0.25 & 0.75 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.42 & 0.58 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

In order to understand this matrix better, let us focus on type 6A, the fourth row of the matrix. Remember from Table 6.1 (and Matrix (6.1)) that this type was replicated correctly by 92% of its learners; the other 8% produced a type 6B output. Of this latter type, 83% of the learners replicated their input correctly; the other 17% produced a type 6D output. Of the type 6D learners, 8% regularised their input to type 6E. All these changes are reflected in the fourth row of Matrix (6.2): 42% of the original type 6A languages are still of type 6A (fourth row, fourth column); some of the original type 6A languages are now of type 6B (26%: fourth row, fifth column); some languages have already gone through this last change and have proceeded to

become type 6D languages (25% of the original type 6A languages: fourth row, seventh column); and an even smaller proportion (8% of the original type 6A languages) have gone through all previous stages and have regularised completely (fourth row, eighth column).

As the number of iterations increases, so as $t \rightarrow \infty$, the transition probabilities will reach a STATIONARY DISTRIBUTION. The transition probabilities in the Markov chain based on Matrix (6.1) approach the values in Matrix (6.3):

$$(6.3) \quad \lim_{t \rightarrow \infty} \mathbf{M}_{B1}^t = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The matrix shows that all types with parameter space size 6 (occupying the bottom five rows in the matrix) eventually regularise, that is, the probability that they become type 6E languages, corresponding to the rightmost column in the matrix, approaches 1; because learners are inclined to add categories rather than remove them, they gradually fill up the available space in its entirety throughout the evolution of the inventory (except in type 4A).

While the values in the rows necessarily add up to 1, the values in the columns do not: these sums indicate the relative frequencies of the types. Over a large number of generations, then, the behaviour of the learners in task B1 would predict a typological pattern in which types 3A, 4A, 4B and 6E remain (that is, the columns that contain non-zeros), with type 6E being five times more frequent than each of the other three remaining types, and in which types 6A, 6B, 6C and 6D have become extinct (these are the columns that contain only zeros).

This prediction can be visualised in the same sort of graph seen before in Figures 4.11, 4.13 and 4.15. As in those figures, the thickness of the circle around a type i is proportional to the probability of self-selection, that is, the transition probability p_{ii} . In the graphs in Chapter 4, these circles could have different widths, but the stationary distribution, there are two possible widths, namely 0 (i.e. no circle) and 1. This reflects the fact that in matrix (6.3), only these two values occur. The thickness of an arrow between two types i and j is proportional to the transition probability between those types, that is, to p_{ij} ; here again, as opposed to Chapter 4, there are only two possible values.

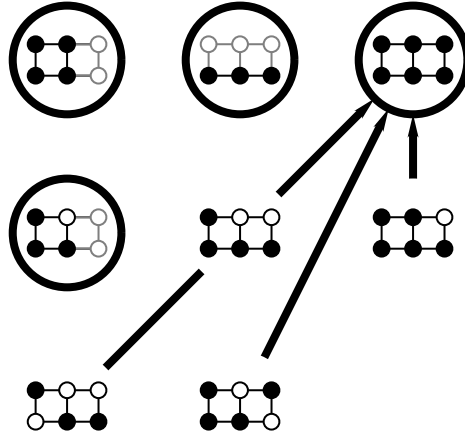


Figure 6.2. A graph showing transitional probabilities between types after the 50th generation of learners who show the same categorisation behaviour as the cohort of learners in task B1 (as in Figures 4.11, 4.13 and 4.15).

In Matrix (6.3), the predicted frequency of a type could be found by adding the numbers in the corresponding column; in Figure 6.2, it can be found by counting the number of incoming arrows and adding the probability of self-selection (either 0 or 1). Note that in the figure, there is partial overlap between the arrows from types 6A and 6B to type 6E: therefore, there are actually four arrows pointing to type 6E, not three.

6.2.2 Inductive biases only: cohort A1

In a Markov chain based on the categorisation behaviour of the learners in task A1 (Table 4.12, p. 92), the transition probabilities after a large number of generations come to approach the values in Matrix (6.4).

$$(6.4) \quad \lim_{t \rightarrow \infty} \mathbf{M}_{A1}^t = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

In this stable state, five types remain, namely 3A, 4A, 4B, 6B and 6E. All type 6A languages have changed to type 3A inventories; all type 6C and 6D languages have changed to type 6E inventories.

In graph form, these transitions to the stationary distribution look as follows:

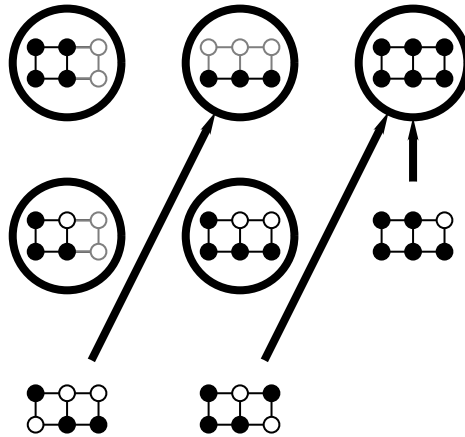


Figure 6.3. A graph showing transitional probabilities between types after the 50th generation of learners who show the same categorisation behaviour as the cohort of learners in task A1.

6.2.3 Inductive biases only: general predictions

Comparing Figures 6.2 and 6.3 (or Matrices (6.3) and (6.4)), we see that in the stable states predicted by tasks A1 and B1 all three regular inventories (3A, 4B and 6E) remain, as well as type 4A; additionally, the transition matrix from task A1 predicts the diachronic stability of type 6B. These types are **ABSORBING STATES**: once a language has reached such a state, it cannot attain any other state anymore. These states are easily recognised in a transition matrix because any absorbing state i meets two criteria: $p_{ii} = 1$ (implying that all other transition probabilities p_{ij} (with $j \neq i$) are 0); and at least one state j exists for which $p_{ji} > 0$.

In terms of the complexity measures, the generalisation seems to be that only inventories with low complexity remain. In both stable distributions, the highest possible value of lc has reduced from 5 to 2; the lowest possible value of E has increased from 0.5 to 0.67 in the A1 Markov chain, and from 0.5 to 0.75 in the B1 Markov chain.

6.3 Practice: attested phonological sound changes

The complexity-reducing behaviour of the cohorts of learners in Chapter 4 suggests that perhaps only those changes are possible that reduce the complexity of the sound system, and those that increase complexity are impossible or very rare. However, sound change obviously does not revolve only around learnability and complexity, but perceptual and articulatory pressures play crucial roles too. In order to test the hypothesis that sound change reduces complexity, I look at attested phonological sound changes (meaning those in which one or more featural representations within

the system change, instead of only phonetic properties, as was true for some of the sound changes in Chapter 3) and compare the complexity scores of the systems before and after the change.

I focus on three sound changes: that of Old English into Middle English into Modern English (§6.4), the First Germanic Consonant Shift (§6.5), and Zulu (§6.6). In the first of these three, I treat its obstruent inventory (§6.4.1) as well as its vowel system (§6.4.2). I use tables like Table 6.2, which shows the obstruents of Bisa (already familiar from Table 1.1 on p. 3). In many of the consonant tables in this section, plosives and affricates are given in the unshaded cells, fricatives are given in the shaded cells. In the vowel tables, short vowels are given in the unshaded cells, long vowels in the shaded cells.

Table 6.2. *Example table of an obstruent inventory: the obstruents of Bisa. Plosives and affricates are written in unshaded cells, fricatives in shaded cells.*

	[LABIAL]		[CORONAL]		[DORSAL]
[VOICELESS]	p	f	t	s	k
[VOICED]	b	v	d	z	g

The inclusion of fricatives introduces a manner contrast, and hence a third dimension, so this example table goes beyond the two-dimensional category structures from Chapters 4 and 5: if all contrasts were binary, the table would capture the Shepard types. However, this added dimension bears no effect on the way we establish the complexity indices of a system. For instance, the obstruent system of Bisa can be described as [PLOSIVE] + [LABIAL] + [CORONAL] and therefore has a logical complexity index of 3; its feature economy index is $10/(3 \cdot 2 \cdot 2) = 0.83$.

For the analysis pursued here, it is not sufficient to know that the sound change $A > B$ occurs frequently, or that in language X in the Yth century phoneme [C] turned into [D]; we need to know exactly in the context of what inventory this happened, in order to be able to establish the system complexities before and after. We thus need well-described phonological changes, if possible even with an analysis in terms of features. This prerequisite may be problematic, because a certain sound system can be analysed in more than one way; additional information about phonological behaviour may be needed to decide whether one analysis is to be preferred over another, but such information is not always available. However, because both complexity measures are feature-based, their values depend entirely on the chosen analysis. For the English data in §6.4, I follow Lass (2000). Some of his tables provide analyses where the features are hierarchically organised: for instance, in the Old English vowels (§6.4.2), the rounding feature can be combined with [FRONT] but not [BACK], because in Old English there was no contrast between unrounded and rounded back vowels. In Lass' analysis, the representation of the place feature is not hierarchical, even though the dental, alveolar, and postalveolar

places of articulation may be seen as further specifications of a single coronal place value (using such features as [anterior] and [distributed]; cf. Chomsky & Halle 1968; Clements & Hume 1995; and the discussion in Hall 1997). I follow Lass' analyses, and for the sake of uniformity, I do not use a hierarchical analysis of place in my analysis of the First German Consonant Shift either.

In the minimal formulas and tables, I use the following abbreviations: [ALV] for alveolar; [ASP] for aspirated; [DENT] for dental; [FRIC] for fricative; [GLOTT] for glottal; [IMPL] for implosive; [LAB] for labial; [LAB.VEL] for labiovelar; [PAL.ALV] for palatoalveolar; [PLOS] for plosive; [UNASP] for unaspirated; [VCD] for voiced; [VCL] or [VL] for voiceless; [VEL] for velar. I group affricates together with plosives; in the data in this chapter, no contrasts exist between them.

Neither complexity measure takes into account the substance of the features, meaning that they both disregard any influence from the phonetics; this influence is assessed in §6.7.1, and in §7.5.

6.4 Old English to Modern English

This subsection focuses on the development of Old English into Middle English into Modern English. I look at both the obstruent systems (§6.4.1) and the vowel inventories (§6.4.2) of these three stages.

6.4.1 Obstruents

Old English had the following obstruent system:¹⁷

Table 6.3. *The obstruents of Old English (Lass 2000: 70).*

	[LABIAL]	[DENTAL]	[ALV.]	[PAL.ALV]	[VELAR]
[PLOS, VOICELESS]	p	t		tʃ	k
[PLOS, VOICED]	b	d		dʒ	g
[FRICATIVE]	f	θ	s	ʃ	x

We see a recurrent pattern in this table: for all places of articulation except alveolar, Old English had voiceless and voiced plosives or affricates; for all places of articulation it also had a fricative category, which did not need to be specified for voicing. The minimal formula of this system is as follows:

[FRIC] + [PLOS] ∧ ([LAB] + [DENT] + [PAL.ALV] + [VEL])¹⁸

¹⁷ Lass assumes that [t] and [d] were dental, not alveolar; even if it were the other way around, the complexity counts would not change.

¹⁸ '∧' takes precedence over '+', so "[A] + [B] ∧ [C]" means "any segment with property [A], plus any segment with properties [B] and [C]".

The logical complexity index of this inventory is 6; its feature economy index is $13/(5 \cdot 3) = 0.87$.

The obstruent system of Middle English contained the new categories {v ð z}, introducing a voicing contrast (and concomitant gaps) in the fricatives too:

Table 6.4. *The obstruents of Middle English (Lass 2000: 71).*

	[LABIAL]	[DENTAL]	[ALV]	[PAL.ALV]	[VELAR]				
[VOICELESS]	p	f	t	θ	s	tʃ	ʃ	k	x
[VOICED]	b	v	d	ð	z	dʒ	ʒ	g	

Because new gaps were introduced, the logical complexity index of the system increased to 9:

$$[\text{LAB}] + [\text{DENT}] + [\text{ALV}] \wedge [\text{FRIC}] + ([\text{PLOS}] + [\text{FRIC}] \wedge [\text{VCL}]) \wedge ([\text{PAL.ALV}] + [\text{VEL}])$$

Although three new categories came into existence, the parameter space grew too, slightly reducing the feature economy index from 0.87 to 0.8.

As Middle English turned into Modern English, a gap was filled by [ʒ]. Also, [x] disappeared, leaving word-initial [h] as its only trace and introducing an additional place feature value [GLOTTAL]:

Table 6.5. *The obstruents of Modern English (Lass 2000: 71).*

	[LABIAL]	[DENTAL]	[ALV]	[PAL.ALV]	[VELAR]	[GLOTT]				
[VOICELESS]	p	f	t	θ	s	tʃ	ʃ	k		h
[VOICED]	b	v	d	ð	z	dʒ	ʒ	g		

The minimal formula of this system is as follows:

$$[\text{LAB}] + [\text{DENT}] + [\text{PLOS}] \wedge ([\text{PAL.ALV}] + [\text{VEL}]) + [\text{FRIC}] \wedge ([\text{ALV}] + ([\text{VL}] \wedge ([\text{PAL.ALV}] + [\text{GLOT}]))$$

This gives a logical complexity count of 10: the complexity-reducing regularisation in the supraglottal consonants has been counteracted by having unpaired [h]. The feature economy index of this system is $16/(6 \cdot 2 \cdot 2) = 0.67$.

6.4.2 Vowels

For the vowels, I only look at monophthongs. Table 6.6 lists those of Old English in Lass' (2000: 68) analysis, who does not make a height distinction between near-open [æ] and open [a:].

Table 6.6. *The monophthongs of Old English (Lass 2000: 68).*

	[FRONT], [UNROUNDED]		[FRONT], [ROUNDED]		[BACK]	
[CLOSE]	i	i:	y	y:	u	u:
[MID]	e	e:	ø	ø:	o	o:
[OPEN]	æ	a:				ɑ:

There are no gaps in the close and mid vowels, only in the open vowels. The minimal formula of this system is as follows:

$$[\text{CLOSE}] + [\text{MID}] + [\text{OPEN}] \wedge ([\text{UNROUNDED}] + [\text{BACK}] \wedge [\text{LONG}])$$

Therefore $lc = 6$; $E = 15/(3 \cdot 3 \cdot 2) = 0.83$.

The monophthongs of Middle English are presented in Table 6.7. The vowel height contrast is now quaternary; the rounding feature has become redundant.

Table 6.7. *The monophthongs of Middle English (Lass 2000: 68).*

	[FRONT]		[BACK]	
[CLOSE]	i	i:	u	u:
[CLOSE-MID]	e	e:	o	o:
[OPEN-MID]		ɛ:		ɔ:
[OPEN]		a:		ɑ:

All long vowels occur, as well as all close and close-mid vowels. The minimal formula of this inventory is quite short: $[\text{LONG}] + [\text{CLOSE}] + [\text{CLOSE-MID}]$. Therefore, lc is now 3; $E = 12/(4 \cdot 2 \cdot 2) = 0.75$.

Finally, let us turn to the monophthongs of Modern English:

Table 6.8. *The monophthongs of Modern English (Lass 2000: 69).*

	[FRONT]		[CENTRAL]	[BACK]	
[CLOSE]	ɪ	i:		ʊ	u:
[MID]	ɛ		ɜ:	ʌ	ɔ:
[OPEN]	æ			ɒ	ɑ:

Even in Lass' fairly simple representation, which lacks schwa as well as any distinctions between near-close and close or near-open and open vowels, we see a lot more gaps, suggesting that the vowel system has become more complex. Indeed it now has a logical complexity of 7:

$$[\text{FRONT}] \wedge ([\text{CLOSE}] + [\text{SHORT}]) + [\text{CENTRAL}] \wedge [\text{MID}] \wedge [\text{LONG}] + [\text{BACK}]$$

The feature economy index of this system is $11/(3 \cdot 3 \cdot 2) = 0.61$.

6.5 The First Germanic Consonant Shift

The transition of Proto-Indo-European (PIE) into Proto-Germanic is another excellent testing ground for the hypothesis that sound change reduces complexity, because it again provides us with not only an initial and a final stage, but also two intermediate steps. I assume that this set of changes, also known as the First Germanic Consonant Shift or Grimm's Law (named after Jacob Grimm, but already described by Danish linguist Rasmus Rask in 1818), is a set of pull chains, not push chains. In a pull chain, a segment or group of segments move/s towards a different, empty location in an auditory space, after which another segment or group of segments fill/s the newly vacated space; in a push chain, a segment or group of segments move/s towards a location that is already occupied, causing the already present segment(s) to move towards a new location too. Representing a sound change in separate steps, as I do here, entails treating those steps as pull chains.

For PIE, I assume the obstruent inventory in Table 6.9 (Lehmann 1952), with a voicing contrast in the plosives, an additional aspiration contrast in the voiced plosives, plus the alveolar fricative [s]. Because there is no voicing contrast in the fricative system, I assume that [s] has no voicing specification; I also assume that the aspiration feature is dependent on the [VOICED] feature value in the plosives, because aspiration was only contrastive in voiced plosives.

Table 6.9. *The First Germanic Consonant Shift: stage 1 (initial stage, before step 1).*

	[LAB]	[DENT]	[ALV]	[VEL]	[LAB.VEL]
[PLOS], [VCL]	p	t		k	k ^w
[PLOS], [VCD], [UNASP]	b	d		g	g ^w
[PLOS], [VCD], [ASP]	b ^{fi}	d ^{fi}		g ^{fi}	g ^{wfi}
[FRICATIVE]			s		

The minimal formula for this obstruent system is as follows:

$$[\text{PLOS}] \wedge ([\text{LAB}] + [\text{DENT}] + [\text{VEL}] + [\text{LAB.VEL}]) + [\text{FRIC}] \wedge [\text{ALV}]$$

Therefore, its logical complexity index is 7; its feature economy index is $13/(5 \cdot 4) = 0.65$. If PIE would not have had [s], the complexity of its obstruent system would have been 1; then again, the existence of a [FRICATIVE] feature value may have paved the way for the first step of the consonant shift, the spirantisation of the voiceless stops:

Table 6.10. *The First Germanic Consonant Shift: stage 2 (after step 1).*

	[LAB]	[DENT]	[ALV]	[VEL]	[LAB.VEL]
[PLOS], [UNASP]	b	d		g	g ^w
[PLOS], [ASP]	b ^h	d ^h		g ^h	g ^w h
[FRICATIVE]	ϕ	θ	s	x	x ^w

This system has voiceless fricatives for all its places of articulation, plus labial, dental, velar and labialised velar plosives (these are voiced, but voicing is not contrastive anymore):

[FRIC] + [LAB] + [DENT] + [VEL] + [LAB.VEL]

This inventory is less complex: $lc = 5$, $E = 13/(5 \cdot 3) = 0.87$.

In the second step of the consonant shift, the unaspirated voiced stops devolve: the pairs of plosive now differ in voicing and aspiration. For the sake of continuity, I analyse this difference as an aspiration contrast. This step only changes the substance of the features, the complexity measures are unaffected ($lc = 5$, $E = 0.87$):

[FRIC] + [LAB] + [DENT] + [VEL] + [LAB.VEL]

Table 6.11. *The First Germanic Consonant Shift: stage 3 (after step 2).*

	[LAB]	[DENT]	[ALV]	[VEL]	[LAB.VEL]
[PLOS], [UNASP]	p	t		k	k ^w
[PLOS], [ASP]	b ^h	d ^h		g ^h	g ^w h
[FRICATIVE]	ϕ	θ	s	x	x ^w

In the final step of the consonant shift, the aspirated stops become unaspirated, rendering the aspiration feature obsolete. I include here the segments that are explained by Verner's Law (Verner 1877), which states that fricatives immediately following an unaccented vowel in PIE became voiced in Proto-Germanic after the pitch accent system was lost. This gives the obstruent system in Table 6.12, with manner and voicing contrasts:

Table 6.12. *The First Germanic Consonant Shift: stage 4 (final stage, after step 3).*

	[LABIAL]		[DENTAL]		[ALV]	[VELAR]		[LAB.VEL]	
[VOICELESS]	p	ϕ	t	θ	s	k	x	k ^w	x ^w
[VOICED]	b	β	d	ð	z	g	ɣ	g ^w	ɣ ^w

This inventory can be described as follows:

[LAB] + [DENT] + [VEL] + [LAB.VEL] + [FRIC] \wedge [ALV]

This minimal formula yields a logical complexity index of 6; the feature economy increases considerably, to 0.9.

A certain degree of controversy surrounds the First Germanic Consonant Shift; many of the assumptions that I made in this subsection may be somewhat contentious. For instance, the assumption that the separate steps formed a pull chain was questioned by Kretschmer (1932) and Luick (1940). The initial inventory reported in Table 6.9 (Lehmann 1952) is sometimes analysed as having ejectives instead of voiceless stops, and voiceless stops instead of voiced ones, including the aspirates (Hopper 1973; Vennemann 1985), because it is unusual for an inventory to have aspirated voiced plosives but not aspirated voiceless ones (Martinet 1955; Jakobson 1958). Under this analysis, Verner's Law may have applied before Grimm's, as defended by Vennemann (1985) and Noske (2012); this last source offers a comprehensive discussion of these (and alternative) assumptions.

6.6 Zulu

The last test case here is the (im)plosive inventory of Zulu, with the data in Tables 6.13 and 6.14 taken from Clements (2003: 317). This inventory contains plosives, ejectives, and an implosive: disregarding the discussion in Clements (2003) about the feature specifications of these latter two classes across languages, I simply characterise ejectives as [EJECTIVE], and implosives as [IMPLOSIVE], which seems sufficient for Zulu. Neither of these classes have a voicing contrast, so I assume they do not carry a voicing specification.

The sound change described by Clements consists of two stages: in the first stage, about a century ago, Zulu had the inventory from Table 6.13, which can be summarised as follows:

$$[\text{EJECTIVE}] + [\text{VCL}] \wedge ([\text{ASP}] + [\text{DOR}]) + [\text{VCD}] + [\text{IMPL}] \wedge [\text{LAB}]$$

So $lc = 7$, and $E = 11/(5 \cdot 3) = 0.73$.

Table 6.13. *The (im)plosives of Zulu: former stage (Clements 2003: 317).*

	[LABIAL]	[CORONAL]	[DORSAL]
[EJECTIVE]	p'	t'	k'
[PLOS], [VCL], [UNASP]			k
[PLOS], [VCL], [ASP]	p ^h	t ^h	k ^h
[PLOS], [VCD]	b	d	g
[IMPLOSIVE]	ɓ		

The current inventory is as follows:

Table 6.14. *The (im)plosives of Zulu: current stage (Clements 2003: 317).*

	[LABIAL]	[CORONAL]	[DORSAL]
[EJECTIVE]	p'	t'	k'
[PLOS], [VCL], [UNASP]	p	t	k
[PLOS], [VCL], [ASP]	p ^h	t ^h	k ^h
[PLOS], [VCD]	b		g

The minimal formula of the current inventory is [EJECTIVE] + [VCL] + [LAB] + [DOR], so $lc = 4$; the table makes it fairly easy to see that $E = 11/(4 \cdot 3) = 0.92$.

This sound change, then, made the inventory less complex, which was exactly Clements' point, and although he and I compute feature economy in different ways, I agree with his conclusion. However, Zulu cannot have gone from the former to the current inventory in one step: there are too many differences between them. These differences are: (1) the implosive that existed in stage 1 does not exist anymore in stage 2; (2) the gaps in the voiceless unaspirated plosives were filled; (3) |d| existed in stage 1 but not anymore in stage 2. Clements cites a study by Louw (1962) suggesting a pull chain in which the voiced plosives devoiced before the implosive became a voiced plosive. This entails that |k| and |g| switched places, which is an unexpected turn of events, because |k| and |g| would have merged temporarily before splitting again: this situation may be explained by |k| being "marginal" and "largely restricted to affixes" (Clements 2003: 317), so the environments in which they occurred probably did not overlap. The intermediate stage would have looked as in Table 6.15. The minimal formula of this inventory is:

[EJECTIVE] + [VCL] + [PLOS] \wedge [DOR] + [IMPL] \wedge [LAB]

Its logical complexity index, therefore, is 6; its feature economy index is $11/15 = 0.73$.

Table 6.15. *The (im)plosives of Zulu (transitional stage).*

	[LABIAL]	[CORONAL]	[DORSAL]
[EJECTIVE]	p'	t'	k'
[VOICELESS], [UNASP]	p	t	k
[VOICELESS], [ASPIRATED]	p ^h	t ^h	k ^h
[VOICED]			g
[IMPLOSIVE]	ɓ		

In an alternative scenario, only {b d} devoiced, after which |ɓ| became |b|; however, this hypothesis would contradict the assumption that language change affects natural classes, an assumption that does hold in the attested scenario (cf. §1.1.2).

6.7 Evaluation

Figure 6.4 summarises the trajectories of both complexity measures throughout all four sound changes described in this chapter. The lc values are connected with a solid line; the E values are connected with a dashed line. Complexity-reducing sound changes, then, are those for which the solid lines go down or the dashed lines go up.

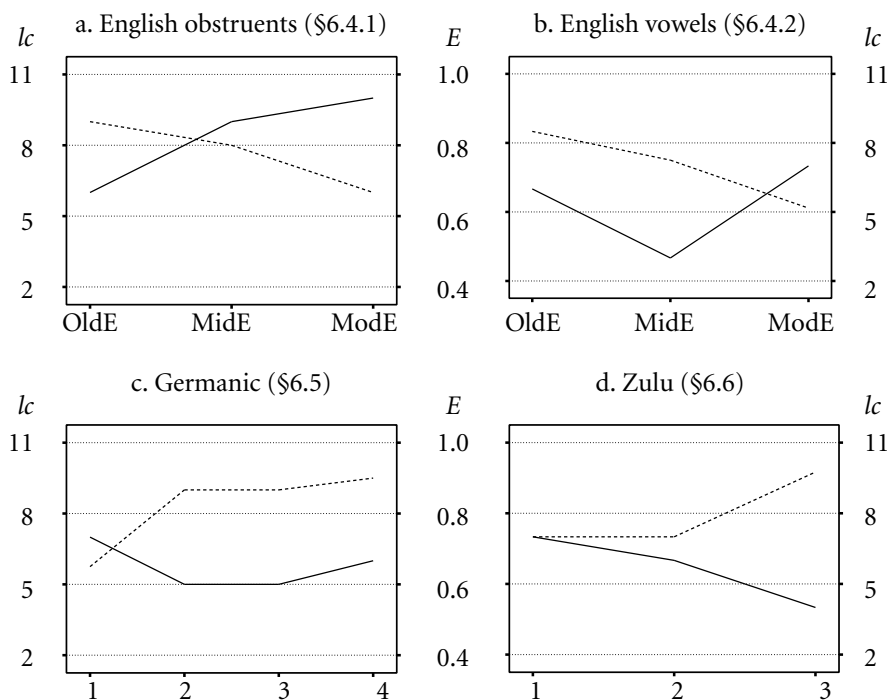


Figure 6.4. Plots of the complexity measures throughout the four changes described in this chapter. Logical complexity values are drawn with solid lines, feature economy values with dashed lines.

None of the sound changes were regularising; the highest value of E is 0.92 (the current (im)plosive system of Zulu), and the lowest value of lc is 3 (the vowels of Middle English). The results also show that sound change does not necessarily reduce complexity as quantified by these two complexity measures. Of course this sample is too small to do any sort of meaningful statistical testing; a larger sample would shed more light on the question whether sound change is complexity-reducing more often than complexity-increasing.

It bears repeating that both complexity measures depend entirely on the specifics of the formalisation in terms of phonological features. For instance, decisions on hierarchical relations between such features influence the values of the

measures. In Lass' analysis of the vowels of Old English, the roundedness contrast was dependent on the [FRONT] feature (cf. Table 6.6 on p. 137), probably because it was only distinctive in front vowels. Imagine an alternative analysis in which roundedness could have been combined with [BACK] too, as in Table 6.14.

Table 6.14. *The monophthongs of Old English (alternative analysis).*

	[FRONT], [UNROUNDED]		[FRONT], [ROUNDED]		[BACK], [UNROUNDED]		[BACK], [ROUNDED]	
[CLOSE]	i	i:	y	y:			u	u:
[MID]	e	e:	ø	ø:			o	o:
[OPEN]	æ	a:				ɑ:		

A possible argument for such an analysis could be that a contrast between rounded and unrounded back vowels exist in some languages, such as Turkish; however, the tables in this chapter are established on a language-specific basis. Another argument could be that the minimal formula should also tell us which categories are *not* part of the language, and if we do not exclude unrounded back vowels explicitly, we might erroneously infer that Old English had a vowel like [u]; I return to this question in §8.2.1 (p. 159).

In this analysis, *lc* would remain stable at 7:

$$[\text{UNROUNDED}] \wedge ([\text{FRONT}] + [\text{OPEN}] \wedge [\text{LONG}]) + [\text{ROUNDED}] \wedge ([\text{CLOSE}] + [\text{MID}])$$

However, *E* would be reduced greatly, because there are now six additional unused feature combinations: it would equal $15/(3 \cdot 2 \cdot 2 \cdot 2) = 0.63$ instead of 0.83, as it was in §6.4.2.

6.7.1 The role of phonetics

As I mentioned in §6.3, the complexity measures do not take into account the substance of the phonological features involved (§2.8.1, p. 47), which entails, among other things, that they do not distinguish between feature combinations that are deemed impossible to articulate and combinations that are not. Similarly, these measures cannot capture a difference in the voicing contrast between alveolar and velar plosives, even though the perceptual confusion between the former is much smaller than between the latter, which has been the probable cause of sound changes in many languages (Ohala & Riordan 1979; Ohala 1983; cf. Boersma 1998: 384–386 for a discussion of the various strategies that languages have employed to resolve this perceptual confusion). This insensitivity to substance also entails that we cannot make any predictions about the direction of a sound change that leaves complexity unaffected: depending on the inventory in which the change takes place, $|s| > |h|$

might not be preferred over $|h| > |s|$, even though the former change is fairly frequent while the latter has not yet been attested (Kümmel 2007; Honeybone 2016).

Nevertheless, the relevance of these phonetic pressures is clearly visible in the sound change data. If Proto-Indo-European, for instance, would have wanted to reduce its complexity, it could easily have done so by simply abolishing $|s|$, or by implementing a contrast between dental and alveolar stops, in which case $|s|$ would not be as much of an exception; however, such a fine-grained distinction between dental and alveolar plosives is probably difficult to produce and to perceive. The consecutive steps in the First Germanic Consonant Shift also neatly illustrate how only sufficient contrast was maintained: for instance, once the voiced unaspirated plosives had devoiced (stage 3), it had become unnecessary for the voiced aspirated plosives to maintain their aspiration, so this feature was dropped (stage 4). Similarly, it is telling that of all the features that Zulu could get rid of, it chose its [IMPLOSIVE] feature. This tendency towards optimal dispersion is also reflected in the number of features that is needed to describe an inventory: in the analysis presented here, PIE had binary manner, voicing and aspiration contrasts in stage 1, but Proto-Germanic retained only manner and voicing contrasts. This reduction of the number of necessary features is another perspective on complexity reduction, to which neither feature economy nor logical complexity is very sensitive; counting the number of features in an inventory is a definition of phonological complexity pursued by Moran and Blasi (2014). Changes in the substance of the necessary features are another way to look at complexity in sound change. Complexity in sound systems is the topic of the next chapter.

Complexity in sound systems

In 1921, Sir Richard Paget wrote that “[i]n speech, man, absurdly, used unvoiced sounds as well as voiced, though the carrying power and musical and emotional quality of unvoiced speech are very inferior to those of voiced speech” (p. 172). This is indeed true for vowels, but the growing availability of typological data has made it clear that the presence of voiced consonants in a phoneme inventory strongly implies the presence of unvoiced ones, while the reverse is not true. In this chapter, I compare the roles that feature economy and logical complexity play in the typology of attested plosive systems in spoken languages. After a brief discussion of previous literature on the complexity of sound systems (§7.1), I explain the analysis pursued in this chapter in §7.2, and present its results in §7.3; I then compare these results to the experimental findings from Chapter 4 in §7.4, and discuss the implications of this comparison in §7.5.

7.1 Definitions of complexity

The complexity of sound systems can be defined in various ways. A fairly straightforward way is to count segments, either consonants, vowels, or both (Hockett 1955; Maddieson 1980, 1984, 2007; Moran & Blasi 2014), or compute the consonant-to-vowel ratio; segment counts have also often been used as part of correlational studies, for instance with population size, as Lupyan and Dale (2010) did for morphological complexity (cf. §6.1.1). Hay and Bauer (2007) found such a correlation in their sample of 216 languages, but they neglected to correct for language family in the analysis (cf. §7.3.1), and the sizes of the language families in their sample varied greatly; Atkinson (2011) also found a significant correlation in his sample of 504 languages. Pericliev (2004) did not establish a significant correlation in his sample of 428 languages, and neither did Donohue and Nichols (2011) in a larger database of 1350 languages. In 969 languages from the PHOIBLE database, Moran, McCloy and Wright (2012) did find a significant effect, but its size is so small that they consider it uninteresting.

In this chapter, we only establish frequency distributions of complexity measures; we do not compute their correlations with possible explanatory factors. In that sense, this chapter is more similar to Coupé, Marsico and Pellegrino (2009),

The UPSID data and the comparison of complexity measures (§§7.2–3) will be published as Seinhorst and Van de Leur (under review).

who drew “phonological graphs” for 451 sound systems: these graphs contained the shortest distances in terms of features between the segments in these sound systems. The complexity measure was based on the average distance between the segments. Moran and Blasi (2014) counted the number of binary features that are needed to define the inventories in PHOIBLE; they chose features rather than phonemes following Kabak’s (2004) argument that phonological acquisition involves the acquisition of contrasts. This definition of complexity is similar to how Dahl (2004: 45) defines “conceptual complexity”, namely as the number of semantic features needed to describe a meaning. The number of features needed to define a phoneme inventory, however, is not necessarily indicative of the number of segments in that inventory: all five types 6A–6E (cf. Figure 4.7, p. 79) are identical in terms of the necessary feature values, but the number of categories in them ranges from three to six. Indeed, Moran and Blasi also look at feature economy, and show that languages tend to be uneconomical (p. 234). Clements (2003) tested the hypothesis that languages tend to recombine their features, for instance by counting pairs of segments that share properties with other segments within the same inventory (“attractors”) and segments that do not (“isolated sounds”); he found that attractors occur significantly more often than chance, while isolated sounds occur significantly less often than chance. Clements’ findings were replicated by Coupé, Marsico and Pellegrino (2017).

7.2 Feature-based analyses of sound systems

The remainder of this chapter presents data from the 317 sound systems in the first edition of the UCLA Phonological Segment Inventory Database (UPSID: Maddieson 1984), more specifically their plosive sets. For reasons of practical feasibility we chose the first, smaller edition of UPSID, rather than the extended database with 451 languages (used by Coupé, Marsico & Pellegrino 2009, mentioned above). We focused on plosives exclusively, because all spoken languages seem to use them, meaning that we can use the database in its entirety. We only analysed supralaryngeal segments, thus excluding the glottal stop; we also excluded affricate segments and stops with a double place of articulation.

In the discussion of analyses of sound systems in §6.3, I mentioned that Lass (2000), in his analyses of English, does not divide coronal segments further into smaller subgroups such as dentals, alveolars, alveolopalatals, palatals and retroflexes, contrary to a number of scholars; instead, he treats all places of articulation on par with each other. Maddieson does so too, and we follow his analyses, as I did with Lass’ in Chapter 6. This results in relatively “flat” representations, in which an inventory {p t k}, with a labial, a coronal and a dorsal, has the same structure (and the same complexity indices) as an inventory {t k q}, with a coronal and two dorsals. We treated secondary articulations, such as palatalisation or velarisation, as indepen-

dent of the place feature values, meaning that we assumed secondary articulations to be able to combine with all possible place features. We consider the palatalisation feature as rather abstract and separate from the place feature, because it affects different places in different ways: in labials, palatalisation entails an added articulatory gesture, while in velars it merely fronts the location of the constriction. Our decision to consider secondary articulations independently affects both complexity measures equally: it may generate more possible but unused feature combinations, thus lowering E , and in the calculation of lc it may require the specification of the place value(s) that the secondary articulation combines with, thus lengthening the minimal formula. From the resulting matrices, we computed the complexity indices of all 317 languages. Consider the following (fictional) plosive inventory:

Table 7.1. *A fictional plosive inventory.*

	[LABIAL]	[ALVEOLAR]	[RETROFLEX]	[DORSAL]
[VOICELESS]	p	t	ʈ	k
[VOICED]	b	d		

This inventory has a four-way place of articulation contrast and a binary voicing contrast. Of these eight possibilities, six are used, so $E = 0.75$. The shortest description of this inventory is “[VOICELESS] + [LABIAL] + [ALVEOLAR]”, so its logical complexity index is 3.

All languages in UPSID have place contrasts. If languages implement one additional contrast, it is usually laryngeal. In a few cases, other features are used, such as gemination (e.g. in Maranungku, mentioned in §4.3.1, and Delaware), or secondary articulations: for instance, Ahuslay contrasts plain and velarised segments, and Cheremis contrasts plain and palatalised segments. Chipewyan has plain, velarised and palatalised stops; we did not analyse this contrast as two separate binary articulatory contrasts (plain vs. palatalised, and plain vs. velarised), but rather as three possible values of a single contrast, based on the auditory continuum of F_2 (cf. Padgett 2001, 2003 about the velarised–palatalised contrast in Russian).

7.3 Complexity indices of plosive inventories in UPSID

In this section, we look at the frequency distributions of complexity values in UPSID. Figure 7.1 shows how feature economy is distributed in the database.

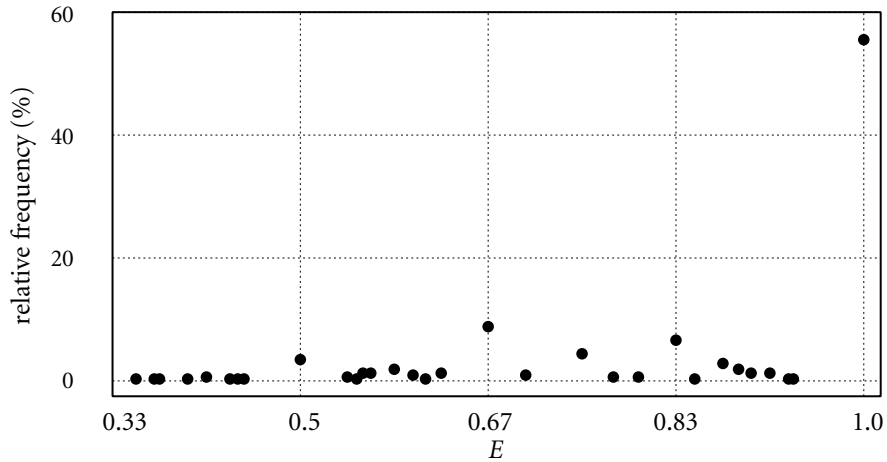


Figure 7.1. Relative frequencies of feature economy indices in UPSID.

Clements (2003: 289, 290; 2009: 28) and Moran and Blasi (2014) already noted that many sound systems are not fully economical; the conclusion from Chapter 6 that sound change is not necessarily regularising complies with this observation. Indeed, this frequency distribution follows a more or less exponential function, reminiscent of the relative frequencies of the 250 cross-linguistically most frequent phonemes from Figure 1.1 (p. 11): while a tight majority of languages in the sample has $E = 1$ (namely 176 out of 317 languages, or 55.5%), the values of E range from 0.355 to 1. Within this range, we see slightly higher frequencies for such values as 0.5, 0.67, 0.75 and 0.83, which result from common denominators.

Figure 7.2 shows how logical complexity is distributed in the database.

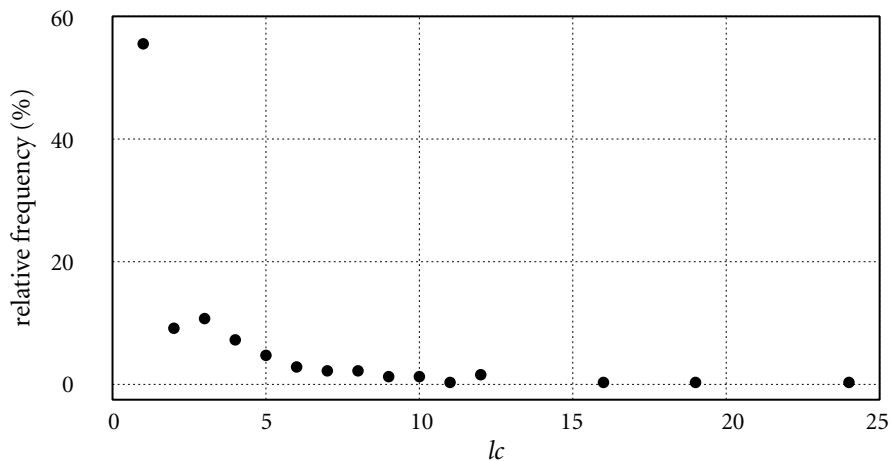


Figure 7.2. Relative frequencies of logical complexity indices in UPSID.

This distribution also seems to follow a power law, perhaps more clearly than the feature economy values — after all, it lacks the bumps in the distribution of E values, corresponding with certain fractions. The same 176 languages with $E = 1$ also have $lc = 1$, and higher values of lc tend to get increasingly less likely. The highest value found in the sample is 24.

7.3.1 The role of language family

Figures 7.1 and 7.2 do not take into account an important factor that may influence our observations, namely relatedness. It is easily imaginable that a specific complexity value occurs solely in the languages of one language family, but perhaps just because those languages evolved from a single ancestor with that complexity value, not necessarily because that value is intrinsically advantageous for some reason. To avoid such a bias, we need to group the observations from the same language family using a mixed-effects model in our analysis: otherwise we would be considering all data points to be unrelated, making the results of our analysis entirely unreliable.

Figures 7.3 (below) and 7.4 (next page) plot the complexity values, both for feature economy and logical complexity, for each individual language family. The shading of a circle corresponds with its relative frequency within the language family, with darker circles indicating higher proportions. The figure also lists how many languages are in a language family; these numbers are given in parentheses.

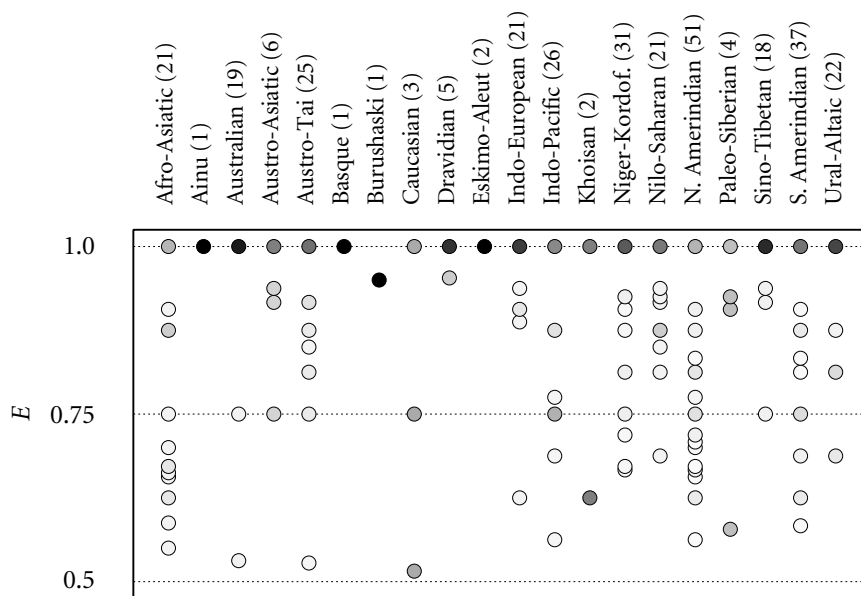


Figure 7.3. Relative frequencies of feature economy indices in UPSID. Darker circles correspond with higher relative frequencies.

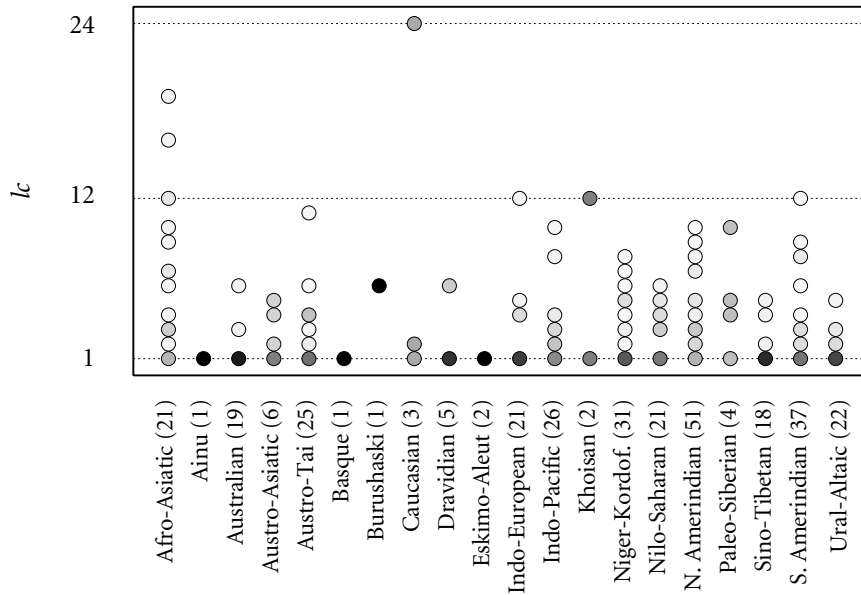


Figure 7.4. Relative frequencies of logical complexity indices in UPSID. Darker circles correspond with higher relative frequencies.

These figures do not really seem to change the picture that Figures 7.1 and 7.2 painted: across families, the darkest circles tend to be at the lowest complexity indices ($E = lc = 1$). We conclude from these figures, then, that more variation exists within language families than between them.

7.3.2 Feature economy versus logical complexity (again)

We can once again ask the question which of the two complexity measures best explains our results. We therefore carried out linear regressions with the natural logarithms of the count data (i.e. the number of languages with a certain complexity index) as the outcome variable, the complexity indices as predictors, and language family as a fixed effect. We found a significant effect of both E and lc on the natural logarithm of language count; every increase in E by 1 multiplies the language count by a factor 7.65 (95% CI 2.94...18.31, $p = 7.4 \cdot 10^{-6}$), and every increase in lc by 1 divides the language count by a factor 1.11 (95% CI 1.06...1.15, $p = 9.6 \cdot 10^{-6}$). The AIC of the E model is so much higher than that of the lc model (252.21 and 230.94, respectively), that we conclude that logical complexity is a better predictor of the distribution of plosive inventories in UPSID.

7.4 Category structures

The 317 languages in UPSID can be divided into eighty different category structures, or more precisely, there are eighty unique combinations of inventory size, E index, and lc index. In these eighty types, feature economy and logical complexity are strongly negatively correlated (Spearman's rho: -0.701). The correlation between size and feature economy is weak ($\rho = 0.035$), but larger types tend to have higher logical complexity values ($\rho = 0.242$). There are six types that occur more than ten times, many of which are already familiar from Chapter 4. Figure 7.5 repeats the eight category structures from that chapter (it is Figure 4.7 from p. 79):

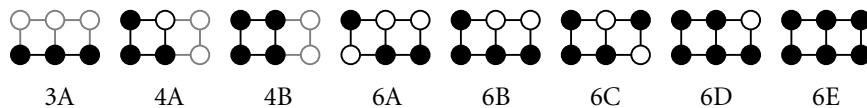


Figure 7.5. *The eight plosive category structures from Chapter 4 (p. 79).*

By far the most frequent type in UPSID, with 83 instances, has six categories, $E = 1$ and $lc = 1$. In all but three cases, this is type 6E from Chapter 4. One of these three exceptions is Nunggubuyu (Maddieson 1984: 325), which has four types of coronal segments (dental, alveolar, palato-alveolar and retroflex) in addition to labial and dorsal plosives.

A type with eight categories, $E = 1$ and $lc = 1$ was found 29 times. These are inventories with a binary laryngeal contrast and a quaternary place contrast.

There are 19 instances of inventories with four categories, $E = 1$ and $lc = 1$. This kind of inventory lacks the binary laryngeal contrast from the previously described type.

We found another 19 languages with three categories, $E = 1$ and $lc = 1$. These are languages of type 3A, with no laryngeal contrast and a ternary place contrast. In most cases, these inventories have {p t k}. Bandjalang is listed as having {b d g}; according to Sharpe (2005), there is an additional palatal stop, and these segments are not specified for voicing or manner because they are phonetically voiceless and are often realised as affricates or fricatives. Aleut has {t k q}, lacking a labial stop; it does have a labial nasal.

Again 19 instances were found of languages with five categories, $E = 0.83$ and $lc = 3$. These languages are of type 6D from Chapter 4, with a ternary place contrast, a binary laryngeal contrast, and one lacking feature combination. In all cases except for one, these languages contrast labials, coronals and dorsals; Hupa (Maddieson 1984: 370) only has {t c q^h ch^h} but does have a labial nasal. According to Golla (1970: vii), Hupa does have a sound intermediate between [b] and [p], which seems to have phonemic status but is “rare”.

The database also contains 14 languages with four categories, $E = 0.67$ and $lc = 2$. This is type 6B from Chapter 4, with a binary laryngeal contrast and a ternary place contrast, that uses one laryngeal feature value and one place of articulation exhaustively.

Table 7.2 lists how often the eight types from Chapter 4 occur in UPSID:

Table 7.2. *Frequencies of occurrence of the eight category structures from Chapter 4 in UPSID.*

	type							
	3A	4A	4B	6A	6B	6C	6D	6E
count	19	0	0	1	14	1	19	80

Of the 317 languages in UPSID, a total of 134 can be classified as one of the eight category structures. In a chaotic universe where the transition probabilities between all types are random, a Markov chain will always reach a stationary distribution in which each type makes up exactly one-eighth of the languages; in such a universe, the 134 languages in Table 7.2 would be scattered evenly across types, and we would thus expect each type to consist of $134/8 = 16.75$ languages. A Fisher's exact test reveals that the difference between these observed and random distributions is statistically significant ($p = 5.0 \cdot 10^{-4}$), meaning that the attested distribution is not due to chance.

7.4.1 Comparison with experimental data

The experimental data from Chapters 4 and 5 made typological predictions; these were explored for the category structures from Chapter 4 in §6.2 (p. 126). We can now compare these predictions with the attested data.

Table 7.3 lists the relative frequencies with which the eight types occur within this subset of 134 languages; it also lists the predicted relative frequencies in the stationary distributions predicted by the results from tasks A1 and B1, the tasks in which learners acquired novel handshape inventories, as computed in §6.2.

Table 7.3. *Relative frequencies (in %) of the category structures from Chapter 4 in UPSID, and their predicted frequencies on the basis of tasks A1 and B1.*

	type							
	3A	4A	4B	6A	6B	6C	6D	6E
UPSID	14.5	0.0	0.0	0.76	10.7	0.76	14.5	61.1
task A1	25.0	12.5	12.5	0.0	12.5	0.0	0.0	37.5
task B1	12.5	12.5	12.5	0.0	0.0	0.0	0.0	62.5

Comparing the attested numbers with the predicted ones, we see considerable differences in the two types with parameter space 4: on the basis of the experiments, these were expected to persist (probably because of their low logical complexity), but they are not found in UPSID. On the other hand, type 6D is much more prevalent in UPSID than the experimental results from Chapter 4 predicted. These discrepancies are probably due to phonetic reasons, which I discuss in the following section.

7.5 Inductive biases versus phonetics

The comparison of predicted and attested data provides a window into the interaction between learning biases and phonetic factors. In this section, I discuss some aspects of this interaction.

7.5.1 Places of articulation

UPSID does not contain any languages with three or more plosives and only two places of articulation, as indicated by the absence of types 4A and 4B (Hawaiian has a binary place contrast, but since it only has two plosives, it does not correspond to any of the category structures); clearly there is something attractive about having three or more places of articulation. Abry (2003) argues that in a plane defined by F_2 and F_3 , [b], [d] and [g] form a triangle in the same way [i], [a] and [u] do in the F_1 – F_2 plane, thus explaining their prevalence in terms of optimal auditory contrast. Lindblom et al. (2011) explain this observation as an adaptation to auditory, articulatory and acquisitional factors. They performed an experiment in which native listeners from Korean, English, Hindi and Spanish categorised 35 CV syllables in noise: these syllables consisted of the vowels {i e a o u} preceded by bilabial, dental, alveolar, retroflex, palatal, velar and uvular stops. They simulated inventories of various sizes in which three factors could be optimised: perceptual contrast between the 35 syllables, based on the confusion matrix obtained in the categorisation experiment; articulatory effort; and learning cost, based on the number of shared onsets and endpoints in the inventory (this measure thus reflects the degree of regularity; cf. §8.2.4). Simultaneous optimisation of all three factors always yields inventories with labials, alveolars and velars; if the inventory contains fewer than 16 syllables, no other places of articulation are present, and in larger inventories, additional places are used besides these three. Moulin-Frier et al. (2015) equipped computer agents with auditory and vocal-tract models and let them play “deictic games”. In such games, a dyad of agents is randomly selected from a population along with a random object. One of the agents produces a motor gesture associated with the selected object; this gesture is then converted to an acoustic signal, and the other agent subsequently updates his knowledge of the relation between objects and signals. If the jaw in the vocal-tract model is in a high position, a {b d g} inventory is the most likely three-category system to emerge under various degrees of noise.

As I noted in §7.4, not all plosive inventories in UPSID have labial, coronal and dorsal categories: Aleut, for instance, only has {t k q}, but it does use labial place in its nasals. Maddieson (2013b) mentions Dumo as a language with only labial and coronal consonants, which means that languages with only two places of articulation do seem to exist.

7.5.2 Systems with one gap

The experimental results predicted that systems with one gap, of type 6D, would not be diachronically stable, but would instead regularise. The UPSID data show otherwise: of the 134 languages that correspond to a category structure, 19 (or 14.5%) are of type 6D. In 11 of these 19 languages, it is a labial that is missing; Ohala (1996, 2009) ascribes this to the labial release burst having a relatively low intensity, because there is no down-stream resonator to amplify the burst. Maddieson (2013c) notes that [p] often lacks from Arabic languages, which he explains by the absence of [p] in Classical Arabic: there is not only a genetic relationship between these languages, making them more likely to share properties, but Classical Arabic also has high status. In 5 of the 19 type 6D languages in our sample, a dorsal is lacking, perhaps because of the higher confusability between voiceless and voiced velars (as discussed earlier in §6.7.1); in one language, Mura, the lacking category is a coronal.

7.5.3 Regularisation versus irregularity

While the learning experiments in Chapters 4 and 5 provided evidence of complexity-reducing behaviour, the typological data show a large degree of variation in complexity values: in the 317 languages in our sample, the lowest value of E was 0.355, and lc went as high as 24. If regularising behaviour is due to memory limitations, as argued by Ferdinand, Kirby and Smith (2019), we would expect larger inventories to be less complex. However, mixed-effects models with inventory size as a predictor, family as a random effect and the two complexity measures as outcome variables show that for every additional category within a plosive inventory, its feature economy index actually decreases by 0.015 (95% CI 0.0092...0.021, $p = 1.2 \cdot 10^{-6}$), and its logical complexity index increases by 0.402 (95% CI 0.304...0.500, $p = 2.7 \cdot 10^{-14}$), rejecting the hypothesis that larger inventories are less complex. Of course in larger parameter spaces, more suboptimal contrasts in terms of perception and production may exist, but remember from §5.6 that learners made significantly fewer regularising errors in task C1 (§5.2, with a 3×3 parameter space) than in tasks A1 and B1 (§§4.5–6, with a 3×2 parameter space), because of the higher logical complexity values in the larger space.

Our data are actually rather forgiving, because we investigated only plosive inventories; Moran and Blasi (2014: 234), who looked at whole inventories, plot much lower economy values than the ones reported here. However, already in our

smaller data set it is obvious that irregularity is part and parcel of plosive systems. This observation suggests that human learners, in spite of a regularity bias, are quite capable of acquiring such irregularity. Of course, natural language acquisition, as opposed to learning in the lab, may make it easier to deal with irregularity: the exposure period is defined on a vastly different scale, natural language is used in interaction, the phonological categories are connected to a lexicon instead of being isolated, and so on.

Discussion and implications

In this final chapter, I will single out and discuss a number of results presented in this dissertation, assumptions behind those results, and implications for future research. I will arrange the topics into four sections, laid out somewhat chronologically: computer simulations (§8.1), complexity measures (§8.2), levels of representation (§8.3), and limitations of the data (§8.4). I do not provide a summary of the results here; I refer the reader to pages 199 or 201 for summaries in English or Dutch, respectively.

8.1 Computer simulations

8.1.1 The lexicon and speaker normalisation

The neural network model in Chapters 2 and 3 induces phonological features through statistical learning of auditory and lexical ambient distributions. Auditory information alone suffices to create stable categories (Benders 2013; Chládková 2014; Seinhorst, Boersma & Hamann 2019; §5 in Boersma, Benders & Seinhorst 2020), but the lexicon may influence the nature of these categories too: remember how the categories were reorganised in the simulations with the bimodal category in §2.4.1 once lexical information became available. Also, the lexicon is crucial in the prototype effect, because the auditory distributions would inevitably succumb to the articulatory effect and merge, if there were nothing to tell them apart.

The lexicon also plays a role in the question of SPEAKER NORMALISATION: how is the learner able to abstract away from variation in auditory cues, both within and between speakers? Although this problem is especially relevant in category creation, mature listeners too are constantly adapting to new speakers and new circumstances. In a framework that allows simultaneous bottom-up and top-down activations (i.e. from the phonetics and the lexicon, respectively), such as our neural network model, this is actually not surprising: provided that contextual information is available to the listener, expected lexical items will become active and excite the phonological layer. In category creation, this means that acquiring lexical contrasts helps in the acquisition of phonological contrasts at SF, even if the auditory distributions are not particularly helpful: see the simulations with noise-induced mishearings in §2.6.4, effectively producing very wide peaks in the auditory input distributions. Adaptation in mature listeners is facilitated because expected lexical information activates

already existing categories at SF, so that only the cue connections between SF and AudF need to be adapted. In principle, the network could be extended to represent information about individual speakers or properties of speakers, just as humans do.

8.1.2 The lexicon and diachronic merger

In some of the computer simulations in Chapter 3, diachronic mergers occurred when categories were in close proximity on an auditory continuum. Mergers can happen in the model because it does not contain any elements or mechanisms to prevent them, such as dispersion constraints. Mergers occurred when the valleys in the pooled distributions of these categories were shallow, or altogether absent, causing learners to induce a single feature value for both categories when there is little or no activity in the lexicon yet. Because of this limited availability of lexical information, merger is self-reinforcing and irreversible. If strong activations in the lexicon were available from the onset of learning, representations at SF would always simultaneously reflect auditory similarity (i.e. categories displaying similar behaviour at AudF share nodes at SF) and lexical contrast (i.e. categories displaying contrasting behaviour in the lexicon are represented with unique nodes at SF). The outstar part of the inoutstar algorithm would then enhance the contrast over the generations to a sufficient degree, preferring those nodes at AudF that are the best prototypes of a lexical category: this is what happens in an OT model when categories are close to each other (Boersma & Hamann 2008: 248–252), and in neural networks with outstar learning (Seinhorst 2012: 31–34) as well as inoutstar learning (Seinhorst 2012: 41–42; Boersma, Benders & Seinhorst 2020: 167).

8.1.3 Sequential information and diachronic split

Another phenomenon in sound change is diachronic split, which usually happens when a single phonemic category has multiple allophones (positional variants), and the environment determining the choice between variants is eliminated. As a result, the former variants become distinctive, that is, they become two separate phonemes. This is what happened in the history of English, where the allophone /y:/ in a syllable preceding /i/ became a phoneme when the syllable containing /i/ was dropped, as I mentioned in §6.1. However, the input to the neural network model in Chapters 2 and 3 consisted of pieces of discrete data: it did not contain any sort of sequential information, neither in the phonetics (e.g. formant transitions, pitch contours, muscle movements, etc.) nor in the phonology (e.g. adjacent segments and syllables, long-distance dependencies, etc.). Because sequential information plays a central role in diachronic split, but such information is not encoded in the neural network, the model in its current form is unable to replicate this phenomenon.

8.1.4 Regularisation in the neural network

The computer simulations in Chapter 3 explored feature induction and the evolution of sound systems: the results show that phonological categories become optimally dispersed on all auditory continua. Suppose that a language has five categories, arranged in a quincunx, as in Figure 8.1 (this is type 9J from Chapter 5):

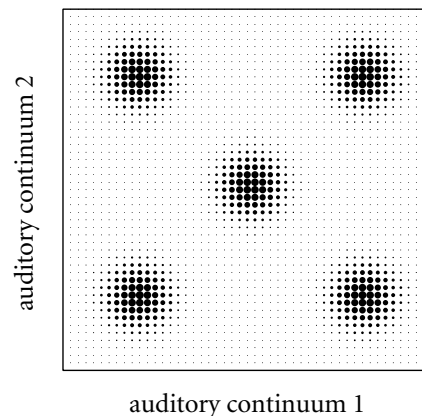


Figure 8.1. *An initial distribution in which the categories are arranged in a quincunx.*

There are ternary contrasts on both continua, and the categories are spaced evenly on both of them, with peaks at 20, 50 and 80 percent of both continua. This means that the contrasts are optimally dispersed, and in computer simulations, this inventory will not change over time. However, the results of the implicit-learning task in Chapter 5 (task C1) show that only 6 out of 12 participants replicated this type correctly; 3 out of the remaining 6 added one or two categories so that one or two feature values were used exhaustively. Also, in UPSID, of the 100 languages with a 3×3 parameter space, none were of type 9J. However, the neural network is incapable of regularisation by design, because the number of lexical categories in the language is fixed in the simulation script: no new categories can be introduced, as might happen in natural language through split or borrowing. The computer simulations and the experiments, then, complement each other.

8.2 Complexity measures

8.2.1 Computing complexity indices

Two complexity measures recurred throughout this dissertation: feature economy and logical complexity. I motivated the choice for these two measures in §4.3.2 on p. 82: feature economy is a familiar concept in phonology, and logical complexity seems to be a useful concept in explanations of the learning of feature combinations.

A general problem with economy measures, at least in terms of comparability between sources, is that most if not all of them are defined using binary features. The quantifications proposed by Clements (2003) and Mackie and Mielke (2011) only work for analyses in which all features have the same valency; of Hall's (2007) definitions, only "Exploitation" can be adjusted for features with different valencies, as I did in Chapter 4. A further distinct disadvantage of feature economy is that it does not distinguish between inventories that have identical numbers of categories and parameter space sizes, but in which the categories are laid out differently, such as types 6B and 6C in Chapter 4 (cf. Figure 4.7 on p. 82).

The logical complexity measure is fundamentally identical to Kolmogorov complexity, which is the length of the shortest computer programme that generates the desired data set. While this measure has mostly been applied outside of linguistics, it is not entirely undisputed, because there is no algorithm to calculate it (Shalizi 2006: 53); indeed, all logical complexity values reported in this dissertation were established by hand. Nevertheless, logical complexity seems to be a valuable predictor in the analyses of many of my results (see §8.2.3).

In Chapter 4, I decided not to use negation in the minimal formulas that determine the logical complexity indices, contra Feldman (2000); the formulas in Chapters 4 and 5 list whatever *is* there, they do not negate what is *not* there. For instance, as I wrote in §4.3.2 (p. 82), this entails that type 6D (the 3×2 type with one gap) cannot be defined as "anything except the missing category" but should instead be described in terms of its features and/or feature combinations that are present. This has consequences for the logical complexity values: the former description, $(a_2b_3)'$, would yield $lc = 2$, while the latter, $a_1 + b_1 + b_2$, yields $lc = 3$ (cf. Table 4.6, p. 81). Table 8.1 lists the minimal formulas and lc values of the category structures from Chapter 4 using negation, as Feldman did. Feature a is the (at most) binary feature, usually a laryngeal contrast; feature b is the (at most) ternary place contrast.

Table 8.1. *A modified version of Table 4.6, allowing for negation.*

Type	disjunctive normal form	minimal formula	lc
3A	$a_1b_1 + a_1b_2 + a_1b_3$	A [all]	1
4A ¹⁹	$a_1b_1 + a_1b_2 + a_2b_1$	$a_1 + b_1$	2
4B	$a_1b_1 + a_1b_2 + a_2b_1 + a_2b_2$	A [all]	1
6A	$a_2b_1 + a_1b_2 + a_1b_3$	$a_2b_1 + a_1(b_1)'$	4
6B	$a_1b_1 + a_1b_2 + a_1b_3 + a_2b_1$	$a_1 + b_1$	2
6C	$a_1b_1 + a_1b_2 + a_2b_1 + a_2b_3$	$a_1(b_3)' + a_2(b_2)'$	4
6D	$a_1b_1 + a_1b_2 + a_1b_3 + a_2b_1 + a_2b_2$	$(a_2b_3)'$	2
6E	$a_1b_1 + a_1b_2 + a_1b_3 + a_2b_1 + a_2b_2 + a_2b_3$	A [all]	1

¹⁹ The alternative using negation, $(a_2b_2)'$, has $lc = 2$ as well.

Compared to Table 4.6, these values are different for types 6A, 6C, and 6D. Table 8.2 lists the AICs of the models in the implicit-learning tasks from Chapter 4, without and with the use of negation in the minimal formulas:

Table 8.2. *Comparisons of the models of the implicit-learning tasks from Chapter 4, without and with the use of negation.*

stimuli	task	outcome	AIC _{without}	AIC _{with}	Δ AIC
handshapes	A1	error score	408.50	410.34	1.85
		match	25.71	27.30	1.59
	B1	error score	757.70	758.80	1.10
		match	55.90	55.64	0.26
speech	B2	error score	837.76	837.75	$4.4 \cdot 10^{-3}$
		match	113.47	113.52	0.056

Remember from §4.8.1 (p. 100) that the model with the lowest AIC should be considered the model that best fits the data; Δ AIC indicates the degree of information loss between them. Following Burnham and Anderson's (2004: 271) criteria for model comparison, also mentioned in §4.8.1 (p. 100), I assume that if Δ AIC is less than or equal to 2, the model with the higher AIC has "substantial support"; if Δ AIC is between 4 and 7, this model has "considerably less support"; and if Δ AIC is greater than 10, the model has "essentially no support". In most comparisons in Table 8.2, the model without negation has a lower AIC, but the Δ AIC between both models is never large enough that one can be preferred over the other. In §8.2.3, where I compare feature economy and logical complexity as predictors, I further explore the consequences of the effect of negation in the 3×2 types.

In the 3×3 types, the larger parameter space results in more gaps, and therefore the effects of the decision not to use negation may be larger. For instance, using negation, type 9H would have the minimal formula $a_3(b_1)' + b_1(a_3)'$ and hence a logical complexity of 4, rather than 6 as in Table 5.2. This reduction from 6 to 4 holds for all four types 9H–9L, which were used in tasks C1 and C2 because of their identical lc values without negation. There were three more types in tasks C1 and C2: type 9B, whose lc remains 2 with the use of negation, type 9F, whose lc remains 5, and type 9N, whose lc is reduced from 8 to 6. Models of the results from task C1 with these adjusted values are no improvements over the models without negation: the AIC of the model with correct replication as the outcome (§5.2.1) increases by 1.00, and the AIC of the model with reaction time quantile as the outcome (§5.2.2) increases by 0.48.

For any inventory, both complexity measures abstract away from the required features. All regular systems have $E = lc = 1$, regardless of the size of the system and

the number of constituent features: in terms of these two complexity measures, the inventories $\{p\ t\ k\}$ and $\{p\ p^i\ p^w\ t\ t^i\ t^w\ k\ k^i\ k^w\ b\ b^i\ b^w\ d\ d^i\ d^w\ g\ g^i\ g^w\}$ are entirely indistinguishable. In that respect, a minimal formula “A [all]” is quite uninformative; similarly, a minimal formula [VOICELESS] tells us nothing about the possible places of articulation. For instance, a $\{p\ t\ k\}$ language can be described as [VOICELESS], but this description does not include $\{q\}$, even though $\{q\}$ is voiceless too. In fact, the minimal formula [VOICELESS] for a $\{p\ t\ k\}$ inventory is in a sense perpendicular to the learner’s representation, who does not have to represent a laryngeal feature at all, but should instead represent a place contrast (see Jost 2004: 74–76 for a discussion of this problem). Knowledge of the features that define the parameter space, then, is complementary to logical complexity, and to feature economy as well. The reverse is true too, because many different inventories can be defined with the same features. If we want to draw comparisons between languages, as we did in Chapter 7, these generalisations across different parameter spaces are crucial, because the notion of regularity can apply to inventories of any size.

8.2.2 Correlated measures of complexity

The feature economy and logical complexity measures made different predictions with regard to some of the category structures from Chapters 4 and 5, in spite of the strong correlation between them. Unfortunately, the issue of correlated measures seems impossible to avoid. Consider, for instance, another possible definition of complexity: the average number of contrasts that each category participates in (I briefly discussed this measure in §4.3.2). This concept is related to the notion of functional load, that is, the number of distinctions in the lexicon in which a category participates; categories with a lower functional load are more likely to be lost diachronically (Trubetzkoy 1939; Martinet 1955; Wedel, Kaplan & Johnson 2013). The average number of minimal pairs, however, is defined within an isolated category structure, without a lexicon, while functional load necessarily involves lexical frequency. Table 8.3 lists, for the 22 category structures introduced in this book, the values of five complexity measures: the number of categories, the feature economy index, the logical complexity index, the number of gaps, and the average number of contrasts in which a category participates.

In these 22 types, of the $\binom{5}{2} = 10$ correlations between pairs of complexity measures, eight correlation coefficients have absolute values equal to or greater than 0.4; of these eight, four have absolute values equal to or greater than 0.7. Disregarding the last measure, in the UPSID data, five out of the $\binom{4}{2} = 6$ correlations have coefficients with absolute values equal to or greater than 0.4, and three of these five have absolute values equal to or greater than 0.7. Other measures, such as parameter space and Shannon entropy, can be computed directly from the ones in the table and are therefore likely to be correlated with other measures too.

Table 8.3. Complexity scores of all 22 category structures from Chapters 4 and 5.

type	3A	4A	4B	6A	6B	6C	6D	6E
categories	3	3	4	3	4	4	5	6
feature economy	1.0	0.75	1.0	0.5	0.67	0.67	0.83	1.0
logical complexity	1	2	1	5	2	5	3	1
gaps	0	1	0	3	2	2	1	0
avg. contrasts	2	1.33	2	0.5	2	1.5	2.4	3
type	9A	9B	9C	9D	9E	9F	9G	
categories	9	5	7	6	8	5	3	
feature economy	1.0	0.56	0.78	0.67	0.89	0.56	0.33	
logical complexity	1	2	3	4	4	5	6	
gaps	0	4	2	3	1	4	6	
avg. contrasts	4	2.4	3.14	2.67	3.5	2	0	
type	9H	9J	9K	9L	9M	9N	9P	
categories	4	5	6	7	4	5	6	
feature economy	0.44	0.56	0.67	0.78	0.44	0.56	0.67	
logical complexity	6	6	6	6	7	8	9	
gaps	5	4	3	2	5	4	3	
avg. contrasts	1	1.6	2.33	2.86	1	1.6	2	

Bentz et al. (2016) investigated correlations between measures of morphological complexity in the WALS, finding only highly significant correlations. They argue that the choice for a measure can therefore depend on practical considerations, but even highly correlated measures can still make different predictions, as we saw for types 6B and 6C in Table 4.7 (p. 82), and the model comparisons in this dissertation have shown that one measure may be preferred over another, so correlated measures are not necessarily interchangeable. In cases where individual predictors need to be compared, as in Chapters 4, 5 and 7, the use of Akaike's Information Criterion (AIC) has proven a useful strategy. The following subsection returns to those comparisons.

8.2.3 Feature economy versus logical complexity (one last time)

A topic that recurred throughout this dissertation was a comparison between feature economy and logical complexity as predictors in statistical models. This subsection returns to this topic one last time, by providing a brief overview of the comparisons in all analyses and, if possible, choosing a winner. Table 8.4 (next page) lists short descriptions of all analyses (with task numbers for the implicit learning tasks), the section where the AIC values of the models were given, the outcome variables of the models, and the complexity measure with the lowest AIC followed by the AIC difference between the two models ("Δ", in parentheses). The criteria for model comparison were mentioned in §4.8.1 (p. 100), and in this chapter in §8.2.1 (p. 161). I take into account ΔAICs of 4 or larger; in the table, any ΔAIC < 4 has been greyed out.

Table 8.4. *Comparisons of the two complexity measures as predictors.*

description	task	§	outcome variable	AIC _{min} (Δ)
impl. learning: handshapes	A1	4.8	error score	<i>lc</i> (18.7)
			match	<i>lc</i> (15.6)
	B1	4.8	error score	<i>lc</i> (14.3)
			match	<i>lc</i> (5.5)
	C1	5.2	match	<i>lc</i> (2.1)
			reaction time	<i>lc</i> (4.5)
impl. learning: speech	B2	4.8	error score	<i>E</i> (0.1)
			match	<i>E</i> (0.2)
all impl. learning tasks		5.6	match	<i>lc</i> (32.6)
sound systems		7.3	inventory count	<i>lc</i> (21.3)

If there is one measure that best predicts the data in any of the models, whether it analyses the implicit-learning data or the data from UPSID, it is logical complexity. In those cases where logical complexity was not the (considerably) better predictor, that is, the greyed out cells in the table above, a model with both *E* and *lc* as predictors was not an improvement over the model with the lowest AIC: in all three cases, the AIC of this model was actually the highest of the three.

In the linguistic literature, Kolmogorov complexity is usually applied to corpora of text: Juola (1998) and Ehret (2018) use it to assess morphological complexity in several languages, Benedetto, Cagliotti and Loreto (2002) use it in the phylogenetic classification of translations of the same text, and Ehret and Szmrecsanyi (2019) use it to assess the complexity of the written output of second language learners and correlate this complexity with properties of their input. More typological research will have to show the value of this measure as a predictor of grammatical complexity, both in phonology and in other fields.

If we allow for the use of negation in minimal formulas, the model comparisons from tasks A1, B1 and B2 come to look as follows:

Table 8.5. *Comparisons of the two complexity measures as predictors, allowing for negation in the *lc* values.*

description	task	§	outcome variable	AIC _{min} (Δ)
impl. learning: handshapes	A1	4.8	error score	<i>lc</i> (23.2)
			match	<i>lc</i> (14.0)
	B1	4.8	error score	<i>lc</i> (13.2)
			match	<i>lc</i> (5.8)
impl. learning: speech	B2	4.8	error score	<i>E</i> (0.13)
			match	<i>E</i> (0.31)

Although the exact values in Table 8.5 differ from those in Table 8.4, the conclusions based on Table 8.4 remain valid, also with the use of negation.

8.2.4 Gestural economy and perceptual warping

The implicit-learning experiments in Chapters 4 and 5 were intended to focus on learning biases, minimising the roles of perceptual distinctiveness and articulatory ease (although the former can never be ignored altogether). While an inductive bias towards lower complexity was found in all tasks, this bias is not necessarily the only explanation of regularity in sound systems: Maddieson (1995), for instance, argues instead that economy does not exist at an abstract, featural level, but rather at a gestural level, as a tendency to reuse articulatory gestures: he uses articulographic data from a speaker of Ewe to show that the labial gesture in [p] and the dorsal gesture in [k] are identical to the gestures in labial-dorsal [kp]. Articulatory similarity between members of an inventory was one of the optimisation criteria in the simulations by Lindblom et al. (2011), described in §7.5.1 (p. 153). Bybee (2001: 54) also argues that speakers tend to “reuse a single set of highly entrenched neuromotor patterns and substitute members of this set for novel or less common configurations”, thus extending the implications of this hypothesis to sound change too.

A comparable entrenchment effect may also be present in perception. Kuhl (1991) described a “perceptual magnet effect”: she found that the auditory space occupied by a vowel category has internal structure, that is, listeners judge certain realisations as better or more prototypical than others; she also found that listeners are worse at discriminating between tokens that are close to this prototype than between tokens that are close to a non-prototypical realisation. A significant effect is already present in six-to-seven months old infants; no significant effect is found in rhesus monkeys. Kuhl attributes the effect to exposure to distributions of linguistic auditory cues, but her interpretation that the effect does not exist in rhesus monkeys is not a valid conclusion from a non-significant *p* value; also, she does not attempt any direct comparisons between groups, so her claim of the species-specificity of the effect is unsupported. Nevertheless, it seems clear that attractors exist in human perception, production and learning; they are very likely to have similar consequences in the structure of sound systems.

8.3 Levels of representation

8.3.1 Phonemes and allophones

In §4.3.2 (p. 80), I decided to make a principled distinction between type 3A languages, such as {p t k}, and type 6A languages, such as {p t g}. In both types, the place feature alone suffices to distinguish between the members of the inventory, so

we could say that these types are identical in terms of their contrastive features: we might therefore simply assume a place feature at the phonemic level. However, the description of an inventory such as {p t g} seems to require an additional voicing feature (at least phonetically), even though no two categories exist that differ with respect to only this feature.

If the distinction between types 3A and 6A is somehow meaningful, and the different complexity indices are indeed justified, we might expect to find an effect of type on error score in tasks A1 and B1, and this is indeed true: the average error score of all 18 learners of type 6A in tasks A1 and B1 together is significantly higher than that of all 18 learners of type 3A together (by 12.91 points, 95% CI 4.35...21.47, $p = 0.0043$). The experimental design in tasks A and B does not distinguish between phonemes and allophones: according to the view taken in §1.1.1, allophones are positional variants of phonemes, and since the experimental task did not involve any context or alternations this distinction is impossible to make. The significant difference in error scores between types 3A and 6A might suggest that most learners made an abstraction from the phonetic input to an allophonic level, not a phonemic level; on the other hand, there was one learner in task A1 who produced a type 3A output for their type 6A input, suggesting they ignored the non-distinctive feature.

The level of representation at which individual learners induce features also bears on the comparison between the experimental and typological data, because languages may have different phonemic and allophonic inventories. The plosive system of Dutch, for instance, lacks |g| (a small number of loanwords aside), and its members {p t k b d} form a type 6D inventory; however, /g/ is an allophone of |k| before voiced obstruents, as in the compound of the morphemes |zak| ‘pocket’ and |duk| ‘cloth’, which surfaces as /zagduk/ ‘handkerchief’. The plosive allophones of Dutch, then, constitute a type 6E inventory, rather than a 6D inventory. Positional variants are ignored altogether in the category structures and their complexity counts, as are restrictions on the distributions of segments across syllable positions; indeed, Ohala (2009: 49) warns that, exactly for this reason, the “apparent symmetry found in many languages’ segment inventories [...] obscures a more complicated situation”.

8.3.2 Learnability in other domains versus concrete reality

We may expect learning biases to be reflected not only in phoneme inventories, but in other linguistic domains too, both in- and outside of phonology. Another domain that is analysed using features are personal pronoun systems, with such properties as number, person, gender, in-/exclusive, proximity, and so on. The difference between personal pronouns and phonemes, however, is that pronominal features refer to semantic properties: while the relation between phonemes and their referents is language-specific and unpredictable, perhaps with the exception of some instances of onomatopoeia or other kinds of sound symbolism (a.o. Dingemanse et al. 2016),

combinations of pronominal features tend to refer to (groups of) persons that can exist in the external world, at least the number and person features. Because the combinations of these two features have reflections in reality, we might expect personal pronoun systems to display more regularity than phoneme inventories.

To test this hypothesis, I looked at the Free Personal Pronoun System, a database of personal pronoun paradigms (FPPS; Smith & Van Rijn 2012), more specifically those languages that have a ternary person contrast (usually first vs. second vs. third) and a binary number contrast (usually singular vs. plural). Because of these properties, such pronoun systems can be described in terms of the category structures from Chapter 4. If a combination of person and number feature values is encoded with a unique, monomorphemic pronoun, we can regard this as a category that exists in the category structure, and if such a combination is lacking in the language (i.e. the meaning has to be expressed periphrastically), it can be thought of as non-existent in terms of the category structure. (Languages that neutralise the number contrast, by using the same monomorphemic form for plural and singular referents, are not taken into account here.) Of the 429 pronoun systems that correspond to a category structure, the vast majority (409 systems, or 95.34%) are of Type 6E, the regular type with all six categories; one system is of type 3A (it has no plural pronoun forms); seven systems are of type 4B (altogether lacking one value of the person feature). The remaining twelve languages are of type 6D, in which one feature combination cannot be expressed with a pronoun. In Table 7.2, listing the sound systems in UPSID that correspond to a category structure from Chapter 4, there were 99 regular inventories out of 134 (73.9%); in the FPPS, there are 417 out of 429 (97.2%). A Fisher's exact test reveals that the proportion of regular to irregular systems is significantly larger in pronominal systems (odds ratio: 12.21, 95% CI 5.93...26.81, $p = 1.55 \cdot 10^{-14}$), which may be due to their semantic motivation. However, this test does not incorporate information about language family, which is not available in the FPPS. More typological research would be needed to further investigate this hypothesis, and to explore learnability effects in domains other than phoneme inventories.

8.4 Limitations of the data

8.4.1 The input to the neural network

The neural network received synthetic data as input, in which the auditory distributions were well dispersed and were defined on only two continua, and in which it was always clear to the learner which lexical category was intended (except in §2.6.1). Naturalistic data with more auditory continua and more lexical ambiguity may make feature induction more onerous; although the stress tests in §2.6 confirm that the process is resistant to perturbations of the input, the noise in naturalistic input data

is less likely to decrease during the learning process than it did in the input data of §2.6. Interestingly, computational and experimental evidence suggests that the learning process is actually aided by variability in the presence of cues. Monaghan (2017) simulated the cross-situational learning of word–object mappings in which three additional cues could be present: parts of speech that contained useful statistical information (e.g. a definite article reliably signalling an upcoming noun), prosodic information (i.e. raised F_0), and deictic information (i.e. pointing at the intended object). Monaghan varied the presence of each individual cue between 33 and 100%; the simulated learner failed to recognise the word–objects mappings correctly when the additional cues were not present, if a cue had consistently been present during learning, because the learner had not acquired the mapping itself but instead relied on the presence on the cue alone. Monaghan et al. (2017) replicated these results in a learning experiment with adult human participants, finding that they performed best if each of the individual cues was present in 75% of the cases. In future research, the neural network could be trained on naturalistic data, too, and its effects on word learning could be investigated.

8.4.2 Modality

The sets of handshapes in the implicit-learning tasks were designed to have easily learnable properties: the relevant distinctions did not involve orientation or movement, which are known to be difficult to late learners (cf. §4.4.1). The thumb opposition contrast, however, does not seem to be found in natural sign languages. An experiment with handshapes from a natural sign language would be an interesting follow-up study; ideally, the results of such a study should be compared to typological data about sign languages, although unfortunately, a sufficiently large database of sign language phoneme inventories currently does not exist (Roland Pfau, p.c.).

8.4.3 Adult versus child learners

All participants in the learning experiments in Chapters 4 and 5 were adults without any sign language proficiency: they were late learners in a new modality (“L2M2”). This is not unproblematic, because ideally I would like to say something about L1M1 acquisition too, that is, children who acquire a first language. Earlier evidence suggests that adult learners tend to match the statistics of the input, where children seem to overgeneralise and regularise more (Hudson Kam & Newport 2005, 2009; Culbertson & Newport 2015; Kempe, Gauvrit & Forsyth 2015), perhaps because of processing limitations. However, the tasks in none of these sources resembled the one described here: Kempe, Gauvrit and Forsyth used patterns of dots in a grid as input, and the other sources focused on morphosyntactic properties, such as word

order. It would be interesting to repeat the tasks from this dissertation with children (or even infants, with an adjusted set of tasks), and compare the outcomes.

8.4.4 Regularisation and cognitive load

In experiment B, learners in the handshape learning task (task B1) made more regularising errors than in the speech learning task (task B2), perhaps because of the added cognitive load of inducing novel features in an unfamiliar modality, which they did not have to do in the speech task (cf. also §4.9). These results are in line with Ferdinand, Kirby and Smith's (2019) hypothesis that increased cognitive load is correlated with a higher degree of regularisation. However, in task C1, the implicit-learning task with the 3×3 parameter space, learners made significantly fewer regularising mismatches than the learners in tasks A1 and B1, with the 3×2 parameter space (§5.6). This difference may be attributed to at least two reasons. Firstly, the methodology of the tasks is different, and therefore the tasks are not entirely comparable. In the learning phase of the 3×2 tasks, participants were exposed to more tokens of each category in their input than the learners in the 3×3 task, because the response variable in the test phase of the 3×3 task was binary rather than continuous (as in the 3×2 task): in the 3×2 tasks, learners needed enough input to estimate relative frequencies, while in the 3×3 tasks, they only needed to indicate whether they had seen a category or not. Secondly, as mentioned in §5.6, the bigger parameter space of the types in task C1 resulted in larger types and higher complexity. Perhaps learners lost track of the input pattern and did not oversee the full parameter space because of this increased cognitive load, contrary to Ferdinand, Kirby and Smith's (2019) hypothesis. Interestingly, larger plosive inventories exhibit a similar pattern: remember from §7.5.3 that larger systems tend to be more complex. Also, in the forced-choice task from §5.3, I found a negative effect of complexity on learners' ability to select the complexity-reducing category. Taken together, these findings suggest that added cognitive load may only lead to regularisation under certain circumstances, more specifically when novel features are induced in an inventory of moderate size and/or complexity. More research, both experimental and typological, could shed more light on the relation between cognitive load, regularising behaviour, and regularity in language.

8.4.5 Effect sizes in the Markov chains

The implicit-learning experiments were originally planned to have an iterated-learning component, which had to be abandoned for practical reasons. Instead, I set up Markov chains on the basis of the results from the implicit-learning tasks, assuming the simplest possible model of language transmission, in which every generation learns from an entire previous generation and behaves exactly identically to it. For these chains, I used the match data from the learning experiments in Chap-

ter 4, in which the continuous estimated frequencies were translated to categorical response variables (did the learner indicate having seen the six handshapes or not, and which category structure resulted from that response?). In §4.4.4 (p. 89), I specified strict criteria for answering the first question, and the stationary distributions of the Markov chains depend on them entirely. Using the criteria from §4.4.4, for some types there is only a single learner who did not replicate their input correctly, but this still led to complete regularisation in the long run, as for type 6D in task B1 (compare Table 6.1, Matrix (6.1), and Matrix (6.3) in §6.2.1). Interestingly, if transitions between types that are based on a single learner were removed from the Markov chain, types 6A and 6D would survive in both tasks A1 and B1, lowering the lowest stable feature economy value to 0.5 (the lowest value of all eight types), and raising the highest surviving logical complexity value to 5 (the highest value of all eight types). On the other hand, if the minimum misestimation to consider a category as seen were zero instead of 25, there would be three learners of type 6D in task B1 who regularised their input, instead of only one learner; there would also be one learner of type 4A, instead of zero, who regularised their input to type 4B. Future research could include a more realistic approach to language transmission, in which the information that is transmitted between multiple generations of human learners contains the estimated relative frequencies rather than coarse binary decisions.

8.4.6 Analyses of typological data

The analyses of the typological data in Chapters 6 and 7 of course rely entirely on the original descriptions. For many languages, UPSID lists only one reference. In this subsection, I will focus on two category structures that rarely occurred in UPSID, namely types 6A and 6C. It will turn out that the use of different sources leads to different counts in Table 7.4.

As the only instance of a type 6A language, UPSID has Seneca with {t k b}, citing Chafe (1967). This source gives {t k} as full-fledged phonemes (p. 5) and mentions {b} as “occur[ring] only in a few nicknames” (p. 6). However, Chafe (2015: 10) lists {t k d g}, but he also states that {p b} occur in ideophones and nicknames; pursuing Maddieson’s inclusion criterion, this would mean that Seneca is a type 6E language after all, rather than a 6A language.

The results from tasks A1 and B1 in Chapter 4 suggest that Type 6C is the least learnable of the eight category structures. In UPSID, there is one language of this type, namely Efik, which has {t k b d} (Cook 1969, 1985). Efik also has a labiovelar segment [kp] (Ward 1933; Cook 1969, 1985; Mensah & Mensah 2014), which we did not take into account because of its double place of articulation; this segment is sometimes analysed as labial only (Welmers 1973: 75), although place assimilation of a preceding nasal usually yields /ŋ/ (Cook 1969, 1985). According to both Ward (1933) and Mensah and Mensah (2014), Efik actually has both [p] and [kp], which in

our analysis would classify it as a type 6D phoneme inventory, quite straightforward in lacking [g] (cf. §6.7.1); Ward (1933) also lists /g/ as a possible realisation of [h], suggesting a type 6E allophone inventory.

From the description of a phoneme inventory we cannot infer the existence of any unnatural classes, even though such classes occur quite frequently in Mielke's (2008) survey (cf. §1.1.2, p. 3). If a language has {p t k b d g}, and {p t k b} form a class that triggers or undergoes some phonological process, then these categories will still be listed in a 3×2 (or 2×3) matrix, even though this is not necessarily how the language learner will represent them. It is difficult to predict the implications of the existence of any unnatural classes for the complexity indices.

For practical reasons, the results presented in Chapter 7 come from a fairly small dataset: we looked only at plosive inventories, and only at 317 languages. We did this because all languages in the sample had plosives, and because the effects of articulatory and perceptual factors within this sample are probably comparable between languages. Future research could include a larger sample of languages, and larger subsets of inventories, or entire inventories; nevertheless, even in our small sample, we found robust effects of complexity on frequency, and of inventory size on complexity. We expect these effects to generalise to a larger sample of languages too.

References

- Abry, Christian (2003). [b]–[d]–[g] as a universal triangle as acoustically optimal as [i]–[a]–[u]. *Proceedings of the 15th International Congress of Phonetic Sciences*: 727–730.
- Akaike, Hirotugu (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.
- Al Kendi, Azza & Ghada Khattab (2019). Acoustic properties of foreigner-directed speech. In: Calhoun, Sasha, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences*, paper 650. Canberra: Australasian Speech Science and Technology Association.
- Alem, Sylvain, Clint Perry, Xingfu Zhu, Olli Loukola, Thomas Ingraham, Eirik Søvik & Lars Chittka (2016). Associative mechanisms allow for social learning and cultural transmission of string pulling in an insect. *PLoS Biology* 14 (10): e1002564.
- Atkinson, Quentin (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332: 346–349.
- Augst, Gerhard (1975). *Untersuchungen zum Morpheminventar der deutschen Gegenwartssprache*. Tübingen: Gunter Narr Verlag.
- Baer-Henney, Dinah (2015). *Learner's little helper: strength and weakness of the substantive bias in phonological acquisition*. Doctoral dissertation, University of Potsdam.
- Balota, David & Daniel Spieler (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: beyond measures of central tendency. *Journal of Experimental Psychology: General* 128: 32–55.
- Bartlett, Frederic (1932). *Remembering: a study in experimental and social psychology*. Cambridge: Cambridge University Press.
- Benders, Titia (2013). *Nature's distributional-learning experiment. Infants' input, infants' perception, and computational modeling*. Doctoral dissertation, University of Amsterdam.
- Benedetto, Dario, Emanuele Caglioti & Vittorio Loreto (2002). Language trees and zipping. *Physical Review Letters* 88: 048702.
- Benni, Tytus (1929). Zur neueren Entwicklung des Phonembegriffs. *Donum Natalicium Schrijnen*, 34–37. Nijmegen/Utrecht: N. V. Dekker & Van de Vegt.

- Bentz, Christian, Tatyana Ruzsics, Alexander Kopenig & Tanja Samardžić (2016). A comparison between morphological complexity measures: typological data vs. language corpora. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (Osaka, Japan)*: 142–153.
- Bentz, Christian & Bodo Winter (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3: 1–27.
- Berg, Thomas (1998). *Linguistic structure and change: an explanation from language processing*. Oxford: Clarendon Press.
- Blackwell, H. Richard & Harold Schlosberg (1943). Octave generalization, pitch discrimination, and loudness thresholds in the white rat. *Journal of Experimental Psychology* 33: 407–419.
- Blaho, Sylvia (2008). *The syntax of phonology: a radically substance-free approach*. Doctoral dissertation, University of Tromsø.
- Blasi, Damian, Steven Moran, Scott Moisik, Paul Widmer, Dan Dediu & Balthasar Bickel (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363: eaav3218.
- Bloomfield, Leonard (1933). *Language*. New York: Holt.
- Blutner, Reinhard (2000). Some aspects of optimality in natural language interpretation. *Journal of Semantics* 17: 189–216.
- Bochner, Joseph, Karen Christie, Peter Hauser & Matt Searls. (2011). Learners discrimination of linguistic contrasts in American Sign Language. *Language Learning* 61: 1302–1327.
- Boersma, Paul (1998). *Functional phonology: formalizing the interactions between articulatory and perceptual drives*. Doctoral dissertation, University of Amsterdam.
- Boersma, Paul (2003). The odds of eternal optimization in Optimality Theory. In: Holt, D. Eric (ed.), *Optimality Theory and language change*, 31–65. Dordrecht: Kluwer.
- Boersma, Paul (2006). Prototypicality judgments as inverted perception. In: Fanselow, Gisbert, Caroline Féry, Ralf Vogel & Matthias Schlesewsky (eds.), *Gradience in grammar: generative perspectives*, 167–184. Oxford: Oxford University Press.
- Boersma, Paul (2008). Emergent ranking of faithfulness explains markedness and licensing by cue. *Rutgers Optimality Archive* 954.
- Boersma, Paul (2009). Cue constraints and their interactions in phonological perception and production. In: Boersma, Paul & Silke Hamann (eds.): *Phonology in perception*, 55–110. Berlin: Mouton de Gruyter.

- Boersma, Paul (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In: Benz, Anton & Jason Mattausch (eds.), *Bidirectional Optimality Theory*, 33–72. Amsterdam: John Benjamins.
- Boersma, Paul (2019). Simulated distributional learning in deep Boltzmann machines leads to the emergence of discrete categories. In: Calhoun, Sasha, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences*, paper 808. Canberra: Australasian Speech Science and Technology Association.
- Boersma, Paul, Titia Benders & Klaas Seinhorst (2020). Neural network models for phonology and phonetics. *Journal of Language Modelling* 8: 103–177.
- Boersma, Paul, Paola Escudero and Rachel Hayes (2003). Learning abstract phonological from auditory phonetic categories: an integrated model for the acquisition of language-specific sound categories. *Proceedings of the 15th International Congress of Phonetic Sciences*: 1013–1016.
- Boersma, Paul & Silke Hamann (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25: 217–270.
- Boersma, Paul & David Weenink (2018). *Praat. Doing phonetics by computer*. Software, version 6.0.43, downloaded from www.praat.org on November 1st, 2018.
- Bogacz, Rafał, Marius Usher, Jiayang Zhang & James McClelland (2007). Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences* 362: 1655–1670.
- Breen, Gavan & Rob Pensalfini (1999). Arrernte: a language with no syllable onsets. *Linguistic Inquiry* 30: 1–26.
- Brentari, Diane (1990) *Theoretical foundations of American Sign Language phonology*. Doctoral dissertation, University of Chicago.
- Burnham, Kenneth & David Anderson (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* 33: 261–304.
- Butcher, Andrew (2006). Australian Aboriginal languages: consonant-salient phonologies and the ‘place-of-articulation imperative’. In: Harrington, Jonathan & Marija Tabain (eds.), *Speech production: models, phonetic processes, and techniques*, 187–210. New York: Psychology Press.
- Bybee, Joan (2001). *Phonology and language use*. Cambridge: Cambridge University Press.
- Bybee, Joan, Revere Perkins & William Pagliuca (1993). *The evolution of grammar: tense, aspect and modality in the languages of the world*. Chicago: University of Chicago Press.

- Chafe, Wallace (1969). *Seneca morphology and dictionary*. Washington, DC: Smithsonian Press.
- Chafe, Wallace (2015). *A grammar of the Seneca language*. Oakland, CA: University of California Press.
- Chater, Nick & Morten Christiansen (2010). Language acquisition meets language evolution. *Cognitive Science* 34: 1131–1157.
- Chater, Nick & Morten Christiansen (2018). Language acquisition as skill learning. *Current Opinion in Behavioral Sciences* 21: 205–208.
- Chládková, Kateřina (2014). *Finding phonological features in perception*. Doctoral dissertation, University of Amsterdam.
- Cho, Taehong, Sun-Ah Jun & Peter Ladefoged (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics* 30: 193–228.
- Chomsky, Noam (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- Chomsky, Noam (1981). *Lectures on government and binding*. Studies in Generative Grammar 9. Dordrecht: Foris.
- Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.
- Chomsky, Noam & Howard Lasnik (1993). Principles and Parameters theory. In: Jacobs, Joachim, Arnim von Stechow, Wolfgang Sternefeld & Theo Vennemann (eds.), *Syntax: ein internationales Handbuch zeitgenössischer Forschung*, 506–569. Berlin: De Gruyter.
- Christiansen, Morten & Nick Chater (2008). Language as shaped by the brain. *Behavioral and Brain Sciences* 31: 489–558.
- Clements, G. Nick (2003). Feature economy in sound systems. *Phonology* 20: 287–333.
- Clements, G. Nick (2009). The role of features in speech sound inventories. In: Raimy, Eric and Charles Cairns (eds.), *Contemporary views on architecture and representations in phonological theory*, 19–68. Cambridge, MA: MIT Press.
- Clements, G. Nick & Elizabeth Hume (1995). The internal organization of speech sounds. In: Goldsmith, John (ed.), *The handbook of phonological theory*, 245–306. Oxford: Blackwell.
- Clements, G. Nick & Engin Sezer (1982). Vowel and consonant disharmony in Turkish. In: van der Hulst, Harry & Norval Smith (eds.), *The structure of phonological representations, vol. 2*, 213–255. Dordrecht: Foris.
- Comrie, Bernard (1992). Before complexity. In: Gell-Mann, Murray & John Hawkins, *The evolution of human languages: Proceedings of the Workshop on the Evolution of Human Languages*, 193–211. Redwood, CA: Addison-Wesley.

- Cook, Thomas (1969). *The pronunciation of Efik for speakers of English*. Bloomington, IN: Indiana University Press.
- Cook, Thomas (1985). *An integrated phonology of Efik*. Doctoral dissertation, University of Amsterdam.
- Coupé, Christophe, Egidio Marsico & François Pellegrino (2009). Structural complexity of phonological systems. In: Pellegrino, François, Egidio Marsico, Ioana Chitoran & Christophe Coupé (eds.), *Approaches to phonological complexity*, 141–169. Berlin: Mouton de Gruyter.
- Coupé, Christophe, Egidio Marsico & François Pellegrino (2017). To what extent are phonological inventories complex systems? In: Mufwene, Salikoko, Christophe Coupé & François Pellegrino (eds.), *Complexity in language: developmental and evolutionary perspectives*, 135–164. Cambridge: Cambridge University Press.
- Crasborn, Onno (2001). *Phonetic implementation of phonological categories in Sign Language of the Netherlands*. Doctoral dissertation, Leiden University.
- Cristia, Alejandrina (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition* 170: 312–327.
- Cristià, Alejandrina & Amanda Seidl (2008). Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development* 4: 203–227.
- Culbertson, Jennifer & Elissa Newport (2015). Harmonic biases in child learners: in support of language universals. *Cognition* 139: 71–82.
- Cummings, Alycia, John Madden & Kathryn Hefta (2017). Converging evidence for [coronal] underspecification in English-speaking adults. *Journal of Neurolinguistics* 44: 147–162.
- Dahl, Östen (2004). *The growth and maintenance of linguistic complexity*. Amsterdam/Philadelphia: John Benjamins.
- Dale, Rick & Gary Lupyan (2012). Understanding the origins of morphological diversity: the linguistic niche hypothesis. *Advances in Complex Systems* 15: 1150017.
- de Boer, Bart (2000). Self-organization in vowel systems. *Journal of Phonetics* 28: 441–465.
- de Boer, Bart (2001). *The origins of vowel systems*. Oxford: Oxford University Press.
- de Groot, A. Willem (1931). Phonologie und Phonetik als Funktionswissenschaften. *Travaux du Cercle Linguistique de Prague* 4: 116–147.
- de Groot, A. Willem (1948). Structural linguistics and phonetic law. *Lingua* 1: 175–208.

- Delgutte, Bertrand (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In: Carlson, Rolf & Björn Granström (eds.), *The representation of speech in the peripheral auditory system*, 131–150. Amsterdam: Elsevier.
- Delgutte, Bertrand (1997). Auditory neural processing of speech. In: Hardcastle, William & John Laver, *The handbook of phonetic sciences*, 507–538. Oxford: Blackwell.
- Demany, Laurent & Françoise Armand (1984). The perceptual reality of tone chroma in early infancy. *Journal of the Acoustical Society of America* 76: 57–66.
- Dijkstra, Nienke & Paula Fikkert (2011). Universal constraints on the discrimination of place of articulation? Asymmetries in the discrimination of ‘paan’ and ‘taan’ by 6-month-old Dutch infants. In: Danis, Nick, Kate Mesh & Hyunsuk Sung (eds.), *BUCLD 35: Proceedings of the 35th annual Boston University Conference on Language Development*, 170–182. Somerville, MA: Cascadilla Press.
- Dik, Simon (1989). *The theory of Functional Grammar*. Dordrecht: Foris.
- Dingemanse, Mark, Will Schuerman, Eva Reinisch, Sylvia Tufvesson & Holger Mitterer (2016). What sound symbolism can and cannot do: testing the iconicity of ideophones from five languages. *Language* 92: e117–e133.
- Dingemanse, Mark, Francisco Torreira & Nick Enfield (2013). Is “Huh?” a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PLoS ONE* 8: e78723.
- Dixon, Robert (1977). *A grammar of Yidjñ*. Cambridge: Cambridge University Press.
- Donohue, Mark & Johanna Nichols (2011). Does phoneme inventory size correlate with population size? *Linguistic Typology* 15: 161–170.
- Dryer, Matthew (2013). Polar questions. In: Dryer, Matthew & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at <http://wals.info/chapter/116>.
- Dryer, Matthew & Martin Haspelmath (eds.) (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at <http://wals.info>.
- Edmunds, Charlotte & Andy Wills (2016). Modeling category learning using a dual-system approach: a simulation of Shepard, Hovland and Jenkins (1961) by COVIS. *Proceedings of the Annual Meeting of the Cognitive Science Society* 38: 69–74.
- Ehret, Katharina (2018). Kolmogorov complexity as a universal measure of language complexity. In: Berdicevskis, Aleksandrs & Christian Bentz (eds.), *Proceedings of the First Shared Task on Measuring Language Complexity*, 8–14.

- Ehret, Katharina & Benedikt Szmrecsanyi (2019). Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Learning* 35: 23–45.
- Ernestus, Mirjam (2000). *Voice assimilation and segment reduction in casual Dutch: a corpus-based study of the phonology–phonetics interface*. Doctoral dissertation, Free University Amsterdam.
- Escudero, Paola & Paul Boersma (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition* 26: 551–585.
- Evans, Nicholas & Stephen Levinson (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32: 429–492.
- Everett, Daniel (1986). Pirahã. In: Derbyshire, Desmond & Geoffrey Pullum (eds.), *Handbook of Amazonian Languages 1*, 200–325. Berlin: Mouton de Gruyter.
- Favaro, Livio, Marco Gamba, Eleonora Cresta, Elena Fumagalli, Francesca Bandoli, Cristina Pilenga, Valentina Isaja, Nicolas Mathevon & David Reby (2020). Do penguins' vocal sequences conform to linguistic laws? *Biology Letters* 16.
- Feldman, Jacob (2000). Minimization of Boolean complexity in human concept learning. *Nature* 407: 630–633.
- Feldman, Naomi, Thomas Griffiths & James Morgan (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*: 2208–2213.
- Ferdinand, Vanessa, Simon Kirby & Kenny Smith (2019). The cognitive roots of regularization in language. *Cognition* 184: 53–68.
- Ferguson, Brock & Casey Lew-Williams (2014). Communicative signals promote abstract rule learning by 7-month-old infants. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.
- Ferrara, Lindsay & Anna-Lena Nilsson (2017). Describing spatial layouts as an L2M2 signed language learner. *Sign Language & Linguistics* 20: 1–26.
- Ferrero, Guillaume (1894). L'inertie mentale et la loi du moindre effort. *Revue Philosophique de la France et de l'Étranger* 37: 169–182.
- Fiorito, Graziano & Pietro Scotto (1992). Observational learning in *Octopus vulgaris*. *Science* 256: 545–546.
- Firchow, Irwin & Jacqueline Firchow (1969). An abbreviated phoneme inventory. *Anthropological Linguistics* 11: 271–276.

- Flemming, Edward (1995/2002). *Auditory representations in phonology*. Doctoral dissertation, University of California Los Angeles. New York & London: Routledge.
- Fowler, Carol, Philip Rubin, Robert Remez & Michael Turvey (1980). Implications for speech production of a general theory of action. In: Butterworth, Brian (ed.), *Language Production*, 373–420. New York: Academic Press.
- Fudge, Erik (1967). The nature of phonological primes. *Journal of Linguistics* 3: 1–36.
- Golla, Victor (1970). *Hupa language dictionary*. Hoopa, CA: Na:tinixwe Mixine:whe’.
- Gray, Tyler, Andrew Reagan, Peter Sheridan Dodds & Christopher Danforth (2018). English verb regularization in books and tweets. *PLoS ONE* 13 (12): e0209651.
- Greenberg, Joseph (1963). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, Joseph (ed.), *Universals of language*, 110–113. Cambridge, MA: MIT Press.
- Griffiths, Thomas, Brian Christian & Michael Kalish (2008). Using category structures to test iterated learning as a method for revealing inductive biases. *Cognitive Science* 32: 68–107.
- Grossberg, Stephen (1969). Embedding fields: a theory of learning with physiological implications. *Journal of Mathematical Psychology* 6 (2): 209–239.
- Grossberg, Stephen (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics* 23: 121–134.
- Grossberg, Stephen (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science* 11: 23–63.
- Hall, Daniel Currie (2007). *The role and representation of contrast in phonological theory*. Doctoral dissertation, University of Toronto.
- Hall, Daniel Currie (2011). Phonological contrast and its phonetic enhancement: dispersedness without dispersion. *Phonology* 28: 1–54.
- Hall, T. Alan (1997). *The phonology of coronals*. Amsterdam/Philadelphia: John Benjamins.
- Hammarström, Harald, Robert Forkel & Martin Haspelmath (2019). *Glottolog 3.4*. Jena: Max Planck Institute for the Science of Human History. Available online at <http://glottolog.org>.
- Harris, John & Geoff Lindsey (1995). The elements of phonological representation. In: Durand, Jacques & Francis Katamba (eds.), *Frontiers of phonology: atoms, structures, derivations*, 34–79. London/New York: Longman.
- Haspelmath, Martin (2012). How to compare major word-classes across the world’s languages. In: Graf, Thomas, Denis Paperno, Anna Szabolcsi & Jos Tellings

- (eds.), *Theories of everything: in honor of Edward Keenan*, 109–130. UCLA Working Papers in Linguistics 17.
- Haspelmath, Martin, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.) (2001). *Language typology and language universals*. Berlin/New York: Walter de Gruyter.
- Hauser, Marc, Noam Chomsky & Tecumseh Fitch (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* 298: 1569–1579.
- Hawkins, John (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.
- Hay, Jennifer & Laurie Bauer (2007). Phoneme inventory size and population size. *Language* 83: 388–400.
- Hengeveld, Kees & Sterre Leufkens (2018). Transparent and non-transparent languages. *Folia Linguistica* 52: 139–175/
- Hengeveld, Kees & J. Lachlan Mackenzie (2008). *Functional Discourse Grammar: a typologically-based theory of language structure*. Oxford: Oxford University Press.
- Hobaiter, Catherine, Timothée Poisot, Klaus Zuberbühler, William Hoppitt & Thibaud Gruber (2014). Social network analysis shows direct evidence for social transmission of tool use in wild chimpanzees. *PLoS Biology* 12 (9): e1001960.
- Hockett, Charles (1955). *A manual of phonology*. Baltimore, MD: Waverly Press.
- Hockett, Charles (1963). The problem of universals in language. In: Greenberg, Joseph (ed.), *Universals of language*, 1–29. Cambridge, MA: MIT Press.
- Hoff, Erika (2006). How social contexts support and shape language development. *Developmental Review* 26 (1): 55–88.
- Honeybone, Patrick (2005). Diachronic evidence in segmental theory: the case of obstruent laryngeal specifications. In: van Oostendorp, Marc & Jeroen van de Weijer (eds.), *The internal organization of phonological segments*, 317–352. Berlin: Mouton de Gruyter.
- Honeybone, Patrick (2016). Are there impossible changes? $\theta > f$ but $f \not> \theta$. *Papers in Historical Phonology* 1: 316–358.
- Hopper, Paul (1973). Glottalized and murmured occlusives in Indo-European. *Glossa* 7: 141–166.
- Hopper, Paul & Elizabeth Traugott (2003). *Grammaticalization*. Cambridge: Cambridge University Press.
- Horner, Victoria, Andrew Whiten, Emma Flynn & Frans de Waal (2006). Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences* 103: 13878–13883.

- Hudson Kam, Carla & Elissa Newport (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development* 1: 151–195.
- Hudson Kam, Carla & Elissa Newport (2009). Getting it right by getting it wrong: when learners change languages. *Cognitive Psychology* 59: 30–66.
- Hullebus, Marc, Stephen Tobin & Adamantios Gafos (2018). Speaker-specific structure in German voiceless stop voice onset times. *Proceedings of Interspeech 2018*: 1403–1407.
- Hurford, James & Simon Kirby (2002). The emergence of linguistic structure: an overview of the iterated learning model. In: Parisi, Domenico & Angelo Cangelosi (eds.), *Simulating the evolution of language*, 121–147. Berlin: Springer Verlag.
- Hyman, Larry (2008). Universals in phonology. *The Linguistic Review* 25: 83–137.
- Hyman, Larry (2018). What is phonological typology? In: Hyman, Larry & Frans Plank (eds.), *Phonological typology*, 1–20. Berlin: De Gruyter Mouton.
- Iverson, Gregory & Joseph Salmons (1999). Glottal spreading bias in Germanic. *Linguistische Berichte* 178: 135–151.
- Jacob, Michal & Shaul Hochstein (2008). Set recognition as a window to perceptual and cognitive processes. *Perception & Psychophysics* 70 (7): 1165–1184.
- Jäger, Gerhard (2003). Learning constraint subhierarchies: the Bidirectional Gradual Learning Algorithm. In: Zeevat, Henk & Reinhard Blutner (eds.), *Optimality Theory and pragmatics*, 251–287. Basingstoke: Palgrave Macmillan.
- Jakobson, Roman (1941). *Kindersprache, Aphasie und allgemeine Lautgesetze*. Uppsala: Almqvist & Wiksell.
- Jakobson, Roman (1958). Typological studies and their contribution to historical linguistics. *Proceedings of the 8th International Congress of Linguists (Oslo 1957)*: 17–35.
- Jakobson, Roman, Gunnar Fant & Morris Halle (1952). *Preliminaries to speech analysis: the distinctive features and their correlates*. Cambridge, MA: MIT Press.
- Jakobson, Roman, Serge Karcevsky & Nikolai Trubetzkoy (1928). Quelles sont les méthodes les mieux appropriées à un exposé complet et pratique de la grammaire d'une langue quelconque? In: de Boer, Cornelis, Jacobus van Ginneken & Anton van Hamel (eds.), *Actes du Premier Congrès International de Linguistes*, 33–36. Leiden: A. W. Sijthoff.
- Jelinek, Eloise (1995). Quantification in Straits Salish. In: Bach, Emmon, Eloise Jelinek, Angelika Kratzer & Barbara Partee (eds.), *Quantification in natural languages*, 487–540. Dordrecht: Kluwer Academic Publishers.

- Jessen, Michael & Catherine Ringen (2002). Laryngeal features in German. *Phonology* 19: 1–30.
- Johnson, Keith (2004). Massive reduction in conversational American English. In: Yoneyama, Kiyoko & Kikuo Maekawa (eds.), *Spontaneous speech: data and analysis. Proceedings of the 1st Session of the 10th International Symposium*, 29–54. Tokyo: The National International Institute for Japanese Language.
- Jones, Daniel (1919). The phonetic structure of the Sechuana language. *Transactions of the Philological Society* 28 (1): 99–106.
- Jones, Daniel (1931). On phonemes. *Travaux du Cercle Linguistique de Prague* 4: 74–78.
- Jost, Jürgen (2004). External and internal complexity of complex adaptive systems. *Theory in Biosciences* 123: 69–88.
- Juola, Patrick (1998). Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics* 5: 206–213.
- Kabak, Baris (2004). Acquiring phonology is not acquiring inventories but acquiring contrasts: the loss of Turkic and Korean primary long vowels. *Linguistic Typology* 8: 351–368.
- Kalish, Michael, Thomas Griffiths & Stephan Lewandowsky (2007). Iterated learning: intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review* 14 (2): 288–294.
- Keating, Patricia (1984). Phonetic and phonological representation of stop consonant voicing. *Language* 60: 286–319.
- Kegl, Judy (2002). Language emergence in a language-ready brain: acquisition issues. In: Morgan, Gary & Bencie Woll (eds.), *Language Acquisition in Signed Languages*, 207–254. Cambridge: Cambridge University Press.
- Kempe, Vera, Nicolas Gauvrit & Douglas Forsyth (2015). Structure emerges faster during cultural transmission in children than in adults. *Cognition* 136: 247–254.
- Kemps, Rachel, Mirjam Ernestus, Robert Schreuder & R. Harald Baayen (2004). Processing reduced word forms: the suffix restoration effect. *Brain and Language* 19: 117–127.
- Kiang, Nelson (1980). Processing of speech by the auditory nervous system. *Journal of the Acoustical Society of America* 68: 830–835.
- Kirby, Simon, Hannah Cornish & Kenny Smith (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences* 105 (31): 10681–10686.

- Kirby, Simon & James Hurford (2002). The emergence of linguistic structure: an overview of the iterated learning model. In: Cangelosi, Angelo & Domenico Parisi (eds.), *Simulating the evolution of language*, 121–148. London: Springer Verlag.
- Kirby, Simon, Monica Tamariz, Hannah Cornish & Kenny Smith (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition* 141: 87–102.
- Kirchner, Robert (1998/2001). *An effort-based approach to consonant lenition*. Doctoral dissertation, University of California Los Angeles. Rutgers Optimality Archive 276. Published in 2001, New York & London: Routledge.
- Kolmogorov, Andrey (1963). On tables of random numbers. *Sankhyā, The Indian Journal of Statistics, Series A* 25: 369–375.
- Koopmans-Van Beinum, Florian (1980). *Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions*. Doctoral dissertation, University of Amsterdam.
- Kretschmer, Paul (1932). Die Urgeschichte der Germanen und die germanische Lautverschiebung. *Wiener prähistorische Zeitschrift* 19: 269–280.
- Kruschke, John (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review* 99 (1): 22–44.
- Kuhl, Patricia (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics* 50: 93–107.
- Kümmel, Martin (2007). *Konsonantenwandel: Bausteine zu einer Typologie des Lautwandels und ihre Konsequenzen für die vergleichende Rekonstruktion*. Wiesbaden: Reichert Verlag.
- Kuo, Li-Jen (2009). The role of natural class features in the acquisition of phonotactic regularities. *Journal of Psycholinguistic Research* 38 (2): 129–150.
- Kuzla, Claudia & Mirjam Ernestus (2011). Prosodic conditioning of phonetic detail in German plosives. *Journal of Phonetics* 39: 143–155.
- Kuznetsova, Alexandra, Per Brockhoff & Rune Christensen (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* 82: 1–26.
- Lahiri, Aditi & Henning Reetz (2002). Underspecified recognition. In: Gussenhoven, Carlos & Natasha Warner (eds.), *Laboratory Phonology 7*, 637–676. Berlin: Mouton.
- Lahiri, Aditi & Henning Reetz (2010). Distinctive features: phonological underspecification in representation and processing. *Journal of Phonetics* 38: 44–59.
- Lahiri, Aditi & Sibrand van Coillie (1999). *Non-mismatching features in language comprehension*. Unpublished manuscript, University of Konstanz.

- Lammertink, Imme, Paul Boersma, Frank Wijnen & Judith Rispens (2019). Children with developmental language disorder have an auditory verbal statistical learning deficit: evidence from an online measure. *Language Learning* 70: 137–178.
- Langacker, Ronald (1977). Syntactic reanalysis. In: Li, Charles (ed.), *Mechanisms of syntactic change*, 57–139. Austin, TX: University of Texas Press.
- Lass, Roger (2000). Phonology and morphology. In: Lass, Roger (ed.), *The Cambridge history of the English language (Volume 3: 1746–1776)*, 56–186. Cambridge: Cambridge University Press.
- Lehmann, Winfred (1952). *Proto-Indo-European Phonology*. Austin, TX: University of Texas Press and Linguistic Society of America.
- Levelt, Clara, Niels Schiller & Willem Levelt (1999). A developmental grammar for syllable structure in the production of child language. *Brain and Language* 68: 291–299.
- Levering, Kimery, Nolan Conaway & Kenneth Kurtz (2020). Revisiting the linear separability constraint: new implications for theories of human category learning. *Memory & Cognition* 48: 335–347.
- Liberman, Alvin, Franklin Cooper, Donald Shankweiler & Michael Studdert-Kennedy (1967). Perception of the speech code. *Psychological Review* 74: 431–461.
- Liberman, Alvin, Katherine Safford Harris, Howard Hoffman & Belver Griffith (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54: 358–368.
- Liberman, Alvin & Ignatius Mattingly (1985). The motor theory of speech perception revised. *Cognition* 21: 1–36.
- Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin Nowak (2007). Quantifying the evolutionary dynamics of language. *Nature* 449: 713–716.
- Liljencrants, Johan & Björn Lindblom (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Linguistics* 48: 839–862.
- Lindblom, Björn, Randy Diehl, Sang-Hoon Park & Giampiero Salvi (2011). Sound systems are shaped by their users: the recombination of phonetic substance. In: Clements, G. Nick & Rachid Ridouane (eds.), *Where do phonological features come from? Cognitive, physical and developmental biases of distinctive speech categories*, 67–98. Amsterdam/Philadelphia: John Benjamins.
- Lisker, Leigh (2001). Hearing the Polish sibilants [š ś s]: phonetic and auditory judgments. In: Grønnum, Nina & Jørgen Rischel (eds.), *To honour Eli Fischer-Jørgensen. Travaux du cercle linguistique de Copenhague XXXI*, 226–238. Copenhagen: Reitzel.

- Lisker, Leigh & Arthur Abramson (1964). A cross-linguistic study of voicing in initial stops: acoustical measurements. *Word* 20: 384–422.
- Long, Michael (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics* 4: 126–141.
- Lorente de Nó, Rafael (1938). Cerebral cortex: architecture, intracortical connections, motor projections. In: Fulton, John (ed.), *Physiology of the nervous system*, 291–327. London: Oxford University Press.
- Louw, Jacobus (1962). On the segmental phonemes of Zulu. *Afrika und Übersee* 46: 43–93.
- Love, Bradley (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review* 9 (4): 829–835.
- Luce, R. Duncan (1986). *Response times: their role in inferring elementary mental organization*. New York: Oxford University Press.
- Luick, Karl (1940). *Historische Grammatik der englischen Sprache*. Leipzig: Bernhard Tauchnitz.
- Lupyan, Gary & Rick Dale (2010). Language structure is partly determined by social structure. *PloS ONE* 5 (1): e8559.
- Mackie, Scott & Jeff Mielke (2011). Feature economy in natural, random, and synthetic inventories. In: Clements, G. Nick & Rachid Ridouane (eds.), *Where do phonological features come from? Cognitive, physical and developmental biases of distinctive speech categories*, 43–66. Amsterdam/Philadelphia: John Benjamins.
- Maddieson, Ian (1980). Phonological generalizations from the UCLA Phonological Segment Inventory Database (UPSID). *UCLA Working Papers in Phonetics* 50: 57–68.
- Maddieson, Ian (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.
- Maddieson, Ian (1995). Gestural economy. In: Elenius, Kjell & Peter Branderud (eds.), *Proceedings of the 13th International Congress of Phonetic Sciences*, 574–577. Stockholm: Stockholm University Press.
- Maddieson, Ian (2007). Issues of phonological complexity: statistical analysis of the relationship between syllable structures, segment inventories, and tone contrasts. In: Solé, Maria-Josep, Patrice Speeter Beddor and Manjari Ohala (eds.), *Experimental approaches to phonology*, 93–103. Oxford: Oxford University Press.
- Maddieson, Ian (2013a). Consonant inventories. In: Dryer, Matthew & Martin Haspelmath, (eds.), *The World Atlas of Language Structures Online*, chapter 1. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at <https://wals.info/chapter/1>.

- Maddieson, Ian (2013b). Absence of common consonants. In: Dryer, Matthew & Martin Haspelmath, (eds.), *The World Atlas of Language Structures Online*, chapter 18. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at <https://wals.info/chapter/18>.
- Maddieson, Ian (2013c). Voicing and gaps in plosive systems. In: Dryer, Matthew & Martin Haspelmath, (eds.), *The World Atlas of Language Structures Online*, chapter 5. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at <https://wals.info/chapter/5>.
- Marcus, Gary, Keith Fernandes & Scott Johnson (2007). Infant rule learning facilitated by speech. *Psychological Science* 18: 387–391.
- Marcus, Gary, Sujith Vijayan, Shoba Bandi Rao & Peter Vishton (1999). Rule learning by seven-month-old infants. *Science* 283: 77–80.
- Martinet, André (1955). *Économie des changements phonétiques: traité de phonologie diachronique*. Berne: Francke.
- Martinet, André (1960). *Éléments de linguistique générale*. Paris: Armand Colin.
- Martinet, André (1962). *A functional view of language*. Oxford: Oxford University Press.
- Mathesius, Vilém (1931). Zum Problem der Belastungs- und Kombinationsfähigkeit der Phoneme. *Travaux du Cercle Linguistique de Prague* 4: 148–152.
- McCarthy, John & Alan Prince (1995). Faithfulness and reduplicative identity. In: Beckman, Jill, Laura Walsh Dickey & Suzanne Urbanczyk (eds.), *Papers in Optimality Theory*, 249–384. University of Massachusetts Occasional Papers 18. Amherst, MA: Graduate Linguistic Student Association.
- McCloy, Daniel & Adrian Lee (2019). Investigating the fit between phonological feature systems and brain responses to speech using EEG. *Language, Cognition and Neuroscience* 34: 662–676.
- McGurk, Harry & John MacDonald (1976). Hearing lips and seeing voices. *Nature* 264: 746–748.
- McMurray, Bob, Richard Aslin & Joseph Toscano (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science* 12: 369–378.
- McWhorter, John (1998). Identifying the creole prototype: vindicating a typological class. *Language* 74: 788–817.
- McWhorter, John (2001). The world's simplest grammars are creole grammars. *Linguistic Typology* 5: 125–166.

- Medin, Douglas & Paula Schwanenflugel (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory* 7: 355–368.
- Mensah, Eyo & Eyamba Mensah (2014). The adaptation of English consonants by Efik learners of English. *English Language Teaching* 7: 38–49.
- Mehr, Samuel, Manvir Singh, Dean Knox, Daniel Ketter, Daniel Pickens-Jones, Stephanie Atwood, Christopher Lucas, Nori Jacoby, Alena Egner, Erin Hopkins, Rhea Howard, Joshua Hartshorne, Mariela Jennings, Jan Simson, Constance Bainbridge, Steven Pinker, Timothy O'Donnell, Max Krasnow & Luke Glowacki (2019). Universality and diversity in human song. *Science* 366: eaax0868.
- Meisel, Jürgen (2011). Bilingual language acquisition and theories of diachronic change: bilingualism as cause and effect of grammatical change. *Bilingualism: Language and Cognition* 14: 121–145.
- Meissner, Martin, Stuart Philpott & Diana Philpott (1975). The sign language of sawmill workers in British Columbia. *Sign Language Studies* 9: 291–308.
- Mesch, Johanna (2001). *Tactile sign language: turn taking and questions in signed conversations of Deaf-Blind people*. International Studies on Sign Language and Communication of the Deaf 38. Hamburg: Signum.
- Mielke, Jeff (2008). *The emergence of distinctive features*. Oxford Studies in Typology and Linguistic Theory. Oxford: Oxford University Press.
- Moisik, Scott & Dan Dediu (2017). Anatomical biasing and clicks: evidence from biomechanical modeling. *Journal of Language Evolution* 2: 37–51.
- Monaghan, Padraic (2017). Canalization of language structure from environmental constraints: a computational model of word learning from multiple cues. *Topics in Cognitive Science* 9: 21–34.
- Monaghan, Padraic, James Brand, Rebecca Frost & Gemma Taylor (2017). Multiple variable cues in the environment promote accurate and robust word learning. *Proceedings of the 39th Cognitive Science Society Conference*: 817–822.
- Moore, Brian & Brian Glasberg (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America* 74: 750–753.
- Moran, Steven & Damián Blasi (2014). Cross-linguistic comparison of complexity measures in phonological systems. In: Newmeyer, Frederick & Laurel Preston (eds.), *Measuring grammatical complexity*, 217–240. Oxford: Oxford University Press.
- Moran, Steven & Daniel McCloy (eds.) (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. Available online at phoible.org.

- Moran, Steven, Daniel McCloy and Richard Wright (2012). Revisiting population size vs. phoneme inventory size. *Language* 88: 877–893.
- Moreton, Elliot, Joe Pater & Katya Pertsova (2017). Phonological concept learning. *Cognitive Science* 41: 4–69.
- Moulin-Frier, Clément, Julien Diard, Jean-Luc Schwartz & Pierre Bessière (2015). COSMO (“Communicating about Objects using Sensory-Motor Operations”): a Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics* 53: 5–41.
- Mufwene, Salikoko (2001). *The ecology of language evolution*. Cambridge: Cambridge University Press.
- Mugdan, Joachim (2014). More on the origins of the term *phonème*. *Historiographica Linguistica* 41: 185–187.
- Muysken, Pieter & Norval Smith (2008). The study of pidgin and creole languages. In: Arends, Jacques, Pieter Muysken & Norval Smith (eds.), *Pidgins and creoles: an introduction*, 3–14. Amsterdam: John Benjamins.
- Nichols, Johanna (1992). *Linguistic diversity in time and space*. Chicago: Chicago University Press.
- Nordlinger, Rachel & Louisa Sadler (2004). Nominal tense in cross-linguistic perspective. *Language* 80: 776–806.
- Noske, Roland (2012). The Grimm–Verner push chain and Contrast Preservation Theory. In: Botma, Bert & Roland Noske (eds.), *Phonological explorations: empirical, theoretical and diachronic issues*, 63–86. Berlin: De Gruyter.
- Nosofsky, Robert, Mark Gluck, Thomas Palmeri, Stephen McKinley & Paul Glauthier (1994). Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland and Jenkins (1961). *Memory & Cognition* 22: 352–369.
- Nosofsky, Robert & Thomas Palmeri (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review* 3: 222–226.
- Ohala, John (1980). Moderator’s summary of ‘Symposium no.1: phonetic universals in phonological systems and their explanation’. *Proceedings of the 9th International Congress of Phonetic Sciences*: 181–194.
- Ohala, John (1983). The origin of sound patterns in vocal tract constraints. In: MacNeilage, Peter (ed.), *The production of speech*, 189–216. Heidelberg/Berlin: Springer Verlag.
- Ohala, John (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America* 99: 1718–1725.

- Ohala, John (2009). Languages' sound inventories: the devil in the details. In: Pellegrino, François, Egidio Marsico, Ioana Chitoran & Christophe Coupé (eds.), *Approaches to phonological complexity*, 47–58. Berlin: Mouton de Gruyter.
- Ohala, John & Carol Riordan (1979). Passive vocal tract enlargement during voiced stops. In: Wolf, Jared & Dennis Klatt (eds.), *Speech communication papers presented at the 97th meeting of the Acoustical Society of America*. New York: Acoustical Society of America.
- Ortega, Gerardo & Gary Morgan (2010). Comparing child and adult development of a visual phonological system. *Language, Interaction and Acquisition* 1: 67–81.
- Ouddeken, Nina (2018). *Voicing in transition: laryngeal characteristics in West-Germanic and Italo-Romance dialects*. Doctoral dissertation, Radboud University, Nijmegen.
- Padgett, Jaye (2001). Contrast dispersion and Russian palatalization. In: Hume, Elizabeth & Keith Johnson (eds.), *The role of speech perception in phonology*, 187–218. San Diego, CA: Academic Press.
- Padgett, Jaye (2003). Contrast and post-velar fronting in Russian. *Natural Language and Linguistic Theory* 21: 39–87.
- Paget, Richard (1931). The gestural origin of language. *Actes du Deuxième Congrès de Linguistes*: 172–176. Paris: Adrien Maisonneuve.
- Pajak, Bozena, Klinton Bicknell & Roger Levy (2013). A model of generalization in distributional learning of phonetic categories. In: Demberg, Vera & Roger Levy (eds.), *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics*, 11–20. Sofia: Association for Computational Linguistics.
- Palmer, Stephanie, Laurel Fais, Roberta Golinkoff & Janet Werker (2012). Perceptual narrowing of linguistic sign occurs in the 1st year of life. *Child Development* 83: 543–553.
- Passy, Paul (1890). *Étude sur les changements phonétiques et leurs caractères généraux*. Paris: Firmin-Didot.
- Pellegrino, François, Egidio Marsico, Ioana Chitoran & Christophe Coupé (eds.) (2009). *Approaches to phonological complexity*. Berlin: Mouton de Gruyter.
- Pericliev, Vladimir (2004). There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language. *Linguistic Typology* 8: 376–383.
- Pierrehumbert, Janet (2001). Exemplar dynamics: word frequency, lenition and contrast. In: Bybee, Joan & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–157. Amsterdam/Philadelphia: Benjamins.

- Pinker, Steven & Paul Bloom (1990). Natural language and natural selection. *Behavioral and Brain Sciences* 13: 707–726.
- Prince, Alan & Paul Smolensky (1993/2004). *Optimality Theory: constraint interaction in Generative Grammar*. Malden, MA: Wiley-Blackwell.
- Pycha, Anne, Pawel Nowak, Eurie Shin & Ryan Shosted (2003). Phonological rule-learning and its implications for a theory of vowel harmony. In: Tsujimura, Mimura & Gina Garding (eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics* 22, 101–114. Somerville, MA: Cascadilla Press.
- R Core Team (2019). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org>
- Ravignani, Andrea, Bill Thompson, Thomas Grossi, Tania Delgado & Simon Kirby (2018). Evolving building blocks of rhythm: how human cognition creates music via cultural transmission. *Annals of the New York Academy of Sciences* 1423: 176–187.
- Real, Florencia & Thomas Griffiths (2009). The evolution of frequency distributions: relating regularization to inductive biases through iterated learning. *Cognition* 111: 317–328.
- Ren, Jie, Uriel Cohen Priva & James Morgan (2019). Underspecification in toddlers' and adults' lexical representations. *Cognition* 193: 103991.
- Rialland, Annie (2005). Phonological and phonetic aspects of whistled languages. *Phonology* 22: 237–271.
- Roberts, Séan & Stephen Levinson (2017). Conversation, cognition and cultural evolution: a model of the cultural evolution of word order through pressures imposed from turn taking in conversation. *Interaction Studies* 18: 402–442.
- Rodriguez-Cuadrado, Sara, Cristina Baus & Albert Costa (2018). Foreigner talk through word reduction in native/non-native spoken interactions. *Bilingualism: Language and Cognition* 21: 419–426.
- Rothermich, Kathrin, Havan Harris, Kerry Sewell & Susan Bobb (2019). Listener impressions of foreigner-directed speech: a systematic review. *Speech Communication* 112: 22–29.
- Rumelhart, David & David Zipser (1985). Feature discovery by competitive learning. *Cognitive Science* 9: 75–112.
- Saffran, Jenny, Richard Aslin & Elissa Newport (1996). Statistical learning in 8-month-old infants. *Science* 274: 1926–1928.
- Saffran, Jenny & Erik Thiessen (2003). Pattern induction by infant language learners. *Developmental Psychology* 39: 484–494.

- Samuels, Bridget (2011). A minimalist program for phonology. In: Boeckx, Cedric (ed.), *The Oxford handbook of linguistic minimalism*, ch. 25. Oxford: Oxford University Press.
- Sandler, Wendy (1989) *Phonological representation of the sign: linearity and non-linearity in American Sign Language*. Dordrecht: Foris.
- Sandler, Wendy, Irit Meir, Carol Padden & Mark Aronoff (2005). The emergence of grammar: systematic structure in a new language. *Proceedings of the National Academy of Sciences* 102: 2661–2665.
- Sapir, Edward (1921). *Language: an introduction to the study of speech*. New York: Harcourt Brace.
- Sapir, Edward (1933). La réalité psychologique des phonèmes. *Journal de Psychologie* 30: 247–265.
- Savage, Patrick, Steven Brown, Emi Sakai & Thomas Currie (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences* 112: 8987–8992.
- Scharinger, Mathias, Ulrike Domahs, Elise Klein & Frank Domahs (2016). Mental representations of vowel features asymmetrically modulate activity in superior temporal sulcus. *Brain and Language* 163: 42–49.
- Schluter, Kevin, Stephen Politzer-Ahles & Diogo Almeida (2016). No place for /h/: an ERP investigation of English fricative place features. *Language, Cognition and Neuroscience* 31: 728–740.
- Schuttenhelm, Lisan (2013). *Perception of the non-native phone [g] in Dutch: where lies the VOT boundary?* MA thesis, University of Amsterdam.
- Seifart, Frank, Julien Meyer, Sven Grawunder & Laure Dentel (2018). Reducing language to rhythm: Amazonian Bora drummed language exploits speech rhythm for long-distance communication. *Open Science* 5: 170354.
- Seinhorst, Klaas (2012). *The evolution of auditory dispersion in symmetric neural nets*. MA thesis, University of Amsterdam.
- Seinhorst, Klaas (2016a). Mind the gap: inductive biases in phonological feature learning. In: Roberts, Séan, Christine Cuskley, Luke McCrohon, Lluís Barceló-Coblijn, Olga Fehér & Tessa Verhoef (eds.), *The Evolution of Language: Proceedings of the 11th International Conference*, paper 155.
- Seinhorst, Klaas (2016b). System complexity and (im)possible sound changes. *Papers in Historical Phonology* 1: 238–249.
- Seinhorst, Klaas (2017). Feature economy versus logical complexity in phonological pattern learning. *Language Sciences* 60: 69–79.

- Seinhorst, Klaas, Paul Boersma & Silke Hamann (2019). Iterated distributional and lexicon-driven learning in a symmetric neural network explains the emergence of features and dispersion. In: Calhoun, Sasha, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences*, paper 779. Canberra: Australasian Speech Science and Technology Association.
- Seinhorst, Klaas & Floor van de Leur (under review). *Feature economy versus logical complexity in plosive inventories*.
- Selten, Reinhard & Massimo Warglien (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences* 104: 7361–7366.
- Senghas, Ann, Sotaro Kita & Asli Özyürek (2004). Children creating core properties of language: evidence from an emerging sign language in Nicaragua. *Science* 305: 1779–1782.
- Shalizi, Cosma (2006). Methods and techniques of complex systems science: an overview. In: Deisboeck, Thomas & J. Yasha Kresh (eds.), *Complex systems science in biomedicine*, 33–114. New York: Springer.
- Sharpe, Margaret (2005). *Grammar and texts of the Yugambeh-Bundjalung dialect chain in Eastern Australia*. München: Lincom.
- Shepard, Roger, Carl Hovland & Herbert Jenkins (1961). Learning and memorization of classifications. *Psychological Monographs* 75 (13).
- Singleton, Jenny & Elissa Newport (2004). When learners surpass their models: the acquisition of American Sign Language from inconsistent input. *Cognitive Psychology* 49: 370–407.
- Siviter, Harry, Charles Deeming, Marjolein van Giezen & Anna Wilkinson (2017). Incubation environment impacts the social cognition of adult lizards. *Royal Society Open Science* 4: 170742.
- Skoruppa, Katrin & Sharon Peperkamp (2011). Adaptation to novel accents: feature-based learning of context-sensitive phonological regularities. *Cognitive Science* 35: 348–366.
- Smith, Kenny, Katya Abramova & Simon Kirby (2012). *Gesture and the iterated learning paradigm*. Presentation, Amsterdam Gesture Center (Free University), Amsterdam.
- Smith, Kenny & Elizabeth Wonnacott (2010). Eliminating unpredictable variation through iterated learning. *Cognition* 116: 444–449.
- Smith, Norval & Marlou van Rijn (2012). *Free Personal Pronoun System*. <https://doi.org/10.17026/dans-z5k-ub4q>

- Smolensky, Paul (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry* 27: 720–731.
- Solomonoff, Ray (1960). *A preliminary report on a general theory of inductive inference*. United States Air Force, Office of Scientific Research, Report V-131.
- Sommerfelt, Alf (1928). Sur la nature du phonème. *Norsk Tidsskrift for Sprogvidenskap* 1: 22–26.
- Stevens, Kenneth (1972). The quantal nature of speech: evidence from articulatory-acoustic data. In: Denes, Peter & Edward David Jr. (eds.), *Human communication: a unified view*, 51–66. New York [etc.]: McGraw-Hill.
- Stevens, Kenneth (1989). On the quantal nature of speech. *Journal of Phonetics* 17: 3–46.
- Stevens, Kenneth & Samuel Keyser (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics* 38: 10–19.
- Stokoe, William (1960) *Sign language structure. An outline of the visual communication systems of the American Deaf*. Silver Spring, MD: Linstok Press.
- Strenski, Ivan (2006). *Thinking about religion: an historical introduction to theories of religion*. Malden, MA / Oxford: Blackwell.
- Taatgen, Niels, Marcia van Oploo, Jos Braaksma & Jelle Niemantsverdriet (2003). How to construct a believable opponent using cognitive modeling in the game of Set. In: Detje, Frank, Dietrich Dörner & Harald Schaub (eds.), *The logic of cognitive systems: Proceedings of the Fifth International Conference on Cognitive Modeling*, 201–206. Bamberg: Universitäts-Verlag Bamberg.
- ten Bosch, Louis (1991). *On the structure of vowel systems: aspects of an extended vowel model using effort and contrast*. Doctoral dissertation, University of Amsterdam.
- ter Schure, Sophie (2016). *The relevance of visual information on learning sounds in infancy*. Doctoral dissertation, University of Amsterdam.
- Tesar, Bruce (1997). An iterative strategy for learning metrical stress in Optimality Theory. In: Hughes, Elizabeth, Mary Hughes & Annabel Greenhill (eds.), *Proceedings of the 21st Annual Boston University Conference on Language Development*, 615–626. Somerville, MA: Cascadilla.
- Tesar, Bruce & Paul Smolensky (2000). *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.
- Tomasello, Michael (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

- Tomasello, Michael (2009). The usage-based theory of language acquisition. In: Bavin, Edith (ed.), *The Cambridge Handbook of Language Acquisition*, 69–88. Cambridge: Cambridge University Press.
- Trails, Anthony (1985). Phonetic and phonological studies of !Xóõ Bushman. *Quellen zur Khoisan-Forschung* 5. Hamburg: Helmut Buske Verlag.
- Trask, R. Lawrence (2000). *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh University Press.
- Trubetzkoy, Nikolai (1939). Grundzüge der Phonologie. *Travaux du Cercle Linguistique de Prague* 7.
- Tsuji, Sho, Reiko Mazuka, Alejandrina Cristia, Paula Fikkert (2015). Even at 4 months, a labial is a good enough coronal, but not vice versa. *Cognition* 134: 252–256.
- Twaddell, William (1935). *On defining the phoneme*. Language Monographs 16. Baltimore, MD: Waverly Press.
- Usher, Marius & James McClelland (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review* 111: 757–769.
- van de Ven, Marco, Benjamin Tucker & Mirjam Ernestus (2011). Semantic context effects in the comprehension of reduced pronunciation variants. *Memory & Cognition* 39: 1301–1316.
- van de Vijver, Ruben & Dinah Baer-Henney (2012). Voice more, front less: on the development of knowledge of voicing and vowel alternations in German nouns by 5-year-olds, 7-year-olds and adults. In: Biller, Alia, Esther Chung & Amelia Kimball (eds.), *Proceedings of the 36th Boston University Conference on Language Development*, volume 2, 660–672. Sommerville, MA: Cascadilla Press.
- van der Harst, Sander (2011). *The vowel space paradox: a sociophonetic study on Dutch*. Doctoral dissertation, Radboud University Nijmegen.
- van der Hulst, Harry (1993). Units in the analysis of signs. *Phonology* 10: 209–241.
- van der Hulst, Harry (2016). Monovalent ‘features’ in phonology. *Language and Linguistics Compass* 10: 83–102.
- van der Kooij, Els (2002). *Phonological categories in Sign Language of the Netherlands: the role of phonetic implementation and iconicity*. Doctoral dissertation, Leiden University.
- van Witteloostuijn, Merel, Paul Boersma, Frank Wijnen & Judith Rispens (2019). Statistical learning abilities of children with dyslexia across three experimental paradigms. *PLoS ONE* 14 (8): e0220041.
- Vennemann, Theo (1985). The bifurcation theory of the Germanic and German consonant shifts: synopsis and some further thoughts. In: Fisiak, Jacek (ed.),

- Papers from the 6th International Conference on Historical Linguistics*, 527–547. Amsterdam: Benjamins.
- Verner, Karl (1877). Eine Ausnahme der ersten Lautverschiebung. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen* 23: 97–130.
- Vet, Dirk Jan (2013). *ED (Experiment Designer)*. Software. www.fon.hum.uva.nl/dirk/software.php
- Vogt, Paul & Federico Divina (2009). Social symbol grounding and language evolution. In: Belpaeme, Tony, Stephen Cowley & Karl MacDorman (eds.), *Symbol grounding*, 33–53. Amsterdam: John Benjamins.
- Wanrooij, Karin, Paul Boersma & Titia van Zuijlen (2014a). Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study. *Frontiers in Psychology* 5: article 77.
- Wanrooij, Karin, Paul Boersma & Titia van Zuijlen (2014b). Distributional vowel training is less effective for adults than for infants. A study using the mismatch response. *PLoS ONE* 9 (10): e109806.
- Ward, Ida (1933). *The phonetic and tonal structure of Efik*. Cambridge: W. Heffer and Sons.
- Wedel, Andrew (2006). Exemplar models, evolution and language change. *The Linguistic Review* 23: 247–274.
- Wedel, Andrew, Abby Kaplan & Scott Jackson (2013). High functional load inhibits phonological contrast loss: a corpus study. *Cognition* 128: 179–186.
- Welmers, William (1973). *African language structures*. Berkeley, CA: University of California Press.
- Werker, Janet, John Gilbert, Keith Humphrey & Richard Tees (1981). Developmental aspects of cross-language speech perception. *Child Development* 52: 349–355.
- Werker, Janet & Richard Tees (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7: 49–63.
- Whelan, Robert (2008). Effective analysis of reaction time data. *The Psychological Record* 58: 475–482.
- Willoughby, Louisa, Stephanie Linder, Kirsten Ellis & Julie Fisher (2015). Errors and feedback in the beginner Auslan classroom. *Sign Language Studies* 15: 322–347.
- Wright, Anthony, Jacqueline Rivera, Steward Hulse, Melissa Shyan & Julie Neiworth (2000). Music perception and octave generalization in rhesus monkeys. *Journal of Experimental Psychology* 129: 291–307.

Zipf, George (1935). *The psychobiology of language: an introduction to dynamic philology*. Boston, MA: Houghton-Mifflin.

Zipf, George (1949). *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Summary of “The complexity and learnability of phonological patterns: simulations, experiments, typology”

Chapter 1. The research in this dissertation focuses on the roles that complexity and learnability play in the typology of sound systems. By “sound systems”, I mean phoneme inventories, that is, sets of the smallest units that distinguish between different meaning in spoken languages. Such systems are analysed in terms of phonological features: phonemes are usually seen as bundles of such features. In Chapter 1, the introduction, I explain some basic concepts that recur throughout this dissertation.

Chapters 2 and 3: computer simulations. Having set the scene in the introductory chapter, I explore the question how learners acquire phonological features in Chapters 2 and 3. I report on computer simulations with neural network models that are trained on probability distributions of auditory cues and lexical information. At the beginning of the learning process, the computer model behaves randomly, but after exposure to these distributions, it comes to display discrete activation patterns in the phonological layer: the emergence of these patterns means that phonological categories, that is, features, have emerged. Because activation flows through the entire network, these categories incorporate both bottom-up (i.e. auditory) and top-down (i.e. lexical) information. This process of feature emergence is robust: it happens even when we try to disrupt the model in several ways, for instance by inhibiting the nodes at the lexical layer, by making the auditory cues much noisier, or by delaying the development of the lexicon. While the networks in Chapter 2 learn categories by listening only, I extend the networks in Chapter 3 with an articulatory layer, enabling them to speak as well. This layer inhibits the production of peripheral auditory cues more than the production of central auditory cues, because the former are articulatorily more effortful than the latter. I use such bidirectional networks to set up chains of learners that learn from each other: they acquire a perception grammar first and then use the same knowledge to speak, and their speech is used as the input to a new learner. In such chains, learners end up with a language in which the auditory correlates of phonological categories are still distinct, but this contrast does not come at an excessive articulatory cost.

Chapters 4 and 5: experiments. Having made feasible that learners can induce features from auditory and lexical input, in Chapters 4 and 5 I report on experiments investigating the relation between complexity and learnability, in which human adults learn patterns of feature combinations. In Chapter 4, I explore inventories that were defined using one feature with at most two contrasts, and a second feature

with at most three contrasts; in Chapter 5, I explore inventories that were defined using two features with three contrasts. In both chapters, I use handshapes instead of spoken language: because the participants had already acquired a phonology, and I want to investigate the emergence of a new system of contrasts, I switch to a different modality. The inventories differ in terms of their complexity, which I use as a predictor of learning success. I investigate two measures of complexity: feature economy, defined as the proportion of possible feature combinations that the inventory employs, and logical complexity, the minimal length of the description of the inventory. In most tasks, both predictors yield significant effects; however, logical complexity tends to be a better predictor of learnability than feature economy. In Chapter 4, most errors result in regular inventories, that is, they fill up the gaps in the input; in Chapter 5, with the larger parameter space, learners make significantly fewer regularising errors than the ones acquiring the smaller parameter space in Chapter 4. In both experiments, however, learners significantly reduce the complexity of their input, usually by adding unseen categories rather than by leaving out seen ones.

Chapters 6 and 7: typology. In Chapter 6 I use the results from Chapter 4 to make predictions about the role of complexity in sound change. To do so, I set up Markov chains, treating the cohorts of learners from Chapter 4 as a single generation and assuming that every subsequent generation showed the exact same behaviour. These chains reach a stationary distribution in which only inventories with low complexity survive, because the learners in Chapter 4 reduced the complexity of their input. However, these results only paint a partial picture, because in natural spoken language, perceptual and auditory considerations are involved too. I explored the interaction between cognitive and phonetic factors by establishing complexity indices of the various stages of four attested sound changes, and found that sound change is not necessarily complexity-reducing; rather, it proceeds in a way that ensures sufficient contrast (remember the results from Chapter 3).

The observation that sound change is not necessarily complexity-reducing entails that sound systems do not necessarily have low complexity, and indeed this is what we find in Chapter 7: only just over half of the 317 plosive inventories in the UPSID database are regular. These results suggest that, while a bias towards low complexity exists in learning, this bias is frequently overruled by articulatory and perceptual factors.

Chapter 8: discussion and implications. In the last chapter, I further discuss some of the findings, and their implications for further research. Returning to the two complexity measures mentioned above, for instance, both the experimental and typological results suggest that logical complexity is a better predictor than feature economy, but more work is needed to verify this claim.

Samenvatting van “Complexiteit en leerbaarheid van fonologische patronen: simulaties, experimenten, typologie”

Je kan taalwetenschappers op veel verschillende manieren onderverdelen. Er is bijvoorbeeld een groep taalwetenschappers die wil begrijpen hoe taal wordt opgeslagen en verwerkt in de hersenen: dat zijn psycholinguïsten. Er is ook een groep taalwetenschappers die onderzoekt wat voor eigenschappen er vaak – of juist niet vaak – voorkomen in de talen van de wereld (zover die beschreven zijn), en waarom: die noemen we typologen. Er zijn taalwetenschappers die geïnteresseerd zijn in hoe klanken of gebaren worden gearticuleerd en waargenomen, en hoe ze bijvoorbeeld betekenissen van elkaar onderscheiden: zulke mensen heten fonologen en/of fonetici. En zo zijn er nog veel meer groepen taalwetenschappers — degenen die kijken hoe taal als communicatiemiddel wordt gebruikt, degenen die computermodellen bouwen, degenen die onderzoeken hoe kinderen taal verwerven, enzovoort. Alle taalwetenschappers behoren tot meer dan één groep: de onderverdelingen zijn gedefinieerd in meerdere, onafhankelijke dimensies. Voor het onderzoek dat ik in dit proefschrift beschrijf, heb ik bijvoorbeeld computersimulaties gedaan, experimenten uitgevoerd, en naar een database gekeken met informatie over echte talen. Heel algemeen geformuleerd heb ik de vraag willen beantwoorden: *komen dingen die mensen makkelijker kunnen leren ook vaker voor in talen?*

Het woord “dingen” is natuurlijk niet heel specifiek. In hoofdstuk 1 van dit proefschrift leg ik de achtergrond van mijn promotieonderzoek nader uit. Voor dit onderzoek heb ik naar foneeminventarissen gekeken, de verzamelingen klanken die gesproken talen inzetten om verschillende betekenissen van elkaar te onderscheiden. In het Nederlands verschillen de woorden *dal* en *bal* bijvoorbeeld maar in één klank van elkaar, maar ze verwijzen naar heel andere dingen. Deze observatie vertelt ons iets over het Nederlands: [d] en [b] zijn FONEMEN van het Nederlands.^a We kunnen fonemen, net als taalwetenschappers, ook in verschillende groepen indelen. Sommige Nederlandse fonemen, zoals {b f m w}, worden uitgesproken met één of beide lippen: zulke klanken heten labialen of lipklanken. Sommige Nederlandse fonemen, zoals {b d z l}, spreek je uit met trillende stembanden: zulke klanken noemen we stemhebbend. Deze eigenschappen heten FONOLOGISCHE KENMERKEN, en de namen van die kenmerken zijn doorgaans gebaseerd op de manier waarop we ze uitspreken of waarnemen. De [b] komt voor in het rijtje labialen, maar ook in het rijtje stemhebbende klanken: een foneem heeft meerdere kenmerken tegelijk. Maar hoe leren taalgebruikers die kenmerken? Er bestaat een redelijk populaire theorie, die

^a Fonemen noteer ik in [sluistekens], groepen klanken in {accolades}, en realisaties van klanken in [rechte haken].

van de Universele Grammatica, die stelt dat we worden geboren met kennis over die kenmerken: die kennis zou in ons DNA zitten. Die theorie is gebaseerd op de observaties dat kinderen schijnbaar makkelijk taal verwerven, en dat de talen van de wereld veel op elkaar lijken. Als deze theorie klopt zou het echter wel betekenen dat die kenmerken zich in alle talen precies hetzelfde moeten gedragen, en we weten dat dat niet zo is; sterker nog, hoe meer talen er door typologen beschreven worden, hoe meer onverwachte dingen we vinden, dingen die als het ware de poten onder de stoel van de universaliteit wegzagen.

In hoofdstuk 2 van dit proefschrift gebruik ik computersimulaties om te kijken hoe fonologische kenmerken ontstaan in een NEURAAL NETWERK, een model van een taalleerder dat gebaseerd is op hoe de hersenen werken. Zo'n netwerk bestaat uit zogeheten knopen, die met elkaar verbonden zijn: die knopen kunnen geactiveerd worden, en door de verbindingen andere knopen ook actief maken ("activeren") of juist minder actief maken ("inhiberen"), net zoals synapsen in onze hersenen elkaar activeren of inhiberen. Zo kan activiteit door het model stromen. In het model dat ik gebruik zijn de knopen ingedeeld in drie lagen. Twee daarvan stellen in zekere zin dingen buiten de leerder voor: een van de lagen is een representatie van verschillende betekenissen (zoals objecten die de leerder kan waarnemen), en een andere is een representatie van geluiden die de leerder kan horen, zoals frequenties (met oneindig veel mogelijke waarden). Daartussenin zit nog een derde laag, die een representatie in de hersenen van de leerder voorstelt. Door het computermodel combinaties van objecten en geluiden aan te bieden, en de activaties te laten stromen, ontstaat er gaandeweg op die tussenliggende laag een abstract representatieniveau: de oneindig vele mogelijke geluiden worden in de tussenlaag teruggebracht tot een beperkt aantal activatiepatronen, en dat zijn de kenmerken waar ik het hierboven over had. Die kenmerken kan het model dus leren uit zijn omgeving; we hoeven helemaal niet aan te nemen dat het ze meteen al kent, zoals de Universele Grammatica aanneemt.

Het netwerk leert door alleen te luisteren, maar in hoofdstuk 3 laat ik het ook spreken: daarvoor is nog een vierde laag nodig in het netwerk, een articulatorische laag. Ik laat netwerken ook van elkaar leren, een beetje zoals in een doorfluister-spelletje — een netwerk leert eerst van zijn omgeving en spreekt daarna, waarna een nieuw netwerk van de output van het vorige leert. Zo ontstaan ketens van opeenvolgende "generaties" netwerken die van elkaar leren, zoals echte talen ook van generatie op generatie worden overgedragen. In dit proces van overdracht spelen twee effecten een belangrijke rol: het PROTOTYPE-EFFECT en het ARTICULATORISCH EFFECT. Het prototype-effect houdt in dat je als luisteraar sommige realisaties van fonemen prototypischer, eenvoudig gezegd misschien "beter", vindt dan andere: een [s] die goed sist, met een hoge frequentie, is een duidelijke realisatie van dat foneem, die je eigenlijk onmogelijk kunt verwarren met een andere klank. Zulke prototypische realisaties van klanken kosten echter meestal ook meer moeite om uit te

spreken, en als spreker wil je niet meer moeite doen dan strikt noodzakelijk: de twee effecten zijn dus tegengesteld aan elkaar. In het netwerk is dat geïmplementeerd door geluiden die meer articulatorische moeite kosten sterker te inhiberen dan geluiden die minder moeite kosten. De simulaties in Hoofdstuk 3 laten zien dat de twee effecten gedurende de overdracht tussen meerdere generaties een optimale balans vinden: als je bijvoorbeeld begint met een taal waar sprekers heel veel moeite moeten doen om betekenissen uit elkaar te houden, dan verandert zo'n taal totdat het onderscheid zo klein is als nodig, en dat dat eenmaal is gebeurd, verandert hij niet noemenswaardig meer. Dat talen in sommige opzichten vaak op elkaar lijken kan dus komen doordat ze zo evolueren, zonder dat daar iets doelmatigs voor hoeft te gebeuren, en zonder dat er iets aangeboren hoeft te zijn. Het idee dat mensen – en eigenlijk niet alleen mensen – zo min mogelijk moeite doen om hun doel te bereiken zie je trouwens overal om je heen: denk bijvoorbeeld maar aan de “olifantenpaadjes” die langzaam uitslijten als mensen ergens een stukje willen afsteken.

Terug naar de vraag die centraal staat in dit proefschrift. Om die te beantwoorden moest ik uitzoeken welke dingen mensen makkelijk kunnen leren, en dat heb ik gedaan in hoofdstukken 4 en 5. Ik schreef eerder dat foneeminventarissen eigenlijk altijd worden geanalyseerd aan de hand van de kenmerken die relevant zijn. Als voorbeeld zie je hieronder een tabel met een inventaris van plosieven, klanken die je maakt door heel kort een volledige afsluiting in het spraakkanaal te maken en die vervolgens weer los te laten, zoals [p], of [t], of [d]. Om deze inventaris te beschrijven hebben we twee kenmerken nodig: stemhebbendheid (trillen de stembanden of niet?) en plaats van articulatie (waar wordt de afsluiting in het spraakkanaal gemaakt?). De stembanden kunnen niet of wel trillen (respectievelijk “stemloos” en “stemhebbend”), en de plaats van articulatie kan “labiaal” zijn (bij de lippen), “coronaal” (bij het harde verhemelte), of “dorsaal” (met de achterkant van de tong). Dat levert de zes mogelijkheden in Tabel 1 op (het foneem [g] is de eerste klank in het woord *goal*, niet die in *goed*). Het Engels heeft deze inventaris, bijvoorbeeld.

Tabel 1. *Plosiefinventaris 1.*

	<i>labiaal</i>	<i>coronaal</i>	<i>dorsaal</i>
<i>stemloos</i>	p	t	k
<i>stemhebbend</i>	b	d	g

Dit is een overzichtelijke inventaris: alle combinaties van kenmerken bestaan. Nu een andere mogelijke inventaris (“patroon”):

Tabel 2. *Plosiefinventaris 2.*

	<i>labiaal</i>	<i>coronaal</i>	<i>dorsaal</i>
<i>stemloos</i>	p	t	k
<i>stemhebbend</i>			

Dit is ook overzichtelijk: er zijn alleen maar stemloze plosieven, geen stemhebbende (daarom is dat woord grijs gemaakt). Maar neem nu een taal als het Efik, gesproken in Nigeria:

Tabel 3. *Plosiefinventaris 3.*

	<i>labiaal</i>	<i>coronaal</i>	<i>dorsaal</i>
<i>stemloos</i>		t	k
<i>stemhebbend</i>	b	d	

Dit is al een stuk ingewikkelder: deze taal heeft allebei de coronalen, de stemloze dorsaal, en de stemhebbende labiaal. De notie “ingewikkeld”, of zoals we liever zeggen: “complex”, kunnen we op verschillende manieren preciezer maken: bijvoorbeeld door te kijken hoeveel fonemen, ofwel combinaties van kenmerken, een taal gebruikt, en dat aantal te delen door het aantal combinaties dat *in principe* mogelijk is. Dat getal heet KENMERKZUINIGHEID. In deze laatste taal, bijvoorbeeld, worden alle waarden die de kenmerken kunnen aannemen gebruikt (er zijn stemloze klanken, stemhebbende, labiale, coronale, én dorsale): er zijn dus $3 \times 2 = 6$ mogelijke combinaties. Daarvan worden er maar 4 gebruikt, dus de kenmerkzuinigheid is $4/6 = 2/3$. Inventarissen 1 en 2 zijn maximaal zuinig, want ze gebruiken alle mogelijke combinaties, dus hun kenmerkzuinigheid is 1.

Een andere mogelijkheid om complexiteit te definiëren is om het aantal kenmerken te tellen dat je nodig hebt om een inventaris te beschrijven: voor inventaris 2 is dat er maar eentje (“stemloos”), voor inventaris 3 zijn het er wel vijf (“coronaal”, “stemloos en dorsaal”, “stemhebbend en labiaal”). Deze definitie heet LOGISCHE COMPLEXITEIT. In het algemeen kun je zeggen dat een taal zonder gaten in het systeem, dus een taal waarin geen mogelijke combinaties van kenmerken ontbreken, minimaal complex zijn; zulke inventarissen heten ook wel “regelmatig”. Inventarissen 1 en 2 zijn voorbeelden van zulke inventarissen. In het algemeen geldt dat hoe meer gaten er zijn, hoe hoger de complexiteit is, zoals in Tabel 3.

Uit eerder onderzoek weten we dat mensen meer moeite hebben om complexere systemen te leren: ze hebben bijvoorbeeld meer moeite om te beslissen of een combinatie van kenmerken in de inventaris zit, of ze hebben meer moeite om de regelmaat in de inventaris te beschrijven. Ik heb experimenten gedaan waarin mensen inventarissen van verschillende complexiteit moesten leren, en gekeken hoe goed ze dat deden. Aan mijn onderzoek deden alleen volwassenen mee, die in hun moedertaal of -talen al een heel systeem van kenmerken hadden geleerd; om te

zorgen dat ze niet op die kennis konden leunen moest ik dus iets anders dan spraak gebruiken, en op aanraden van collega's heb ik daarvoor een set heel simpele gebaren gebruikt, die dus andere kenmerken gebruikte dan stemhebbendheid en plaats van articulatie, maar wel op dezelfde manier kon worden gerepresenteerd (een foto van alle handvormen in de taal is te zien op pagina 106).

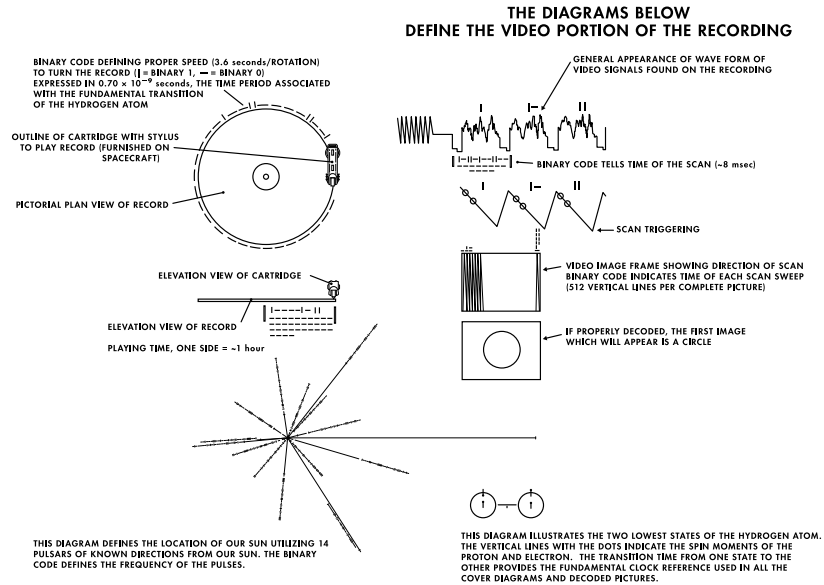
De resultaten van mijn experimenten laten zien dat complexiteit doorgaans inderdaad goed voorspelt hoe mensen het doen (logische complexiteit weliswaar nog iets beter dan kenmerkzuinigheid): hoe hoger de complexiteit, hoe slechter het resultaat. En als mensen de complexiteit blijkbaar te hoog vinden, doen ze er iets aan, door de inventaris dusdanig aan te passen dat hij minder complex wordt: meestal voegen ze een of meerdere fonemen toe die niet in hun input zaten, en vaak resulteert dat in een regelmatig systeem. Een leerder van plosiefinventaris 3, bijvoorbeeld, zal dus geneigd zijn om een [p] toe te voegen, of een [g], of allebei; en hij of zij zal zich daar niet eens bewust van zijn.

Het is mooi dat mensen in een experiment hun input minder complex maken, maar bij echte taalverwerving komt natuurlijk nog veel meer kijken: dat duurt veel langer, is interactief, enzovoort — en, heel belangrijk: in foneeminventarissen zijn het prototype-effect en het articulatorisch effect ook relevant, waar ze dat niet echt waren in de experimenten. Van de drie paren stemloze en stemhebbende klanken in Tabel 1, bijvoorbeeld, lijken de [k] en de [g] in zekere zin meer op elkaar dan de andere twee paren, dus verwarren luisteraars ze sneller met elkaar en moeten sprekers relatief veel moeite doen om het verschil duidelijk te maken: dat zijn geen nuttige eigenschappen, en het is ongetwijfeld de reden dat het Nederlands geen [g] heeft (alleen in een paar leenwoorden, waaronder dus *goal*). Daarom heb ik ook gekeken naar klankveranderingen in echte talen (hoofdstuk 6), om te kijken of die het systeem minder complex maakten, en naar plosiefinventarissen van echte talen (hoofdstuk 7), om te kijken hoe complex die zijn. En zo overtuigend als mijn proefpersonen in het lab de complexiteit van hun input reduceerden, zo vaak vinden we uitzonderingen in natuurlijke talen: klankveranderingen vullen lang niet altijd gaten in het systeem op, en hoewel een krappe meerderheid van de 317 plosiefinventarissen waar ik naar heb gekeken regelmatig is, zijn er ook genoeg inventarissen vol gaten. Dat is niet per se wat we misschien zouden verwachten, maar het is wel heel interessant: het laat namelijk zien wat mensen blijkbaar toch kunnen leren, ondanks dat het niet ideaal is voor ze, en het laat ook zien hoe verschillende factoren (zoals leerbaarheid, perceptie en articulatie) in elkaar grijpen.

Curriculum vitae

Klaas Seinhorst belandde na omzwervingen via de planologie en de Latijnse taal en cultuur bij de opleiding Taalwetenschap aan de Universiteit van Amsterdam, en is daar vooralsnog niet meer vertrokken. Na het cum laude behalen van zijn bachelor (met een minor Poolse taalverwerving) en onderzoeksmaster begon hij aan een promotietraject, waarvan dit proefschrift het resultaat is. Tijdens dat promotietraject presenteerde hij zijn werk in 15 steden in tien landen op drie werelddelen, en onderwees hij vijf cursussen en tien gastcolleges.

EXPLANATION OF RECORDING COVER DIAGRAM



Explanation of the symbols on the Golden Record cover. https://en.wikipedia.org/wiki/File:Voyager_Golden_Record_Cover_Explanation.svg. Source: National Aeronautics and Space Administration (NASA). Public domain.