universität wien

MASTERARBEIT / MASTER'S THESIS

„Excavating Fossils of Proto-Prosodic Processing“

verfasst von / submitted by

Duro Rosic, BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 11.05.2018

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 812
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	English Language and Linguistics
Betreut von / Supervisor:	Univ.-Prof.Mag.Dr. Nikolaus Ritt

Table of contents

Acknowledgements

I thank my mother and father, Ana & Vladimir, for being the best parents anyone could wish for and for always supporting me with their love and wisdom.

I thank my little brother Filip, for inspiring me by growing into one of the coolest persons I know and for giving me so much strength by always looking up to me.

I thank Mag. Hannes Pirker from ÖAW's ACDH, without whose expertise and knowledge of voice-synthesizers and programming I could not have set up my experiment design.

I thank the library staff from the English department, for having allowed me to turn their TV-room into a phonetic recording lab, and all the participants who dared to enter it.

I thank all my friends and colleagues who listened to me rambling on about my project, who gave me feedback on my thesis, who shared happiness with me when things worked out and who lifted me up when things went wrong.

List of Abbreviations

CI	Conceptual-Intentional System
DTE	Designated Terminal Element
FI	Feet Interval
FLB	Broad Faculty of Language
FLN	Narrow Faculty of Language
FOXP2	Forkhead Box P2
GT	Genotype
IH	Integration Hypothesis
II	Imitation Inaccuracy
IL	Iterated Learning
IOI	Inter-Onset Interval
ITL	Iambic-Trochaic Law
KDE	Kernel Density Estimation
LAD	Language Acquisition Device
LED	Levenshtein Edit Distance
LF	Logical Form
ML	Modern Language
NDS	Neo-Darwinian Synthesis
PC	Prominence Center
PCI	Prominence Center Interval
PF	Phonological Form
PL	Protolanguage
PLD	Primary Linguistic Data
POS	Poverty of the Stimulus
PT	Phenotype
SM	Sensory-Motor System
UG	Universal Grammar

Introduction

What are the evolutionary origins of language and how did the species Homo sapiens become so specialized for speech? Frankly, to answer these questions would by far outstrip the limits of a master thesis. Instead, this thesis focuses only on a fragment of this linguistic specialization that is dedicated to the perception and decoding of auditory input: the cognitive sensitivity towards rhythmic regularities in prosodic processing.

Linguistics struggles with the fact that language does not fossilize and that even the earliest linguistic records do not tell us enough about the evolutionary origins of language. Furthermore, theories and definitions of language itself remain to some extent highly controversial, albeit much was learned about the linguistic system in the last years. This only exacerbates the problem since it is easier to explain language evolution if there exists a consensus on how to understand and define language. The desire for consensus in linguistic theory gains even more weight given the fact that language evolution sciences are becoming increasingly interdisciplinary. It seems plausible that the ontological status of language needs to fit into natural science frameworks such as biology to be relatable to the biological concept of evolution. Lastly, although the science of language evolution has become more active, it currently needs more empirically testable hypotheses in order to not stagnate in speculation.

The main claim of the thesis is that there exist cognitive processing biases in the prosodic domain of human phonology, and we suggest that these biases could represent remnants of a linguistic system older than modern human language. The existence of these processing biases is exhibited in a phonetic experiment. In this experiment we show that participants are sensitive to structural regularities inside a phonetic stimulus they hear during the experimental task. Participants preferred stimuli that were rhythmically organized to those which were not. Preference in this sense means that the former structure type was more stable and easier to learn during transmission than the latter. Ultimately, we propose that this preference must have been present in the protolanguage of our ancestors.

To make this line of argumentation work, we will do the following. We start by presenting a literature survey for our methodology in the first three sections. The general idea is to first define what is meant by language in section one, then to do the same for evolution in section two, and then define what is meant by language evolution in section three. After the methodological foundations are laid out, we turn to hypotheses about language evolution in section four and methods for empirical testing of structure emergence in section five. More specifically, in section four we present the concept of protolanguage as a hypothetical stage in language evolution and address the idea that cognitive remnants of protolanguage can be found in modern language. The idea we want to propose is that such cognitive remnants can also be found via experimental studies such as presented in section five. However, none of these experiments were designed with specifically this in mind. This is why in section six we replicate and readjust such an experimental study in order to exhibit cognitive processing biases in the phonological domain of prosody. After presenting the experiment design and our results, we discuss the implications of our findings in the context of language evolution research in section seven. Our argumentation maintains that the prosodic processing biases for rhythmic regularities are indeed likely to have been present very early in the evolutionary history of language.

1. Language

Since this is a thesis on the evolution of language, it makes sense to define what language is before setting out to define what is meant by the evolution of language. Although our human capacity to speak and understand language seems to be as much part of us as our ability to breath or walk or use our hands for various tasks, to explain what language is ontologically will demand more than just a superficial description of collected words and utterances or a categorization of languages and dialects into various nationalities. After all, language is an extremely complex and multi-layered system where various functions interact with, are constrained by, and themselves constrain different

formal properties of linguistic structure, as linguistic research in the last 50 years has demonstrated.

From a survey of the relevant literature, we try to define what language is with regard to its functional and formal properties. This is necessary because we will need phonological terminology in order to analyze our data. Additionally, we need a systematic understanding of language in order to compare it to the communication systems of other species. In this context we also have to discuss to what extent language is a naturalistic or cultural entity, addressing the issues of Poverty of the Stimulus and Universal Grammar along the way. This first section aims to do three things: exhibit the multi-layered complexity of the linguistic system with a focused attention on phonology, demonstrate the great cognitive demand necessary to be able to speak, and determine the ontological status of human language and the linguistic capacity of the human species.

1.1. Function

Language is used everyday. Human interaction seems to be full of communicative behavior making use of language, both on spoken/auditory as well as written/visual channels. Clearly then, a function of language is to be a means of communication. To some extent this function is being efficiently executed, albeit communicative interaction can sometimes be troubled by misunderstandings. Nonetheless, speakers of a human language can talk about a seemingly endless plethora of things: about past and future events, about what is or is not present currently, and even about things which do not exist at all but only in our minds.

Additionally, language can be a carrier to express identity (Labov 1972). Examples include dialects, idiolects, youth-language, net-speak, and further linguistic manifestations that relate to various socio-cultural factors such as nationality, social hierarchy, and communities of practice. To some extent, the way we speak is an important indicator of social background and personality. From this follow functions for exclusion and inclusion, depending whether one is able to adopt the linguistic code of a speech community. Issues of social prestige

have always motivated language change, both synchronically and diachronically, where vernaculars, dialects and linguistic variables function as group-identity markers.

Lastly, language can function as a cognitive tool for categorization (Taylor 1989). This seemingly universal human urge to name and categorize things might well be an expression of our desire to make sense of reality. It appears plausible to assume that these processes are constrained by minimal computation principles of cognitive economy (Rosch 1978). In this sense, the formation of categories serves "to provide maximum information with the least cognitive effort" (Rosch 1978: 28). Instances of linguistic categorization can be found in all formal properties of language, such as the categorical feature distinction of phonemes in phonology (Taylor 1989: 247, 248).

1.2. Form

The formal properties and features of language are complex and multi-layered. When dealing with formal aspects of something as complex as language, it is recommendable to start with minimal assumptions. Formal units of language can be defined minimally as a mapping between signal and meaning. Under this definition, recent generative theories adopting methodological motivations from minimalism resurge basic linguistic ideas dating back to Saussure (1916). Latest linguistic theories influenced by the Minimalist Program (Chomsky 1992) distinguish between Logical Form (LF) and Phonological Form (PF), which are a mapping between meaning and signal respectively. Generally speaking, the majority of linguistic theories share agreement on this mapping relation.

The linguistic interface structure between meaning and signal can be further formalized into multiple layers. Modern language theories such as Functional Discourse Grammar distinguish between a pragmatic layer, a semantic layer, a morphosyntactic layer, and a phonological layer (Hengeveld & Mackenzie 2008). Other architectural formalizations of language structure such as Jackendoff's generative grammar notate for semantic structure, syntactic structure (including morphology) and phonological structure (2002: 6). Similarly, in his latest works

on evolutionary biolinguistics, Fitch repeatedly applied a minimal tripartite differentiation between semantics, syntax and phonology (2005: 194; 2010: 141). He concluded that to include these three in any formal structure of language should be uncontroversial (2010: 141). Differences arise when pragmatics is either excluded from semantics or linguistic structure altogether, and when theories disagree on the cognitive importance of mapping procedures taking place in morphology, syntax, and syntax-semantics interfaces. Some of these differences have also led to disputes. Among these is Jackendoff's critique of "syntacto-centric" generative grammars and his proposal for a parallel architecture of language which acknowledges generative potential in all linguistic structures and layers of language (2002: 107-111). But even among the layers themselves we can find further sub-layers.

1.2.1. Phonological Structure & Interface Systems

Since this thesis deals predominantly with phonology, we will focus in on the structure of this layer in particular and give examples accordingly. What is more, we intend to analyze the data in section six with the following terminology. Adopting Jackendoff's parallel architecture, phonological structure itself can be further organized into four tiers: prosodic, syllabic, segmental and morphophonological (2002: 6).

The segmental structure of phonology is prototypically seen as being organized by phonemes. Phonemes, which can be further decomposed into their distinctive features, are the structurally smallest meaningful building blocks of a language. Meaningful in this sense means that phonemes differentiate words from another, such as the vowels in *bit* and *bat*. When phonemes are not meaningful in this sense they are referred to as phones (Fitch 2010: 97). The distinctive features of a phoneme or phone convey the place and manner of articulation; they are phonetic realizations of articulation gestures.

Sequences of phonemes are organized into higher-order constituent groups such as syllables. These make up the syllabic structure of phonology. It is generally accepted that syllables are themselves hierarchically organized (Jackendoff

2002: 8; Ritt 2004: 141; Fitch 2010: 94). As such, the core of a syllable is known as the nucleus, which is the most sonorous unit, typically a vowel. Any following units, called the coda, build together with the nucleus the rhyme of the syllable. All units preceding the rhyme are termed as the onset.

The highest phonological tier is made up by the prosodic structure, which conveys a hierarchical structure as well (Fitch 2010: 99). This tier both groups syllables into intonational phrases and organizes them on a metrical grid, e.g. into feet. The former are marked by pauses in the speech signal while the latter "indicates relative stress of syllables" (Jackendoff 2002: 8).

Lastly, the morphophonological tier organizes the phonological speech signals into words. Interestingly, on this tier a unit such as *the* or *a* is not considered to be a word but rather a clitic, attaching to adjacent words to build a larger word constituent. Similarly, the copula *is* which sometimes realizes phonetically as [s] or [z] can also be considered a clitic.

Typically, these tiers interact with each other in speech. Jackendoff gives an example of interaction between distinctive features of phonemes in the segmental structure and pluralization in the morphophonological structure of English (2002: 7). Pluralization in English typically functions in the form of adding the sibilant /s/ to the right-boundary of a word, with the exception of irregular plurals. However, the simple addition of the plural suffix motivated by the morphophonology of the language can have different phonetic realizations. These differences are apparent in the distinctive features of the phonemes in phonology's segmental structure. For instance, Jackendoff presents the plurals *cats*, *dogs*, and *horses* (2002: 7). The plural suffix for the first is realized as [s], for the second as [z] and for the third as [əz]. The distinctive feature which differentiates [s] from [z] is that the former is regarded as voiced whereas the second is regarded as unvoiced. This means that for the production of [z] the vocal folds vibrate to create pitch, whereas for [s] they do not. We assume that this is so because the distinctive feature of the preceding consonant, voiced for [g] or unvoiced for [t], is carried over to the following unit in order to make

articulation easier. Contrastingly, for words already ending with a sibilant themselves such as *horse* an unstressed vowel is inserted before the plural morpheme that then acquires the voiced feature of the preceding unstressed vowel. Note that this is also the case for words that end with affricatives, such as *ditch – ditches* and *bridge – bridges*. The crucial observation here is that the phonetic environment in segmental structure determines how morphemes conveying a notion of 'plural' manifest in speech.

Similarly to the tiers in phonology, all broader systems of language such as semantics and syntax show various forms of interaction with each other (Jackendoff 2002: 13). According to Jackendoff "language comprises a number of independent combinatorial systems, which are aligned with each other by means of a collection of interface systems" (2002: 111). Originally, these ideas of independent tiers and layers stem from work in autosegmental phonology by Goldsmith (1979) and Liberman & Prince (1977). The crucial insight was that the metrical tier and the syllabic tier seem to be two independent systems; or put differently: "the metrical grid is not derived from syllabic structure" (Jackendoff 2002: 112). Apparently, both structures have units with "different formation rules"; therefore they must be aligned by an interface (2002: 112). Consequently, it is assumed that this interface between these two tiers operates on stress pattern constraints in a language. As such it negotiates "how a segmental structure is fit into an acceptable metrical grid" (2002: 113). The general idea was that to differentiate linguistic structure into multiple subcomponents was functionalized into a methodological principle. However, to apply notions of "interface constraints" is to account for the fact that "none of these subcomponents of language are strictly encapsulated" (Fitch 2010: 96).

1.3. Nature & Nurture

These formal properties of language can be regarded as forms of "cognitive structure" which are related to "functional activities of the brain" (Jackendoff 2002: 20, 21). This means that aspects of language are somehow instantiated in cerebral behavior. This concurs with Chomsky's assumption that "[t]hings mental, indeed minds, are emergent properties of brains" (1980: 70). Thinking of

language in this way allows us to position language inside a more naturalistic framework since brains are naturalistic entities. Additionally, such rich cognitive and linguistic structure of this kind, in addition to the unlimited meaningful content it can convey, is not found in the communication systems of other species (Jackendoff 2002: 39). This is echoed by Fitch with regard to the acquisition of language, which "requires innate mechanisms present in our species and not in others" (2010: 31). However, he simultaneously adds that "language requires a huge amount of environmental input"(2010: 31).

These comments hint at larger questions whether competence of language is acquired via experience or rather genetically predetermined (Kousta 2009: 96). Competence, in this regard, refers to a speaker's intuitive knowledge of the formal properties of their native language, such as phonotactic constraints on pseudo-words and nonce words, which we will exemplify below (Jackendoff 2002: 29; Fitch 2010: 96). It is important to differentiate competence from performance since performance refers to other cognitive factors during linguistic production and processing, e.g. memory limitations (Chomsky 1965: 4; Jackendoff 2002: 30). Theories on competence "are concerned with what the total structure is for either speaker or hearer", while theories on performance "are concerned with how the structure is built in real time" (Jackendoff 2002: 31).

Naturally, it follows to ask where these capabilities originate from in cognition: are they genetically innate endowments or are they acquired through culture? Historically, nature/nurture debates that addressed this question go back to the philosophical discussions between rationalists and empiricists (Laurence & Margolis 2001: 219). Both schools of thought asked whether the human mind was to be conceived as either a blank tabula rasa shaped mainly by experience or an a priori innate faculty responsible for cognition.

In linguistics, the most influential claims about the existence of innateness were formulated by Chomsky (1965: 47-48; 1988: 597-598). The innate language system is deemed to be responsible for the human creative potential and

capacity to generate an infinite set of output albeit being exposed to a finite set of input (Chomsky 1988: 588-589). One example of this is the phonological system, capable of sequencing a limited set of sounds into endless combinations of these sounds. However, Fitch notes that although phonology represents hierarchical as well as combinatorial structure, it does not compare to the recursive generativity of syntax since phonology lacks self-embedding (e.g. phonemes embedded inside phonemes) (Fitch 2010: 100), albeit some cases presented in Schreuder et al. (2009) may be convincing enough to assume the opposite especially with regard to prosody. Jackendoff in addition concurs that it is necessary to have a cognitive capacity for language (2002: 34). He admits that everyday performance of language seems to "exist in a social context" (2002: 34), e.g. as a negotiation of meanings and cultural habits, rather than be innately predetermined. Yet it is important to see that competence of language permits performance of language. This implies that language users require the necessary proficiencies to, firstly, apply any functions of language, and secondly, cognitively deal with linguistic formal structures (Jackendoff 2002: 34).

Assumptions that aspects of language are innate are not only inspired by language's intricately complex structure. They are also inspired by the generative processes that create these structures through combination of linguistic units. The presence of combinatoriality in linguistic systems seems almost a fundamental necessity, given that human speakers could not memorize all of the unlimited number of possible utterances (Jackendoff 2002: 39). As we have mentioned before, language allows their users to create a virtually limitless array of messages combined from a limited set of linguistic units. According to Jackendoff, such generative processes function by combining a lexicon with a grammar, both constrained by a limited set of rules (2002: 51,52). In phonology and phonotactics, for example, native speakers of English judge the nonce word *blik* as an acceptable and *bnik* as an unacceptable English word, albeit never having encountered them prior to exposition (Jackendoff 2002: 52). This intuition follows from constraints in English phonotactics that rule out certain phoneme sequences such as a plosive followed by a nasal in onset position. One additional example is rhyme. Since it is unlikely that the brain would save all

rhyming units of a language, we assume that they have to be computed on the spot (Jackendoff 2002: 64). Thus native speakers of a language can intuitively think of fitting rhymes for a word they have never heard before: rhyming pairs need not be learned individually.

1.3.1. Poverty of the Stimulus

As we have seen, language is an incredibly complex phenomenon, both in its functions and forms as well as how linguistic structures must be mentally created and processed. 'To do language' seems to be a severely difficult task demanding serious cognitive efforts, yet to humans it comes so naturally. Additionally, the fact that human children acquire the linguistic system so seamlessly in their first four years only exacerbates the challenge to explain this feat. Some scholars have considered this seemingly innate human trait as "an instinct to learn" (Fitch 2010: 30). However, it is widely accepted that children require some sort of linguistic input during their first four years, also known as the critical acquisition phase (Jackendoff 2002: 96). As such, both "a rich biological capacity for language learning and an appropriate social and cognitive environment" are necessities for language (Fitch 2010: 75). The most famous line of argumentation regarding this issue is the so-called Poverty of the Stimulus (POS).

Hornstein et al. give a brief and concise formulation of the POS problem (2005: 2-5). Given that speakers of a language can generate an infinite number of sentences and word-strings, how is this generative capacity acquired during childhood where all primary linguistic data (PLD) the child experiences is finite? Note that this illustrates a logical challenge: how can the potential to produce an **infinite** set of acceptable expressions be acquired from **finite** PLD? From this follows that children are highly unlikely to be radical empiricist learners with regard to their generative potential, and current discussions are concerned with the question whether certain language competences are derived rather from domain-general or domain-specific cognitive processes (Fitch 2010: 87).

In the past and even today, the POS argument and proposals for innate mechanisms of cognitive development were and still remain highly controversial (Laurence & Margolis 2001: 218). Additionally, recent debates are troubled by misunderstandings between varying camps (Lidz 2016; Ibbotson & Tomasello 2016a, 2016b; Chomsky 2014; Fujita 2014; Evans & Levinson 2009). While some scholars like Jackendoff try to clear up misunderstandings and reconcile both domain-specific and domain-general solutions to the POS (Jackendoff 2002: 71-90), others like Evans turn to rather hostile reactions and ridicule (Evans 2016). Most scholars doing empirical studies like Han et al. (2016) admit the importance of domain-general mechanisms during language acquisition. However, their debated findings from their acquisition study on Korean verb-placement suggest that domain-specific knowledge of grammar is additionally necessary to explain why the structure of the acquired grammar is richer than the PLD to which the child is exposed to (Han et al. 2016; Piantadosi & Kidd; Lidz 2016: E2765). Further empirical studies apply computational modeling to reconstruct the acquisition process using Bayesian statistics that represent domain-general mechanisms of language acquisition (Pullum & Schollz 2002; Perfors et al. 2006), but the computer simulations of statistical learners seem to be capable of only acquiring linear instead of hierarchical grammar rules (Berwick et al. 2011: 1235-1239).

As it stands, the POS argument is after 50 years still causing further debate. For one, POS arguments show the implausibility that language is acquired by induction alone; hence it is assumed that learners "must be hard-wired with principles that are specific to language" (Goldberg 2003: 219). Major modern approaches "all agree that a human-specific innate endowment exists that is necessary for language acquisition" (Fitch 2010: 85, 86). No matter whether one follows a nativist, cognitivist, functionalist or general purpose template learning system, both nature as well as nurture related explanations with regard to POS argumentation hold their grounds in linguistics since

> [a] notion of "environment" which includes only stimuli external to the body provides a hopelessly depauperate view of the role of experience in develop- ment. Conversely, a notion of "innateness" which lumps all of these types of developmental experience together as "innate" lacks the precision

and specificity necessary for deeper understanding. (Fitch 2010: 28)

1.3.2. Universal Grammar

It becomes clear that POS argumentations do not give solutions but create puzzles. One partial solution proposed for the puzzle of human language acquisition is an innate language acquisition device (LAD) that supports the child during the critical phase. The interesting bit is that no matter what the origins of any human child, it will acquire a fully acceptable grammar of any language conveyed to it by sufficient exposition to the PLD specific to that language. Thus it is assumed that all natural languages must derive from a Universal Grammar (UG) that guides the LAD with universal principles during the acquisition of a specific language (Chomsky 1981, quoted in Larsen-Freeman & Cameron 2008: 117). Depending on the language-specific PLD, the LAD will "select among the parametric options made available to them by the UG principles" (Larsen-Freeman & Cameron 2008: 118).

This differentiation between innately universal principles and language-specific parameter settings is a crucial categorical distinction in modern generative theories. According to Pinker & Bloom, all languages use the same mental devices; variation follows only from how these are differently weighed across the linguistic competence layers and performance functions (1990: 716). The choice between different parameter settings during acquisition can thus be understood as a form of "learning by forgetting" (Chomsky 2001: 16). In this sense, for example, one part of the acquisition of a phonological system is to memorize certain phonetic distinctions that are important for specific linguistic units in a particular language while forgetting phonetic distinctions that are redundant. Lastly, note that UG is a theory of I-language, as opposed to E-language (Chomsky 1986). The former refers to language as an internal mental state whereas the latter refers to language as an externalization in the form of an individual's or a community's linguistic behavior "external to human minds" (Jackendoff 2002: 29 footnote 6).

Fitch calls to attention regarding UG as a concept and as a term due to the "diversity of interpretations" it receives (2010: 88). Disagreements follow typically from questions to what degree UG is encapsulated and/or specific to the human species. In 2002, Hauser, Chomsky and Fitch published a paper to steer discussions into a more interdisciplinary and conciliatory direction. In this paper, Hauser et al. propose the term Faculty of Language as the innate linguistic system to be researched in biolinguistics. According to Fitch, "the goal of biolinguistics is to gain an understanding of the biological nature of the human capacity for language" (2010: 81). Hauser et al. differentiate between Faculty of Language in the broad sense (FLB) and Faculty of Language in the narrow sense (FLN) (2002: 1570, 1571). FLN is seen as a part of FLB, however with important hypothetical differences. It is hypothesized that FLN is specific to the human species whereas parts of FLB are shared with other vertebrates (2002: 1573).

The core nature of FLN is seen as a "computational system" responsible for the generation of infinite expressions from "a finite set of elements" via recursive operations (2002: 1571). FLB, on the other hand, consists of FLN, a conceptual-intentional system (CI) and a sensory-motor system (SM). Internal representations generated by FLN are mapped as PFs to SM by the phonological system, and as LFs to CI by the semantic system. In this sense, each linguistic expression is "a pairing of sound and meaning" (2002: 1571). Other "organism internal" cognitive processes such as memory are not part of the FLB since they are "not sufficient for language" (2002: 1571). Note, however, that Fitch takes a less radical position in later publications where he includes other modes of cognition such as memory to FLB and refrains from attributing any precise features to FLN (2010: 22, 23). Whatever one's interpretation, Rizzi is right in pointing out that if we want to place "[t]he study of the human language faculty in a biological setting [it] can't avoid coming to terms with the evolutionary aspects" (2016: 11).

2. Evolution

It should become clear that humans are incredibly specialized with regard to language, and we want to be able to explain how this specialization evolved in

the history of our species. Therefore we briefly summarize crucial aspects of recent evolution theory here in the second section. Given that research in the evolution of language is becoming increasingly interdisciplinary, it is recommendable for any evolutionary linguist to have a sufficient grasp of the foundational working methods in biological evolution.

The concept of replication from genetics will have important explanatory value during the discussion of our results. Furthermore, the concept of natural selection will help us understand how and why specific traits are selected for in a species. If we consider the specialization for language to be a trait of the human species, then we want to know how and why this trait was selected for. Additionally, the comparative method will be presented to show how we can compare different species' analogous traits, e.g. their communication systems. Although we can draw only indirect conclusions from this, it is nonetheless an important working method since direct evidence for the linguistic capacities of Homo sapiens' ancestors is incredibly scarce.

Most popularly, the notion of evolution is related to Charles Darwin. His work *The Origin of Species* (1859) is frequently counted among the most influential books ever written in the history of humanity (https://www.goodreads.com/list/show/36714.The_100_Most_Influential_Books_Ever_Written, 1. March 2018). However, the theory we will present here is actually a modernized version of old Darwinian evolutionary theory known as "Neo-Darwinian Synthesis" (NDS) which combines Darwin's insights on natural selection with modern genetics (Ritt 2004: 62; Fitch 2010: 39).

2.1. Neo-Darwinian Synthesis

The first crucial concept to understand in NDS is that of genotype (GT) and of phenotype (PT). Most fundamentally, GTs turn into PTs, where a GT is a collective of genes and a PT represents any kind of organism including actual human bodies. Therefore, we say that a PT is an expression of a GT (Ritt 2004: 63). More specifically, GTs can be seen as DNA-based replicating molecules containing genes that "trigger and direct" the construction of cells by proteins

which in turn are responsible for the construction of organisms (Ritt 2004: 63, 79). The architecture of these DNA molecules is double-stranded, similarly to a ladder or "train tracks" (Fitch 2010: 49). The replication process is the behavior of these DNA molecules, which occurs "every time a cell divides": they split up into two separate strands where each rebuilds into a new ladder "identical to the original" (2010: 50). As such, we may regard GTs as replicators that are always in competition to replicate or disintegrate (Ritt 2004: 80).

On the one hand, replication is never perfect and therefore yielding to diversification due to always changing environmental constraints (Ritt 2004: 80). The evolutionary stability of replicators, on the other hand, explains why life forms are as they are. There are a number of phenotypic traits that promote replication success, among these the genetic means to join "replicator alliances" (Ritt 2004: 74). One example of successful replication where the GT features of a PT are passed on is the inheritance of genes via sexual reproduction. Most importantly, the NDS explains in genetic terms why there are organisms, species, and extended phenotypic behavior such as web spinning, nest building and birdsong: because the very existence of these phenomena promote the replication success of GTs (Ritt 2004: 74, 78).

2.2. Natural Selection

Fitch makes clear that Darwin's primary contribution to evolutionary theory was that of natural selection (2010: 39). As such, the majority of Darwin's main concerns included variation and inheritance (2010: 37), which are mostly explained today genetically via NDS. However, there are still other biological factors involved in evolutionary theory. More specifically, Fitch notes three further categories of natural selection (Fitch 2010: 40 ff.).

Firstly, sexual selection refers to Darwin's idea that evolution cannot be explained only by survivability factors alone (Darwin 1871). Instead, evolution is additionally influenced by the reproductive behavior within a species. In this sense, traits that give a competitive advantage in the mating game are evolutionary preferred. Furthermore, sexual selection can explain fast evolution

particularly when "female choice and male traits are self-reinforcing" within a species (Fitch 2010: 40).

Secondly, kin selection addresses Darwin's failure to explain behavior influenced by altruism (Fitch 2010: 41). The inclusion of kin selection informed by genetics solved this puzzle of altruistic social behavior in species. In this light, helping ones relatives is, genetically speaking never an act of altruism. In the end, these acts would always ensure the replication of genes shared in the family and thereby save them from disintegration. As such, any sacrifice made for kin would never outweigh the evolutionary benefits gained.

Thirdly and lastly, group selection is a concept that attempts to address evolutionary selection on "biological levels above and below the individual" (Fitch 2010: 42). However, Fitch warns readers that the term receives different interpretations, and depending on the interpretation one adopts it may rather contribute to confusion than give clarification (2010: 43).

2.3. Comparative Method

Additionally to categories of natural selection and NDS, evolutionary biology makes use of the comparative method (Fitch 2010: 44 ff.). The comparative method tries to explain the development of adaptive traits of related species usually with regard to how they managed to survive their environment. The relation of species is conceptualized in the form of a branching tree where related groups form a clade that can further subgroup into further clades, e.g. primates being a subgroup of mammals which are in turn a subgroup of vertebrates (Fitch 2010: 45). If a trait is passed on from ancestors to their descendants, this is regarded as a homologous trait, such as fur in mammals or the bone structure of forelimbs in vertebrates.

The comparison of homologous traits among a clade allows us to make inferences about "the common ancestor of that clade" (Fitch 2010: 46). Contrastingly, if a trait evolves "independently in two separate lineages" this is regarded as an analogous trait (2010: 46). As such, birds, bats and butterflies all

share wings as an analogous trait, and the fact that these traits converged independently toward the same function, namely flight, is called convergent evolution (2010: 46).

3. Language Evolution

Having established some fundamental assumptions and conceptions about both language and evolution, we are ready to merge these two notions together. Ultimately and most basically, the question of language evolution is how a system as language might have evolved, and possibly also why it has taken the formal properties it has. One task is to show to what extent the explanatory tools for biological evolution are suitable to explain the evolution of language. Given the complex nature of language, a systematic approach to language evolution needs more than one historical layer of analysis in order to produce satisfactory explanations. As will become clear, what is meant by language evolution can mean different things from a phylogenetic, ontogenetic or glossogenetic perspective. We will define these terms in the following.

Originally, evolution research made use of four "historical levels of description", also known as Tinbergen's Four Whys (1963) that relate to questions about mechanism, function, phylogeny and ontogeny (Fitch 2010: 69). Firstly, the question about mechanism asks what makes a certain behavior possible. In the case of language, this includes but is not limited to a vocal tract, articulators and neural connections to the brain. Secondly, the question about function is concerned with the ultimate goal of a specific behavior. With regard to language, the answer could include communication, identity expression, and categorization. Thirdly, the question about phylogeny addresses the "evolutionary history of the species" (Fitch 2010: 70). Fourthly, the question about ontogeny urges to explain the development of an individual of a species.

Before we address these last two layers in more detail, note that all of Tinbergen's Four Whys relate traditionally to research in behavioral biology. Modern evolutionary linguistics focuses more on phylogeny and ontogeny with the addition of a cultural layer termed 'glossogeny' (Hurford, 1990) "to refer to

historical linguistic change" (Fitch 2010: 34). As such, evolutionary linguistics applies three timelines as hierarchical layers to explain the evolution of language: phylogenetic change taking the longest time to occur, ontogenetic change taking the shortest time to occur, and glossogeny representing "an intermediate level of change" (Fitch 2010: 34). Additionally, "[w]hile ontogeny refers to the individual and phylogeny to the species, glossogeny refers to a set of cultural entities close to what is called 'language' in the everyday sense" (Fitch 2010: 90).

3.1. Phylogeny

As a first crude phylogenetic statement we might propose that the evolution of language is closely tied to the evolution of the human species. This is not to say that research on animal communication is to be excluded from language evolution studies; quite the contrary. It follows exactly from applying the comparative methodology to different species' communication systems that language as we have defined it is ultimately human-specific (Jackendoff 2002: 233). This is so because comparison shows that the whole assembly of multi-layered and multi-functional structure of language is not observable in communication systems of other species. Rather, it currently seems that only some specific features of FLB are apparent in non-humans, such as the acquisition of a lexicon in primate communication and hierarchical structure in birdsong; however for the former it seems that such linguistic capacities only surface when trained to do so, and for the latter one could argue that this resembles recursion typical of FLN (Fitch 2010: 166, 183; Jackendoff 2002: 241).

From this follows to assume that not language as a whole is an evolutionary adaption, but rather that some parts are while others are not (Fitch 2010: 66). Note that there are indeed reasons to attribute some degree of isolation to specific human linguistic domains (e.g. mental dictionary as lexicon and procedural system as grammar) due to "larger principles of neural organization" as studies on aphasia show (Ullman et al. 1997: 267). With regard to phonology, isolating this linguistic layer can be supported by the existence of pseudowords. This is so because the very fact that the phonological system can generate

virtually more symbols than morphology, syntax and semantics need is a sound reason to regard phonology "as a stand-alone system" (Fitch 2010: 467). However, it should be mentioned that neither a factual claim nor an analytic truth is postulated with regard to the isolation of phonology. Rather, isolating phonology can be seen as a methodological principle that could prove to be very effective for addressing evolutionary questions about language in the human species.

One task of language evolution phylogeny is to therefore explain why only humans seem to have language. In 2016, for example, Fitch et al. showed that regarding phonology this could not be strictly due to our articulatory organs because "the primate vocal tract is not as restricted as previously thought" (2016: 4). Since "the basic primate vocal production apparatus is easily capable of producing five clearly distinguishable vowels" - typical of human languages - the researchers rightly assume that the human linguistic capacity rather stems from neural control than anatomy (2016: 4). This alludes to the notion that language appears to be a specifically **cognitive** specialization of humans.

Given the high cognitive demand of such a mental system, scholars frequently address the constant increase in brain size during hominid evolution "over the last 2 million years"; yet fossil records of symbolic thought date back to approximately 80,000 or 100,000 years (Bolhuis et al. 2014: 3, 4 Figure 2). Generally, the whole vertebrate clade evolved faster brains by developing "a fatty insulator material" named myelin that wraps around neural connections to "greatly accelerate signal conduction" in the brain without the neural connections having to increase in diameter (Fitch 2010: 62). Interestingly, for hominids however it seems that the compared to other species vastly outclassing brain growth was preceded by the development of bipedalism (Fitch 2010: 255). Bipedalism would have some further important implications for the creation and usage of tools since hands were not needed for walking anymore. Additionally, some scholars see important connections between the advent of tool making, fire usage and brain growth (Ko 2016) as well as between the structural similarities of tool making and the recursive features of FLN (Fujita 2014).

3.2. Ontogeny

Furthermore, it seems clear that language evolution must address phylogenetic as well as ontogenetic issues about the evolution of human behavior based on both cognitive constraints and cultural environment (Fitch 2010: 31). The richness and complexity of linguistic structure and the POS during acquisition show that humans need some innate support to perform and acquire this kind of cognition; namely FLB and FLN in addition to other domain-general aspects of the mind. It makes sense to ask about the evolution of these genetically endowed systems that support individuals during early ontogeny, especially in the light of NDS.

In genetic studies on specific language disorders the first language related gene to be discovered was the forkhead box P2 (FOXP2), which is apparently present in other vocal learning vertebrates as well but has undergone changes during hominid evolution (Nudel & Newbury 2013). However, Bickerton warns to not attribute too much importance to the FOXP2 since the gene seems to be mainly responsible for articulatory production and perception during vocal communication and learning, and is possibly only one language gene among more to be discovered (2007: 523, 524). What is more, cultural environment does not only hold a function as PLD for some sort of LAD, but might even influence genetic evolution as a form of "Baldwin Effect" (Fitch 2010: 70,71; Jackendoff 2002: 237 footnote 3). This means that if any form of cultural behavior was evolutionary promoted by natural selection, a species could evolve a genetic endowment that cognitively supports the acquisition and performance of that behavior.

Regarding additional ontogenetic issues, there is an important difference between human and non-human ontogenies; namely that human infants are longer dependent on their caretakers than infants of other species (Fitch 2010: 482). It seems plausible to assume that during this phase of high dependency there must be some sort of communication between infant and caretaker. Indeed, Thelen (1991) lists stages where the first signals by the child are

unintentional reflexes in the first two months, which then turn to intentional attempts at vocalization, leading ultimately to a babbling-stage before the usage of first words after its first year. Jackendoff notes that during babbling and the usage of first words the linguistic output of a child conforms to the basic opening and closing mouth movements "while tongue and lip position are held constant" (2002: 243). This linguistic output during the first year displays the child's preference for simple syllables with a "uniform place of articulation", before it continues to experiment with other places of articulation. From this it could follow that syllables are acquired earlier than phonemes as units of linguistic competence in human ontogeny.

3.3. Glossogeny

Current historical records of E-languages show that they do not strictly evolve but rather change through history, whereas "the ability of any child to learn any human language" must have evolved at some point but "has remained frozen for 10,000 years or more" (Bolhuis et al. 2014: 5). Some have suggested that this historical change occurs in cycles where "the process whereby a language alternates between reliance on word order and reliance on affixation typically takes a thousand years" (Pinker & Bloom 1990: 727 footnote 3). However, we can assume that these forms of historical development and cultural evolution must have some important relations to phylogeny and ontogeny, e.g. given the fact that "externalized language" necessarily serves as PLD for the LAD (Fitch 2010: 89). The term glossogeny was introduced to refer to processes and mechanisms of this kind of change, and to reconcile these aspects "with a goal of understanding I-language from a biological, generative perspective" (2010: 89). For instance, it was shown "that patterns of change depend strongly on the frequency with which words are used in discourse" (2010: 90).

Among the most crucial ideas in glossogeny is the attempt to extend the notion of replication and replicators in a Darwinian sense to more than just bodily or genetic entities; namely ideas (Fitch 2010: 89). As such, the mental replicator unit of analysis is called a meme, which was originally introduced by Richard Dawkins (1976). In this sense, a meme can be seen as "a neural structure with

behavioral expressions" that can be observed and recognized by others who are sensitive to the structural regularities of that behavioral expression (Ritt 2004: 219). The question follows whether some linguistic entities can be classified as memetic replicators as well, and how they behave in glossogeny. Thus, Ritt formulates a possible methodology for glossogenetic questions about language evolution and change (2004: 121). The first three questions to ask are the following:

(a) what linguistic competence constituents count as replicators

(b) which mechanics underlie their replication

(c) how is their replicatory success influenced by environmental factors

Note here that 'competence' refers to "some cognitive or mental system […] implemented in the form of brain-states" (Ritt 2004: 20), similarly to the conceptions addressed by Jackendoff, Chomsky and Fitch in section one.

What linguistic entities could count as memetic replicators in phonology? Example candidates for (a) are distinctive phonetic features, phonemes, or phoneme sequences (2004: 132-134). In order to qualify as memetic replicators, they must be entities that have (1) longevity, (2) high enough copying fidelity, and (3) minimal fecundity, and (4) they must actively seek out replication (Ritt 2004: 123, 124). According to Ritt, phonemes meet all four criteria, whereas their distinctive features can be excluded since they do not seem to rely on memetic replication to exist but their replicator status may be rather attestable in the genetic domain (2004: 134). Regarding phoneme sequences, these seem to meet all criteria as well, but this is the case for (1) only when there is a semantic component mapped to them in which case the relevant unit would be the morph or morpheme (Ritt 2004: 135). However, observing and recognizing a morpheme is typically accompanied by further phonotactic competence and being sensitive to its syllabic structure, in which case both competence properties replicate in conjunction (Ritt 2004: 141, 142).

Which mechanics underlie their replication? The answer to (b) is imitation of behavioral expressions (Ritt 2004: 196). First, consider a phonological replicator to be a configuration between two competence constituents of SM, one

responsible for processing "auditory impressions" and the other for executing "articulatory gestures" (Ritt 2004: 170). When someone perceives a behavioral expression, it is either recognized or not (Ritt 2004: 219). If it is recognized, this means that the competence constituents responsible for processing are activated in the form of neural brain states. This promotes stabilization since "the observer's mind already hosts a 'meme' for the observed behavior" (Ritt 2004: 219). If it is not recognized, the observer's mind compares the observed data's structural regularities to those competence constituents it currently contains (Ritt 2004: 220). This means that neural brain states are not only activated, but their configurations restructure to make "plausible hypotheses about the observed behavior", which are in turn tested via virtual or actual articulation and evaluated for their cognitive "costs and benefits" (Ritt 2004: 220). If the evaluation leads to positive results, the neural brain-state acquires a new structural configuration responsible for processing and (re)producing the observed behavior (2004: 220).

How is their replicatory success influenced by environmental factors? To answer (c) Ritt lists three possible types of selective pressures. Firstly, genetic pressures constrain and motivate replicating success based on but not limited to the fact that the SM represents "expressions of human genomes" (Ritt 2004: 222). Since the SM is responsible for production and perception of externalized language, on which memetic replication is based upon, the genetic make-up of this domain promotes replication of units that are easier expressed and perceived than others which are not (2004: 222). Secondly, the replication success of memes is influenced by their own properties and features in the form of memetic pressures. Memes must be adaptive and sensitive to another since they must form networks to be functional and because some memes are derived from other memes (2004: 223). Thirdly, social pressures constrain how successful replication is, given any particular populace (2004: 225). As such, social power relations, prestige, and other cultural values may influence what is worth imitating from whom (2004: 225). In the light of (a), (b) and (c) we thus may assume that glossogenetic change can resemble

> "[…] memetic evolution in terms of variation being brought about by imperfect meme replication and the subsequent selection of competing variants, as the Darwinian paradigms requires." (Ritt 2004: 169)

4. Proto-Language

As we have seen, language evolution can take place on different timescales and in various degrees of scope, be it the individual, a specific community or the whole species. Additionally, these layers show forms of interaction which can have a number of influences on linguistic evolution; the examples we exhibited are only a few. Nonetheless, our records that relate to language evolution are severely limited, especially with regard to phylogeny. Therefore, specific scenarios of phylogenetic language evolution are still firmly in the realm of approximations and speculations.

Current debate in the science of language evolution seems to be divided whether the evolution of language occurred in an abrupt or gradual fashion (Bickerton 2007: 520). Put briefly, the former position argues that language must have evolved abruptly via genetic mutations since it is currently not clear how natural selection would be sensitive to FLN (Hauser et al. 2002; Bolhuis et al. 2014; Jackendoff 2002: 234). Interestingly, there is evidence that properties of SM in FLB such as categorical perception must have already been present in mammals generally before the evolution of Homo sapiens (Bolhuis et al. 2014: 5; Fitch 2010: 99). Thus, it is assumed that what led to the seemingly abrupt evolutionary leap toward modern humans was the appearance of a novel computational system FLN and recursive merge operations. Of course, Fitch is right to point out that this proposal is "plausible, but extremely difficult to test or falsify" given that we lack any further genetic evidence (Fitch 2005: 186, 187).

Of course, there are attempts to explain a gradual language evolution scenario which assume that "the communicative advantages of human language are just the kind of cognitive phenomenon that natural selection is sensitive to" (Jackendoff 2002: 235). Here Jackendoff refers to a paper by Pinker and Bloom (1990) which was the first to argue for a gradual evolutionary emergence of UG under natural selection pressures that relate similarly to "other specialized

biological systems" (Pinker and Bloom 1990: 726). However, what counts as abrupt or gradual is relative to the duration one is ready to attribute to the evolution of language generally, and to the emergence of particular linguistic competences more specifically. It may be more useful to propose different necessarily consecutive stages of language evolution where both gradual and abrupt developments could have had occurred at different points in evolutionary history.

Crucial to this endeavor is the concept of protolanguage (PL). PLs function as hypotheses about possible developmental stages of the human linguistic systems in evolutionary history pre modern Homo sapiens (Fitch 2010: 399, 400). Additionally, these notions of PL are models for possible language evolution scenarios that address the problem of linguistic structure being too complex to have emerged all at once. The general trend here is to assume that PLs were simpler systems than modern language, lacking certain features such as semantics or syntax. We exhibit two PL models and demonstrate justifications for these in accord to our language evolution methodology from section three. Additionally, we present a crucial part of argumentation, namely that there exist cognitive fossils of PL and that it is possible to observe these under certain circumstances in modern language.

4.1. Lexical Protolanguage

The idea of a lexical PL was originally expressed by Bickerton (1990) and later revisited by Jackendoff (2002). Their model differentiates between PL and modern language (ML) so that ML contains hierarchical syntax structure whereas PL does not. According to these lexical PL models, ML can show instances of PL when disrupted or when in development (Jackendoff 2002: 235). For the former, instances of this can be found in language impairments and pidgin languages, and in ape language experiments and language acquisition for the latter. The important claim is that these cases represent cognitive remnants of a lexical PL that did not need syntax to be communicatively effective. Additionally, Jackendoff further claims that such linguistic fossils of PL are still present in modern grammars, e.g. *ouch* and *shh* (2002: 236, 240). Note that he

does not propose such items to be actually historic holdovers of PL, but what is rather a holdover is the potential of any language to contain these units (Jackendoff 2002: 240 footnote 6).

The working method of Jackendoff's lexical PL model is to decompose UG into multiple parts and enquire how they might have evolved "independently or in sequence" in order to enhance communication (Jackendoff 2002: 235, 250). When dealing with the question what came first and what built on what, Jackendoff stresses the importance of not asking whether species X has language or not, but rather what parts of language species X has. For example, he acknowledges that some preconditions for a lexical PL can be observed in primates (2002: 238). Such preconditions include conceptual structure and symbol use, which were used to map meaning to signals (2002: 239). Note here that there are preconditions that are not shared between humans and primates. For example, one crucial difference between primate and human communication is that for the latter it is not situation specific. Additionally, Jackendoff's model assumes an existence of single symbols before the emergence of combinatoriality. This would make sense if the phylogeny of language parallels the ontogeny of language acquisition where a one-word stage precedes usage of longer utterances (2002: 239). Note however that others have criticized such parallelisms between phylogeny and ontogeny (Pinker & Bloom 1990: 719).

Jackendoff proposes two major innovations that led from the single-word stage to an actual lexical PL. The first is for the PL to permit an unlimited large class of symbols in the form of a lexicon where meaningful linguistic units are stored in long-term memory. The second is to permit concatenations of symbols into larger utterances. These two innovations are independent of each other, but it is plausible that both developed concurrently (Jackendoff 2002: 241). According to Jackendoff, besides imitation and pointing behavior, the first innovation required foremost a form of proto-phonology for proto-humans to be able to adapt to learning and inventing a vast amount of symbols combined from smaller units (2002: 241, 242). Difficulties would not only arise due to remembering an increasing number of utterances but also due to the "perceptual problem […] of

making all the utterances discriminable and memorable" (Jackendoff 2002: 242). The second innovation required a process to concatenate symbols without the use of modern syntax where the relation between units is "dictated purely by context", e.g. *Fred apple* or *chicken eat* (Jackendoff 2002: 246). Concatenations of more than two units would additionally need the use of "linear position to signal semantic relations" (Jackendoff 2002: 246).

In summary, the fundamental initial architecture was basically a connection between noises and meaning. Then with the addition of phonology, the interface between phonology and meaning made it possible to have lexical items and linearity principles for semantic relations. The rest of Jackendoff's speculative model describes the further development from PL to ML. To present this only briefly, he starts with hierarchical phrase structure, and re-builds ML with the addition of relational vocabulary, grammatical categories (e.g. nouns and verbs), morphology and grammatical functions; all of which are posited as additional mapping devices between phonology and meaning (2002: 238, 252-261). We believe that the results from our experiment in section seven will be specifically useful for showing how the problems of proto-phonology regarding memorization and perception could have been partly solved via rhythmic factors in the prosodical domain of human language.

4.2. Prosodic Protolanguage

As we have seen, the phonological system appears to have some sort of primacy and importance in the evolution of language. This gives us all the more a reason to concentrate on the evolution of human phonology in particular. Furthermore, the merit of this focus is that human phonological systems are never randomly structured but always structurally constrained, and predictable patterns can be found both in ontogeny as well as in glossogeny (Fitch 2010: 94, 97, 98). For instance, Blevins notes that all human languages contain consonants and vowels in their phonological inventories, with the vowels /a/, /i/ and /u/ being a minimal set for the majority of languages (Blevins 2006: 118). A satisfactory theory of language evolution will have to account for any such structural regularities.

Another core feature of the phonological system is known as duality of patterning, which consists of generativity and arbitrariness (Fitch 2010: 94, 466, 467). Generativity refers to the fact that phonology allows a recombination of a limited set of sound signals into an open-ended set of sound symbols to which meaning can be attached. Arbitrariness refers to the fact that the sound signals and symbols are meaningless on their own, and that the attachment relation between sound and meaning is random; onomatopoeia and phonestemes being the possibly only exceptions. Thus, "[t]he core generative process in phonology is the formation of larger units" from smaller units such as phonemes (Fitch 2010: 468). This process generates hierarchical structures that are constrained depending on the particular language. For example, a very frequent phonological constraint is the permission of CV syllables (Fitch 2010: 468). Not only do all languages permit this syllable configuration; some permit no other configuration at all (Prince & Smolensky 2004: 105). Fitch notes that these features of hierarchical structure show some parallels to music, due to which a number of scholars in the past, among them Darwin (1871), have made proposals for a "musical protolanguage" (2005: 220; 2010: 468-470).

According to Fitch, Darwin's idea was that "the generative aspect of phonology might have emerged before it was put to any meaningful use" (Fitch 2010: 471). Darwin proposed in his model a development in three stages starting with an increase in cognition, followed by the emergence of the phonological system and vocal imitation, and ultimately arriving at ML via a mapping of combined and hierarchically structured signals and meanings (Fitch 2010: 472). Darwin stresses that these developments are rather attributable to the brain than the articulatory organs, and that language evolution is rather concerned with "the language **faculty**" than about any specific language (Fitch 2010: 473; his emphasis). The central argument behind such PL proposals is that song and spoken language share the design features of using a vocal and auditory channel "to generate complex, hierarchical structured signals that are learned and shared across generations" (Fitch 2010: 474).

More specifically, Fitch proposes the notion of a **prosodic** PL, reframing Darwin's proposal in accord with modern evidence (2005: 220; 2010: 474). The reason for this change in terminology is that the modifier *musical* connotes discrete tonal scales and beats, which are rather attributes of modern music than of a precursor communication system of the human species (2010: 475). Therefore, the modifier *prosodic* connoting "syllabic discreteness" is better suited to describe a PL stemming from the "shared mammalian capacity for categorical perception" in vocal and auditory domains (2005: 221; 2010: 475).

One strength of this model is that phonological hierarchy structure could have had functioned as preliminary blueprint for syntactic structure (2010: 475). This assumption is shared by Jackendoff as well, who alludes to Carstairs-McCarthy's (1999) proposition that the noun-verb organization preference for {N {V N}} to {N V N} is derived from the "accidental availability" of asymmetric syllable structure {C {V C}} already present in cognition (Jackendoff 2002: 253, 254 footnote 11). However, one of its weaknesses is that it cannot fully explain how propositional semantics could have had emerged, albeit some non-propositional but meaningful associations can be found in music (Fitch 2010: 476). In this sense, "prosodic protolanguage possessed phonology, and parts of syntax, but lacked lexical, propositional semantics" (Fitch 2010: 476). Note that this is different from Jackendoff's lexical PL model where propositional semantics is assumed to exist before the emergence of a phonological system.

It seems that recent neural data supports the connection between language and music as envisioned by some of Darwin's predictions (Fitch 2010: 477). For example, hierarchical and phonological processing seems to be shared among neural networks whereas "lexical access, phonetic detail, and propositional semantics" appear to be instantiated in separate networks (Fitch 2010: 477). Since music and language share hierarchical structures, this relation offers novel possibilities for empirical experimentation (2010: 478). Given that song and phonology share neural bases we might predict that individuals will co-vary in these aspects, while other processing systems such as syntax and propositional semantics will show more independent variation (2010: 478). For example, Fitch

reports on a study showing correlations between musical and phonological capabilities of Finnish children; again probably due to "shared neural resources" (Milanov et al. 2008; Fitch 2010: 478). Additionally, the argument that song predated speech, and that phonology as well as some features of syntax were possibly influenced by prosodic PL is supported by "the abundant comparative data on the evolution of vocal learning" (Fitch 2010: 481). Indeed, one might ask whether there is something special about human vocal learning in particular and human language learning in general, especially in a more experimental setting to test various models and hypotheses, so as to give research in and explanations of language evolution more of an empirical bite.

5. Experimental Studies

Now we turn to report on recent empirical language evolution studies that apply the concept of iterated learning to psychological experimentations known as diffusion chains. The reason for including these studies is that they represent empirical testing grounds for hypotheses about the emergence of linguistic structure. Here we assume that this structure emergence must be related to aspects of cognition. We would want to incorporate the results of these experiments as evidence for discussions on PL and fossils of PL. However, none of these experiments were designed specifically to exhibit aspects of cognition that could be possible candidates for fossils of PL. This is why we propose to improve the following studies in this regard or at least review them as blueprints for our own experiment.

According to Kirby et al. (2008: 10681) human language can be seen as an "evolutionary milestone" and as a system to transmit human culture. More specifically, it is a system that carries both the meaning of a message as well as information about its own linguistic structure. Kirby and colleagues see language as an "evolutionary system in its own right" and regard it as a testing ground for "theories of cultural evolution" via what is known about language processing and acquisition (2008: 10681). Crucial to this endeavor is the concept of iterated learning (IL). Among the most recent definitions of IL found in the literature is the one given by Kirby et al. (2014: 108):

> The process by which a behavior arises in one individual through induction on the basis of observations of behavior in another individual **who acquired that behavior in the same way**. [their emphasis]

Computational modeling of IL has shown that systems transmitted through repeated transmissions firstly increase in learnability, and secondly increase in structure (Kirby et al. 2008: 10681). However, Kirby and colleagues are aware of critique by Bickerton (2007: 522) who stresses that computer models fail to account for more realistic conditions. Rightly then, Kirby et al. (2008: 10681) state that "[w]hat is needed, therefore, is an experimental paradigm for studying the evolution of complex cultural adaptions using real human participants."

IL traditionally refers to the computer models whereas **diffusion chains** refer to actual studies where IL is researched experimentally. The typical procedure of such experiments is that the experimenter provides a "target behavior" which is observed by a participant (Kirby et al. 2008: 10682). The target behavior is then to be replicated by the participant. This replication subsequently functions as the target behavior for the next participant who likewise observes and replicates it. Repeating this process a number of times forms a diffusion chain, and each repetition is called a generation (2008: 10682). The development observed from generation to generation is known as "diffusion of behavior" (2008: 10682). Typically, the diffusion of behavior is measured and statistically analyzed to test whether predictions can be made with regard to learnability and structure that resemble the findings from computational modeling studies.

Applications of this experimental paradigm were first reported by Bartlett in 1932, but Kirby and colleagues consider the most important application of this methodology to be the recent work on "the cultural transmission of tool-use strategies in populations of chimpanzees and [approximately three to four year old] children" by Horner et al. (2006: 13882; in Kirby et al. 2008: 10682). However, Kirby et al. state that an experiment is needed where a more "direct comparison with human language" can be made (2008: 10682). The following is a summary of three experimental diffusion chain studies making use of iterated learning: Kirby et al. (2008), Verhoef et al. (2014), and Ravignani et al. (2016).

5.1. Kirby et al.'s Artificial Language Learning Experiment

In 2008 Kirby et al. designed a diffusion chain experiment where participants had to learn an artificial language. The task for participants during the diffusion chain was to "[label] visual stimuli with strings of written syllables" (2008: 10682). All participants were human adults and the diffusion chain was made up from 10 generations. Furthermore, it is important to note that no intentional or conscious design was applied by the participants for the emergence of these artificial languages. The artificial language that had to be learnt during the experiment was comprised of a signal-meaning mapping between a set of labels (signal) and colored moving shapes (meaning).

The experiment was split into two stages. During the first stage participants learn an artificial language where units consist of a label and a corresponding visual, as described above. An instance of such a signal-meaning mapping is given below in figure 1. There we can see an example of a label such as the quattro-syllabic *kihemiwi* and a visual that is a green zigzagging square.

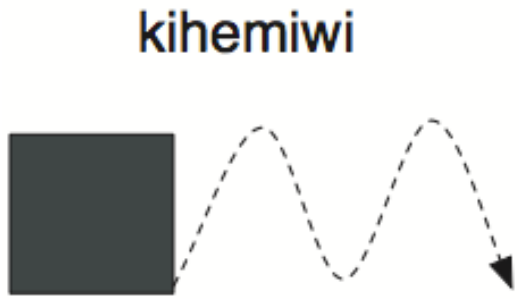

Figure 1 String-picture pair from Kirby et al. (2008: 10682)

All such string-picture pairs were split into two sets: SEEN and UNSEEN (2008: 10682, 10685). Participants were trained only on the former, and subsequently tested on both during the second stage. During training they were presented both the string and the picture, and during testing they were presented the picture only and asked for the string they think would fit. The initial set a participant learns from was randomly generated. The output of this participant

during testing was used as the learning set for the next participant, and this process was repeated until 10 generations were reached. Then the data was statistically analyzed. Interestingly, the outcome of the study matched the results from the IL modeling studies: an increase in learnability and structure.

The increase in learnability can be understood as a decrease in transmission errors. As such, transmission errors were calculated as a "measure of string similarity" (Kirby et al. 2008: 10682). This means that the researchers applied a method to measure the amount of variation between sequential generations in numeric terms. What was computed was the so-called Levenshtein Edit Distance which computes the minimal difference possible between the string-meaning mappings. More explanation and illustration of the Levenshtein Edit Distance will be given once we turn to the actual study of the thesis in section seven since it will function as an important measuring device as well.

The increase in structure was calculated as the "measure of linguistic structure based on measures of compositionality used in [Brighton et al. 2005]" (Kirby et al. 2008: 10682). What is meant by this is that the researchers calculated how systematically signals and meanings were mapped together. To illustrate, the meaning of a visual object in the experiment such as in figure 1 always contained three categorical features: shape, color, and movement. The amount of structure and systematicity was said to increase if it became more predictable how linguistic features of the strings were mapped to meaning features of the visual objects (Kirby et al 2008: 10682). As such, if members of a group of similar strings tend to be mapped to members of a group of similar visuals, then there is high systematicity. In contrast, if these mapping relations are totally random und unpredictable, then there is low systematicity.

Kirby et al. show the results for four independent chains. As predicted by prior modeling experiments, Kirby et al.'s data show an increase in structure and learnability (2008: 10683). Additionally, the overall number of strings decreases, and particular strings are used to refer to more than one visual object presented. Put differently, the language and its meanings become gradually underspecified

(Kirby et al. 2008: 10683). The researchers stress that while this explains the decrease of transmission errors, it does so only to some extent since there were instances where all signals were mapped to the correct meaning, "including meanings drawn from the UNSEEN set" (2008: 10683). Thus it remains to be explained why and how participants were able to correctly map signal and meaning from a set they have never seen during training.

It is plausible that the increase in structure is responsible for this. Initially, the language is unstructured and the mappings between signal and meaning are totally random. Kirby et al. put forward that "nothing in the set of signals gives any systematic clue to the meanings being conveyed", and that rote learning appears to be the only way of acquisition no matter whether the random mappings are underspecified or not (2008: 10683). However, "rote learning for all meanings is impossible" since participants are never trained the UNSEEN set (2008: 10683). Thus the question stands what made successful transmission and generalization possible. We can assume that only an increase in systematicity and structure can allow participants to make successful generalizations for unseen meanings. This is confirmed by the results: specific strings become more regularly mapped according to categorical meaning features of the visual objects throughout generations. For example, all visuals that depict an object moving horizontally are regularly called *tuge* by later generations. Clearly, Kirby et al. 's measurements "confirm that the languages evolve to become more structured", and the researchers argue that this increase in structure supports the induction process of participants (2008: 10684).

Remember that one of the aims of the diffusion chain experiments was to find out whether the data show similar properties to natural human languages. Although the experiment seems to have successfully demonstrated the predicted outcome, underspecification and the ambiguous labeling would reduce the expressiveness of a language. For example, in the most extreme cases participants used only one string for all visuals. Therefore Kirby et al. slightly adjusted their experiment design in order to counteract underspecification, and repeated the study. The adjustment made was a form of filter for "the SEEN set

before each participant's training" (2008: 10684). This filter constrained the set during the training stages so that for each visual there was only one single corresponding signal. Filtering was applied during prior testing stages. All but one of the visuals were erased randomly if any signal had been mapped to more than one visual. Thus, subsequent training sets were never structured by underspecification or ambivalent labeling.

The results of the second experiment were rather similar to the results of the first experiment: both learnability as well as structure increased. However, the filter seemed to have had an impact on the development. For one, the output of participants did not show a decrease in the total number of strings as in the prior experiment. More importantly, given the fact that an increase in structure was observable albeit the filter blocked the emergence of underspecification, the increase in structure could not have been reliant on the emergence of underspecification. Rather, structure increase appeared to be influenced by inferences about the morphological structure of the strings themselves. To illustrate, each signal evolved so as to consist of three morphemes, where each morpheme either referred to a visual object's movement, color or shape; reflecting "regularities in the visual scenes" (2008: 10685). Kirby et al. note in a later paper that "compositional languages developed, where sub-parts of each complex label specified components of the picture that label referred to" (2014: 110).

Lastly, in their discussion Kirby et al. ask whether there is any resemblance between the emerging structures in their diffusion chain experiment and natural human languages. Although underspecification restricted expressivity in the first experiment, there are still similar cases of underspecification in natural human languages. For example, nouns that are not proper names are underspecified in as much as they do not tend to refer to "specific entities" whereas proper names do (Kirby et al. 2008: 10585). Further similarities appear to be a "collapse of distinctions based on color" attested in natural languages due to cognitive biases where object shape is more readily linguistically encoded than color (Kirby et al. 2008: 10685). However, one should be aware that only one single participant

assumed color to be less important in this 'alien' language during posttest discussions (2008: 10685). Hence, it remains questionable to what extent assumptions about similarity are valid here. Nonetheless, the example shown on figure 3 in their paper (Kirby et al. 2008: 10683) does demonstrate this kind of bias where no systematic labeling is present with regard to color, albeit no overt comments about this are made in the text.

A final parallel feature is compositionality. Kirby et al. define compositionality by virtue of "[t]he meaning of an expression normally [being] a function of the meanings of subparts of that expression and of the way the subparts are put together", granting language learnability and expressivity (2008: 10685). Accordingly, an emergence of systematically compositional structure was observable when the experiment was repeated with a filter blocking underspecification.

5.2. Verhoef et al.'s Slide Whistle Experiment

In their study on "experimental iterated learning" (2014: 59) Verhoef et al. created an artificial language devoid of meaning via a slide whistle. Their research focus was specifically the emergence of combinatorial structure, and whether the emergence of combinatorial structure is driven by motivations yielding either to distinctiveness or rather cognitive economy (2014: 57). Four diffusion chains were created, each consisting of 10 generations. Each experimental task took up to 60 minutes where participants heard 12 different whistle sounds one by one during a training phase. Then during recall phase, participants had to recall all twelve whistles in any order. Typically, the output produced during the recall phase became the training input for the next participant. The first stimulus of each chain consisted of selected whistle sounds from a prior pilot study with multiple participants. These whistle sounds made up a set which "was constructed so as to appear to the experimenters to contain minimal combinatorial structure" (Verhoef et al. 2014: 60). Additionally, it was made sure that this set did not contain more than two whistles produced by the same person.

An interesting choice made by the researchers was their "reproduction constraint", reminding of the filter in Kirby et al. 2008 (Verhoef et al. 2014: 60). During recall phase, this constraint prohibited that imitated signals be too similar to each other. Note that what is meant here is the relation among copies, not the similarity between stimulus and copy. For this, the tolerance threshold chosen was set rather high since the main intent was to block "repetitions of the same signal" (Verhoef et al. 2014: 60).

It is somewhat confusing why Verhoef et al. use the term 'homonymy' for identical signals given that the whistles do not contain any meaning. The same can be said for their use of the term 'combinatoriality', which is traditionally referred to the conception that expressivity of human languages stems from the capability to construe a limitless amount of meanings from a limited amount of building blocks. Again, no notions of meaning are found in Verhoef et al.'s data. Contrastingly, for Kirby et al. 2008 these terms makes sense since there is a signal-meaning relation in the data. Similarly, the application of the filter in Kirby et al. 2008 was justified since its function was to prohibit "few words covering lots of meanings" (Verhoef et al. 2014: 60). As it stands, in Verhoef et al. 2014 there is no sound reason given for this procedure besides the fact that it was applied in Kirby et al. 2008.

Frankly, Verhoef et al.'s reasoning that "it is likely that participants in our experiment forget which whistles they already recorded and [that] there is no natural communicative pressure to preserve expressivity" (2014: 60) is merely a necessity of their experimental design and does not sufficiently justify the application of a filter constraint. For one, given that the task is supposed to be difficult in order to demonstrate processing biases, transmission errors and signal assimilation are natural developments of any diffusion chain where a challenging number of signals are to be imitated. But more importantly, expressivity cannot be preserved, whether with or without filter constraints, since expressivity demands semantic content; something which is explicitly absent from Verhoef et al.'s data. Furthermore, in Kirby et al. 2008 the participants never felt immanently any effects of the filter during the experiment

task since the filter was applied to the data afterwards. In Verhoef et al. 2014 on the other hand, filtering was applied during the experimentation task. Thus the possibility of data distortion due to interference factors should not be ruled out. Nonetheless, the automatic filtering procedure presented in Verhoef et al. is impressive, and it is informative to know that something like this is possible.

Verhoef et al.'s results from their qualitative analysis seem to show signs of combinatorial structure (2014: 61). Of course, the question still stands whether the signals are truly combined from smaller units, or are rather restructured so that they become more memorable. To put the question differently, from which direction was combinatorial structure generated: bottom-up or top-down? The former assumes a synthetic combination from smaller building blocks, the latter the segmentation of units from holistic phrases in order to simplify processing. One should be aware that just because the data resemble a combinatorial structure, it does not necessarily mean that the cognitive process of generating said data made use of combinatoriality.

Some cases in Verhoef et al.'s data point towards use of combinatoriality while other cases do not. For example, a participant made a succession of whistles where the first whistle consists of a {high-falling-glide low-falling-glide} configuration and the second of a {high low high low fluctuate} configuration. Interestingly, the next participant repeated this succession as {low-falling-glide high-falling-glide fluctuate} and then {high-falling-glide low-falling-glide fluctuate}. Here, Verhoef et al. assume that the last whistle of the second participant is a combination of the first and second whistle of the prior participant. However, there are two other possible explanations to this not mentioned by the researchers. One is to assume that the second whistle by participant 1 was simplified by participant 2: instead of having separate {high low high low} whistles, participant 2 connected these into two falling glides. The second is to assume a case of metathesis between the second participants first and second whistle where the {low-falling-glide high-falling-glide} is inverted to {high-falling-glide low-falling-glide}. The same case of metathetical inversion can be made for the first participant's first {high-falling-glide low-falling-glide}

whistle and the second participant's first {low-falling-glide high-falling-glide fluctuate} whistle.

There is a problem with such qualitative analyses and Verhoef et al.'s experiment design in general. It is not at all clear which interpretation presented above is the correct one, and parsimony does not seem to help. Given the fact that all participants first heard the whole stimulus set and were then asked to repeat what they have heard in any order obfuscates our choices to map any heard stimulus signal with any copy signal intended by the participant. This leaves mapping choices open to the varying interpretations of any researcher during analysis. Then again, it is quite possible that Verhoef et al. do not attribute much importance to mapping. Rather, what could have been intended by the experiment design was to show which segments remain during memorization and how these recombine into novel configurations; assuming recombination is the cognitive process applied by the participants.

Consider that to segment a phrase would require being able to hold a cognitive representation of a segment in order to decide what chunk of a phrase is to be segmented in the first place. It makes sense that the same would apply to the transformation of segments of a phrase such as in cases of metathesis. It makes even more sense that the same applies to combining two units together. Thus, all of the interpretations presented above would require combinatoriality, since each process of segmentation, transformation or recombination requires the mental representation of a separate chunk, segment or unit. As such, the most likely interpretation would be that the emergence of combinatorial structure was driven by an interaction between these cognitive processes mentioned.

In Verhoef et al.'s quantitative results the decrease in recall error is understood as an increase in learnability, similarly to Kirby et al. 2008. This is measured as the minimal distance between input and output structure. Interestingly, Verhoef et al. did not measure the pitch of the signals but the plunger movement. The plunger is the part of the whistle that can be moved in two directions to determine the pitch during aspiration. The reason for measuring plunger

movement instead of pitch was “the non-linear relation between pitch change and plunger movement” and the fact that “participant behavior was predicted better by the movements of the plunger than by the acoustic signals ”(2014: 63). Indeed, the measurements and quantitative analysis show that there is an increase in learnability over generations.

Regarding the increase of structure, Verhoef et al. applied a measure of entropy, also known as Shannon Entropy (Shannon & Weaver 1949). The Shannon Entropy will be explained later in more detail since it will be applied to the experiment data of this thesis as well. Before measuring entropy, Verhoef et al. segmented the signals at silence points. Then the segments that were similar enough were clustered together to categorical building blocks. Then the researchers computed the Shannon Entropy of the data as the “probability of occurrence of [a] building block” (2014: 64). The results confirm that there is an increase in structure over generations.

Additionally, Verhoef et al. measured the “associative chunk strength” of the signals (2014: 64; from Knowlton & Squire 1994). This was to find out whether the “order of appearance” and the configuration of signal sequences increased in structure as well. Indeed, their results confirm also an increase in “sequential structure” (2014: 64). Further measurements show that the similarity between signals increased, which seems contradicting since Verhoef et al. assume dispersion to happen before combinatorial structure emerges (2014: 65). However, they state that this does not necessarily suggest that no influence with regard to structure increase is to be attributed to dispersion (2014: 65). They show that instances of dispersion can be attested by applying a measure of energy used by Liljencrants & Lindblom (1972) specifically on “the building blocks that construct these signals” (2014: 65). According to their results, dispersion of building blocks becomes increasingly more frequent throughout generations (2014: 66). As such, Verhoef et al. conclude that only one causality process for the emergence of structure is rarely sufficient, since both increases of similarity and dispersion have been attested in their data (2014: 66).

In their discussion Verhoef et al. mention the problem of continuous signals. Continuous signals prove to be difficult to analyze, since the basic elements to construct them are not clear (2014: 67). As will be shown and explained later in our experiment, this problem is solved by the usage of syllables and Prominence-centers as boundary markers. Furthermore, Verhoef et al. try to attribute some meaning to their signals. However, the meanings attributed fail to conform to Hockett's notion of Duality of Patterning (1960) since the meanings in their data always refer to the structure of the signals themselves. Lastly, they make an important note at the end with regard to the function of diffusion chain experiments in general (Verhoef et al 2014: 67):

> [T]he results of this type of work should not be interpreted as an attempt to reconstruct the evolution of language, but as a method to shed light on what mechanisms may be involved in the emergence of structure.

5.3. Ravignani et al.'s Drumming Experiment

In contrast to the two studies presented above, Ravignani et al.'s study is rather concerned with music and rhythm than language. However, as we have learned from prosodic PL theories, there might exist crucial connections between music and language from the perspective of language evolution. Ravignani et al. (2016: 1) claim that there exist six human universals that carry rhythmic features. Additionally, it is stated that percussive behavior is observable in great apes, which are close relatives to the human species. These two postulations are followed by the question of whether percussive behavior was present in human predecessors millions of years ago. The idea is that rhythmicality might be a fundamental feat of human cognition that was present in our ancestors and is still present today in modern Homo sapiens. A way to test whether such rhythmic biases exist is the experimentation on IL via diffusion chains. Thus, the researchers work with basic psychological mechanisms such as working memory, perceptual primitives and categorical perception, and inquire whether the addition of cultural transmissions can lead to cognitive universals.

Through the usage of iterated learning, participants in the study consecutively imitated prior output in a diffusion chain. The researchers observed six developments throughout each diffusion chain (2016: 1-3):

1. increase of systematic structure
2. increase of learnability
3. timing patterns converge to durational categories
4. emergence of motivic structures
5. durations conform to isochrony & metrical hierarchy
6. chains evolve independently

These results show how instances of error during imitation can lead toward patterns that were easier to reproduce (Ravignani et al. 2016: 4). Although each chain resulted in an independent system or 'culture', systematic similarities between patterns within a single chain emerged. The study assumed that perceptual, learning and production biases were possibly responsible for these phenomena.

Regarding participant control, only subjects without existing skills in music performance were allowed to participate. Furthermore, instructions were set as to promote close replication of stimuli, prohibit any forms of deviation, and assure that consecutive sequences were not in any form related. The total count of participants was 48 and all were randomly assigned to 6 different chains. Each chain consisted of 8 generations, with generation 0 being randomly generated via computation. All chains had a different generation 0 as the initial stimulus of each respective chain.

Each participant was played 32 drum patterns, and initially each drum pattern consisted of 12 snare drum hits in generation 0. Moreover, the snare drum hits in generation 0 were assigned random velocity and random Inter-Onset Intervals (the time distances between consecutive drum hits), resulting in chaotic and unstructured rhythm configurations. In each drum pattern there was always a cymbal sound present to mark an end of the signal 1.5 seconds after the last snare drum hit. This cymbal sound was not included in the analysis. The participants listened to and imitated the 32 patterns twice; the first performance

was to briefly practice the task, and the second performance was recorded. The participants reproduced each pattern immediately after hearing it, and they were unaware that they would be listening to stimuli produced by a previous learner. Lastly, participants were asked to complete a questionnaire afterwards. The materials used were headphones, a drumstick, a drum pad, a laptop and an audio interface with a Python code that translated the recorded signals to milliseconds.

6. Thesis Study: Iterated Learning of Syllable and Feet Duration Sequences

What these diffusion chain studies managed to show was that structural regularities matter to human cognition. They matter insofar as that the stimuli in the experiments are processed and reproduced under the constraint of cognitive biases that promote an increase in structure and learnability during the diffusion of behavior. One of the studies showed that there exist cognitive preferences for stimulus structures that were rhythmically organized (Ravignani et al. 2016). Given what we know about the relations between music and language from prosodic PL proposals, we might enquire the status of rhythmicity in a more language-like diffusion chain than the ones presented in Ravignai et al. 2016 and Verhoef et al. 2014.

What follows is a report of a replication of the experiment in Ravignani et al. 2016 with important adjustments. Instead of hearing and replicating drum hits with drumsticks, participants will imitate strings of spoken syllables. The question is whether similar cognitive biases with regard to rhythm apply to the diffusion of behavior in a different production medium; namely speech. Accordingly, the hypotheses are that the syllable strings in the diffusion chains will show an increase in structure and learnability. Additionally, later we will enquire whether similar claims can be made for feet as well.

6.1. Stimuli

Regarding the stimuli, syllables were used instead of snare drum hits. The syllables consisted of two phonemes with maximal distance in sonority: a voiceless plosive obstruent and a vowel, which when put together sounded like /ta/. There are a number of reasons for choosing this type of syllable. Firstly, the syllable itself and its reduplication into multisyllabic units does not carry any established meaning for participants of the study who were all native speakers of German. This made sure that there existed no semantically motivated preference for perception decoding and reproduction during the imitation task. Secondly, the composition of two sonorously maximally distinct phonemes into an open syllable allowed for a clear distinction of pulses in syllable chains, which further ensured that participants were not distracted by factors other than rhythmic perception. Thirdly, the voice onset time (Lisker & Abramson 1964) of a voiceless plosive obstruent followed by a vowel is almost always close to zero, which reduced the degree of idiosyncratic variation of participants. Additionally, /ta/ syllables allow for a more accurate localization of their rhythmically prominent peaks and made annotation with PRAAT (Boersma & Weenik 2018) easier. Lastly, note that stimuli consisting of only /t/ chains without clearly distinguishable sonorants in-between might not only have been too percussive and thus not linguistic enough, but also severely harder to annotate.

Truth be told, the measuring procedure in Ravignani et al. 2016 is superior regarding its precision since it is fully automatized. Thus, a comparatively less precise annotation procedure by eye and ear must approximate as closely as possible with the use of any cues and tools available. What is more, this approximation has to take into account the different acoustic make-up between drum hits and speech: whereas single percussive drum strokes can be located via automatized beat detection, continuous speech signals must convey prominent pulse peaks so that any metrical analysis is possible. It is harder to locate an exact most prominent point in time for a continuous signal than for an instantaneous drum stroke. To locate measuring points for the latter, it is only necessary to time-stamp the moments when the drumstick hits the drum pad. Contrastingly for continuous speech signals, it does not suffice to time-stamp the

beginning of their articulation. As will be elaborated later, reasons for this are grounded on articulatory and psychoacoustic facts about so called Prominence Centers in natural speech signals (Couper-Kuhlen 1993: 17).

The random duration of all phonemes for generation 0 were computed via a Perl-script, with consonants being kept shorter than vowels to appear more natural. The Perl-script then wrote its output into an MBROLA .pho-file, which was then read by an MBROLA synthetic voice speaker (Dutoit et al. 1996). The model speaker chosen was a male adult speaker of Bavarian German conforming to basic German phonotactics. No velocity randomization was applied, since loudness is the least informative stress cue in a natural speech signal when compared to phoneme duration and intonation pitch (Couper-Kuhlen 1986: 24). A string of 12 syllables constituted an intonation phrase with intonation constantly falling by 10 Hz to make the speech signal more natural as well. Each syllable string was followed by an approximately 1.5 seconds long high pitch to low pitch /m/ glide that signaled the end of a string. A total of 6 strings were presented to each participant. At this point it should be noted that the experiment design needed to yield to the fact that a brief experiment taking 10 to 15 minutes increased the likelihood of a person agreeing to participate since no incentive compensation was offered.

The generation of randomly structured stimuli was repeated four times so that each chain A, B, C and D had a different generation 0, with one exception. The generations 0 of C and D were in fact identical with regard to syllable durations, but while the former contained only /ta/ syllables, the latter included random variation in vowel choice between /a/ and /i/ for each syllable. This was to investigate whether an increase of phonological variance would affect performance and how it would differ from the other chains. Note that vowel randomization was constrained in so far as to prohibit sequences where more than three identical vowel variants followed in a row.

6.2. Predictions

One possible prediction with regard to random vowel variation would be that participants could perform worse due to distraction, since they would have had more features to remember and imitate during the task. It is likewise predictable that the different phonetic make-up and articulatory features between /a/ and /i/ could affect performance as well. Both front vowels are known to be rather demanding with regard to articulation, but for different reasons. The vowel /a/ usually demands a lowering of the jaw, and /i/ demands a high tongue position. However, the fact that the preceding dental consonant /t/ demands a closure of the mouth plus contact of the tongue with the upper teeth might make articulation of /ta/ more strenuous than of /ti/. It is not necessarily clear what predictions could follow from this. On the one hand, /ta/ syllables might be kept short, due to their strenuous character, and we could readily assume that in such cases the vowel is demoted to a schwa and loses its jaw-lowering feature. On the other hand, /ta/ syllables might be kept long, since their articulation already demands a high investment in participants' articulatory motor control activation. Or more drastically, one syllable type might be preferred so that the other becomes less frequent or even disappears during transmission.

Note also that in Ravignani et al. 2016 the drum hits of different participants sounded generally the same; human voices generally do not. Such differences in pitch frequency range and timbre are especially perceivable between male and female speakers. However, it is not clear what predictions might follow from this, and assumptions that some participants might have performed differently due to being more or less sympathetic toward the stimulus they perceived seem more like speculation. Yet, the bare fact that a percussive drum stroke is different in acoustic information from a continuant /ta/ syllable might lead to different results. For example, whereas the rhythmic prominence of a drum stroke is influenced by the hit-velocity, a natural speech signal's prominence is more influenced by syllable length and pitch shift, with velocity or volume being the least important stress cue. Thus, since in our study there is no direct correspondent of randomly generated velocity for each unit, it is possible that

some syllables, e.g. those which are relatively longer, will receive more stress and be perceived as carrying more stress than others.

Lastly, if syllable concatenations are processed as feet, it is possible that the last mono –or multisyllabic foot will be increased in duration relative to the preceding feet, similar to intonation phrases in natural speech. Note further that all these predictions deal with duration. One cannot ignore the possibility that suprasegmental effects might influence morphophonological features as well: phonemes might change from /tata/ to /tada/ due to assimilation processes of certain units yielding to phonetic environment, and phonemes might be disconnected and rearranged (e.g.: /ta ta ta/ becomes /tat ata/).

6.3. Procedure

The procedure was similar to that of Ravignani et al. 2016. Initially during every performance, each participant listened to one string and imitated it once to practice production. Then each participant listened to all 6 strings, replicating each string (including the final glide) immediately after hearing it while being recorded. Each such recording of 6 strings was presented to the successive participant with the same procedures. This was repeated 8 times for 4 different chains A, B, C and D, where each chain starts with a different generation 0, with the exception of C and D. As such, the study required a total of 32 participants who were all native speakers of German and who were not musicians, dancers, or actors. The participants were instructed to listen to, memorize and imitate the signals to their best capabilities. After the task, each participant was asked to fill in a questionnaire. The required materials for this study consisted of a pair of headphones, a large membrane condenser microphone, a laptop to operate a digital audio workstation for recording, and an external audio interface as a pre-amp which supported the condenser microphone with phantom power.

6.4. Prominence-Center Localization in PRAAT

After the experiment, the recordings were imported to PRAAT and were annotated for the Prominence Centers of syllables. We assume a Prominence Center (PC) to be the most acoustically salient point in a syllable with regard to

both production as well as perception (Couper-Kuhlen 1993: 17). Note that PCs are not found at the beginning of left-aligned syllable boundaries of consonant-initial units, and that to only determine the nucleus of any syllable is not precise enough for this exact phonometric measuring. Still, the hierarchical position of the nucleus as a core unit in syllable structure is not challenged since PCs are generally found in close proximity to them.

Nonetheless, it has to be stated that the location of a PC can vary according to the phonetic environment. More specifically, it is relative to "the length of the initial consonant of the syllable" (1993: 18). As such, the onsets of words or vowels as well as intensity peaks of vowels can be excluded as possible PCs (1993: 19). Since all syllables in the experiment carry only one unvoiced dental plosive as its onset consonant, PC localization did not need to take this kind of variation into account. Generally speaking, consonant initial syllables carry their PC somewhere at the release point of consonant articulation, closely antecedent to nucleus vowel articulation. This means that in syllables the PC is found in the proximity of the "offset" of the final onset consonant (1993: 21); or put differently: at the transition between /t/ and /a/.

The empirical validity of PCs is well attested by prior experiments on prosodic and rhythmic perception (Couper-Kuhlen 1993: 14-18). More importantly, for this study PCs are crucial measuring points that mark the boundaries of duration intervals. In Ravignani et al. 2016, such units of rhythmic duration are called Inter-Onset Intervals (IOIs) and they encompass the time distance from one beat pulse to the succeeding beat pulse. Since this experiment uses CV-syllables instead of drum strokes, the term IOI will be used when referring to Ravignani et al. 2016, and the term Prominence Center Interval (PCI) will be used for the present study.

Annotation in PRAAT was managed by checking both oscillograms as well as sonograms for relevant cues. The visualization of auditory data made it possible to locate the PCs of the syllables. A visual illustration of the exemplary oscillograms and sonograms is given in figure 2 below.

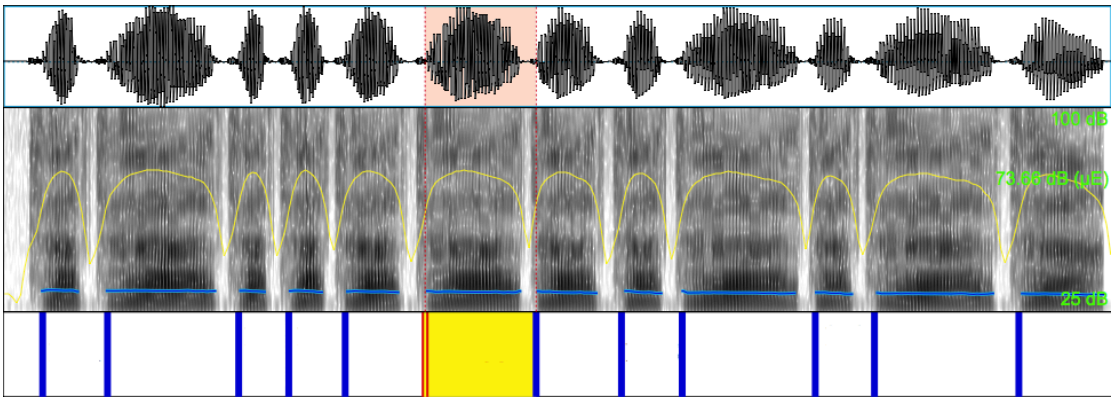

Figure 2 Screenshot of a PRAAT annotation of twelve /ta/ syllables by a synthetic speaker (generation 0 of chain A).

The first row and second row depict the oscillogram and sonogram of the first syllable string consisting of twelve syllables from generation 0 of chain A. The lowest row shows where the measuring marks for PCs (blue bars) were set, which make up eleven PCIs. Note again that the PC locations were not marked at the beginning of any /ta/ speech signal. As discussed before, rather these PCs are annotated for at the offset-moment; this is exactly before the vowel gesture starts and after consonantal closure is released. This moment of release after the consonantal closure is clearly observable with PRAAT's pitch detection.

The PC localization process was mainly supported by pitch detection (blue lines in the sonogram), velocity measurement (yellow lines in the sonogram), and the waveform in the oscillograms. All blue bars were set at exactly such points for all syllables: before the oscillogram wave creates a drastic bump, after velocity had already shown some increase, and right before the occurrence of pitch has been detected. One further auditory cue that was applied helped confirm the validity of the PC location: when playing the recording from the exact point of the presupposed PC, it usually becomes nearly impossible to distinguish the initial /t/ from other voiceless plosives. This is possibly due to the lack of informative feature cues during consonantal closure which pre-inform the perception of consonantal release. After annotation, a PRAAT script was used to calculate the durations of all PCIs and export the data to a text-file. This txt-file was then imported into R Studio for statistical analysis.

6.5. Statistical Analyses & Hypotheses

The statistical analysis was done with R Studio (RStudio Team 2015). Its aims were to account for the statistical development of PCI durations. Two hypotheses, similar to Ravignani et al. 2016, are put forward to predict the outcome:

(i) The distribution of PCIs becomes more categorical

(ii) Imitation becomes more accurate

Generally, it was predicted that the results would be similar to those in Ravignani et al. 2016. Thus (i) relates to an increase in structure and (ii) to an increase in learnability. In this sense, (i) means that the durations of PCIs will increasingly group into, for example, short and long durations where any PCI duration resembles either the short or long category. Further, (ii) implies that participants make fewer mistakes during imitation and that the signals become easier to learn. Additionally, (i) projects two further sub-hypotheses that are more precise:

(i-a) PCI duration categories approximate to relational multiples

(i-b) PCI distribution of later generations reduce in a measure of statistical entropy

What is meant by (i-a) is that if short category PCI durations tend to cluster around 10 milliseconds, long category PCI durations will cluster around 20 or 30 milliseconds. Likewise, if the former cluster around 20 milliseconds, the latter will cluster around 40 or 60 milliseconds. To briefly address the second sub-hypothesis, (i-b) refers to a method of measuring how predictable the sequences of PCI durations become throughout the diffusion chain, and will be explained in more detail later in the results.

6.6. Results:

Of all participants, 23 were female and 9 were male, and the mean age was 23. After the task, all participants answered 5 questions on a questionnaire:

1. How difficult was the task?
2. How similar were the patterns to each other?
3. How musical did the patterns sound?
4. How familiar did the patterns sound?

5. How 'language-like' did the patterns sound?

These questions were taken from Ravignani et al.'s study (2016: supplementary materials), with the exception of question number 5 being my own. All five questions were supposed to be answered with a Likert scale ranging from 1 (not at all) to 7 (extremely).

Generally speaking, all five questions were answered ranging between approximately 3 and 4 on average. Furthermore, the questionnaire results show that the task was generally considered to be rather difficult than not. This correlates with the loss of the number of PCIs due to syllable omission, as is clearly visible in figure 3 below.

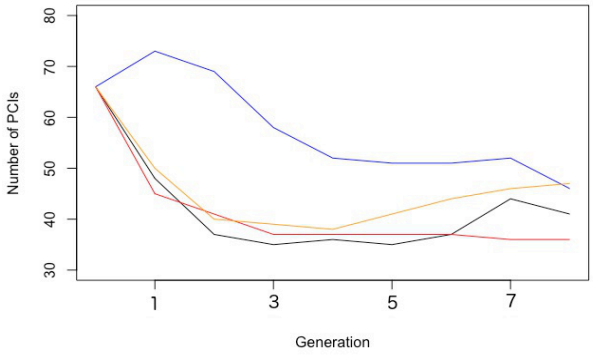

Figure 3 Loss of the number of PCIs for chain A (black), chain B (blue), chain C (red), and chain D (orange) throughout generations 0 to 8.

All generation 0 performances consisted of 66 PCIs in total. Typically, participants could not memorize all syllables given in the stimulus, which resulted in a loss of number of PCIs. However, in some instances, participants added syllables during imitation. Note that the task was meant to be difficult in order to elicit cognitive biases during memorization and imitation. As such, these effects were to be expected. Additionally, it is visible that drastic omission of syllables ceased approximately after generations 2, 3 and 4. This is in accord with the questionnaire answers on task difficulty: later generations found the task to be less difficult than earlier generations. Furthermore, D is the only chain where there is a constant increase in number of syllables. This increase starts at generation 4 and continues until the last generation. Moreover, questionnaire

results show that the sound patterns in chain D were considered to be more 'language-like' and more difficult to imitate than in chain C. Finally, the sound patterns in chain A were increasingly considered to sound more and more musical with each subsequent participant. Note that this trend is to some extent visible in the other the chains, however chain A is the best example since it shows the least amount of fluctuation. At this point it could be interesting to look for correlations between specific questionnaire answers and particular performances. However, this will be postponed until after the Kernel Density Estimation plots have been presented since they can greatly inform this analysis.

6.6.1. Kernel Density Estimation of PCI durations

How does one corroborate (i)? For this, we will have to show how densely the PCI durations are distributed among a single generation, and how this density of any particular PCI duration group develops throughout a chain. This development is shown in figure 4 below, where the PCI durations are mapped to the x-axis and their density is mapped to the y-axis. As such, the more similar PCI durations are and the more frequent any similar PCI durations are, the higher their density. This means that if all PCI durations were absolutely equal, a KDE plot would show a straight vertical line. However, if all PCI durations differed from each other by the exact same value, the KDE plot would show a horizontal line.

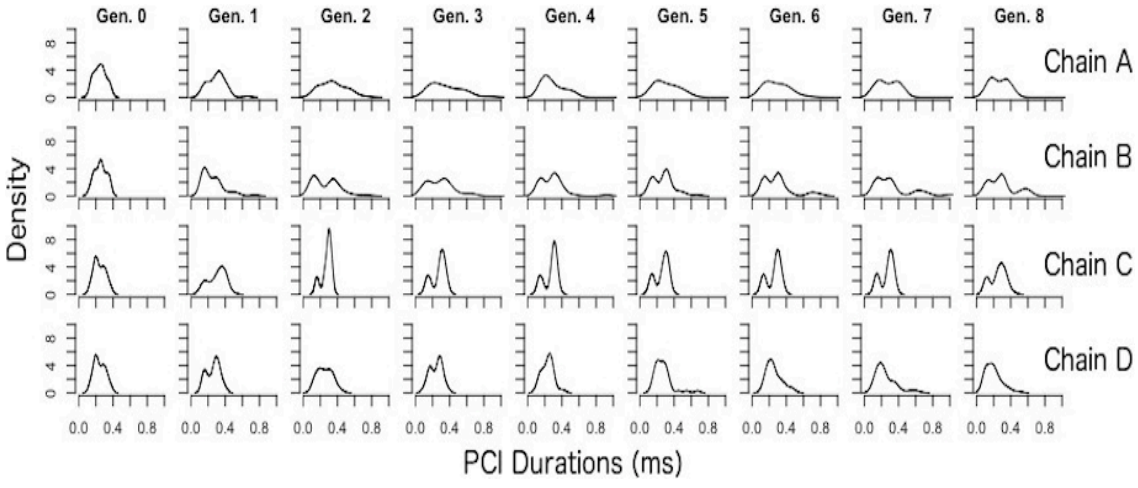

Figure 4 KDE plots showing the Density developments of PCI durations throughout generations 0 to 8 for chains A, B, C and D.

The KDE plots in figure 4 show similar results to Ravignani et al. 2016. On a first observation of chains A, B and C, the data seems to corroborate hypothesis (i) and (i-a): PCI durations tend to cluster around two durational categories in later generations. For chains B and C, the first cluster lies approximately at 15 milliseconds and the second lies approximately at 30 milliseconds. Further, chain A appears to develop clusters around 20 and 40 milliseconds; a development which occurs in relatively later generations (7 and 8) when compared to B and C.

Interestingly, all chains show that participants of every generation 1 already start to group the durations into two categories. For chain A, this is somewhat disrupted throughout subsequent generations, but two density peaks reemerge and are visible again in generation 7 and 8. More importantly, the development of relational multiples seems to be usually delayed and is observable in later generations. Here, one development of the data needs to be highlighted: although chain B and C start with very distinct generation 0 density distributions, both maintain very similar relational multiples of PCI duration categories since generation 4, although the peaks are higher in C than in B. Furthermore, chain C appears to be most stable in maintaining two durational categories. Contrastingly, the stability of chain B is interrupted by the emergence of a third category approximating between 60 and 70 milliseconds in generations 6, 7 and 8.

Note that such stable development is generally not visible for chain D: although both chain C and D start with the exact same PCI durations in generation 0, the development of durational categories is not as stable in the latter as it is in the former. Only in generation 1 and 3 are two observable density peaks attestable, whereas the rest of chain D does not stabilize palpably around any duration cluster throughout generations. Indeed, one possible explanation for this might be phonological interference due to random vowel variation, as participants need not only memorize PCI durations but the sequence of alternations between /ta/ and /ti/ syllables as well. Again, this appears to be in accord with the questionnaire results, where the patterns in chain D were considered to be harder to replicate than in chain C.

6.6.2. Questionnaire Revisit

Analyzing the questionnaire results and the KDE plots in combination can lead to further interesting insights especially when scrutinizing over specific answers in correlation to particular performances. Keep in mind that we can ask at least two questions concerning correlation factors: questionnaire results by participant X correlating with either the performance given by person X or the performance which person X heard as stimulus. We will concern ourselves with the latter since participants were asked to give their answers on they sound patterns they have heard and not the sound patterns they had produced.

Regarding task difficulty, the lowest scores were given by participants 6 and 8 from chain A, and by participants 4, 5 and 6 from chain C. Although both participant 6 and 8 from chain A gave the same scores, the stimuli they had heard were different. As can be seen from the KDE plots, the stimulus for participant 6 does not show peaks in density of PCI durations while the stimulus for participant 8 does. Does from this follow that the development of categorical durations in the form of density peaks has no correlation with task difficulty? After all we could simply assume that the low scores were given due to the fact that the overall number of PCIs decreased as the graphs in figure 2 show. Not necessarily. Consider the fact that the decrease of PCI quantity ceased after generations 2 and 3. Hence the number of PCIs cannot be the only factor. After all, the number of PCIs actually increased at generation 7 and the participant found the task to be very easy nonetheless.

So does an emergence in peak density structure make the task easier after all? It does seem so because participants 4, 5 and 6 from chain C all found the task to be very easy, and their stimuli all displayed PCI duration peaks in the KDE plots. The lowest scores on the other hand were given by the first participants of chain A and chain B. Given the correlation between perceived task difficulty, and number of PCIs and density structure of duration types, this does appear plausible: both participants heard generation 0 as a stimulus, and both were overwhelmed by the high count of PCIs as well as by the lack of clear

distinguishable duration category types. Still, future research should be targeted at being able to explain how the two factors of quantity and distributional structure affect task difficulty in more detail.

The second question dealt with the similarity of the 6 strings in a stimulus. Lowest scores were given by generation 4 from chain B and generation 6 from chain D, while the highest scores were given by generation 4 from chain C. Additionally, chains A and C generally reached the highest scores. Indeed, chain C displays the most stable development of KDE plots, but keep in mind that the question is not whether the generations are similar to each other, but whether the syllable strings of a single generation are similar to each other. Taking a look at the actual composition of strings in PRAAT it becomes apparent that the first and second string of chain B generation 3 have identical PCI duration configuration: {long long short short} throughout; an example of this for string 1 is given below in figure 5.

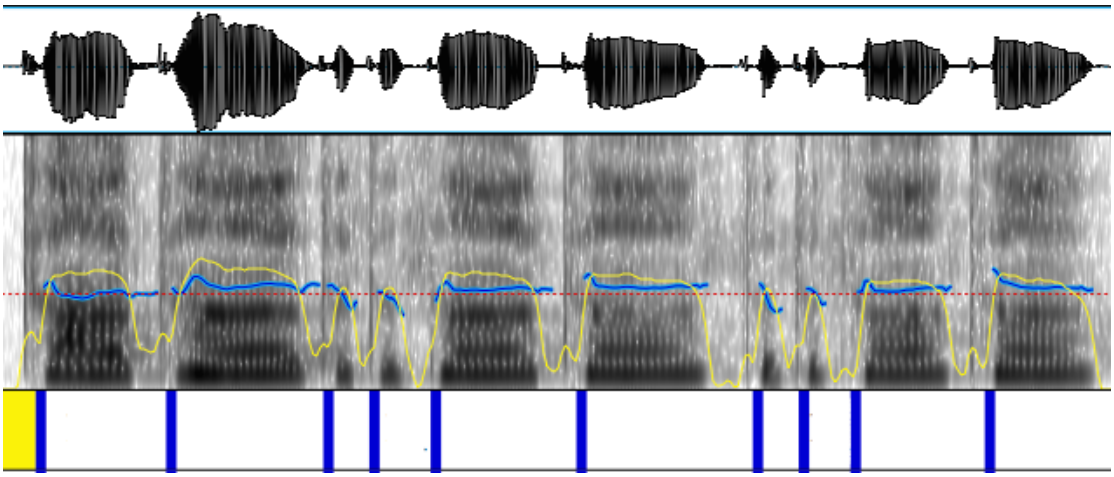

Figure 5 {long long short short} syllable string composition from chain B generation 3 string 1

Strings 4, 5 and 6 however have each at least one instance of {short short long long} sequences. Besides this we find no structural similarities. Contrastingly, all strings but the fifth from chain D generation 5 are {long short} configurations throughout, string 5 being the only one with only identically long PCI durations. An example of the {long short} alternation is given below in figure 6.

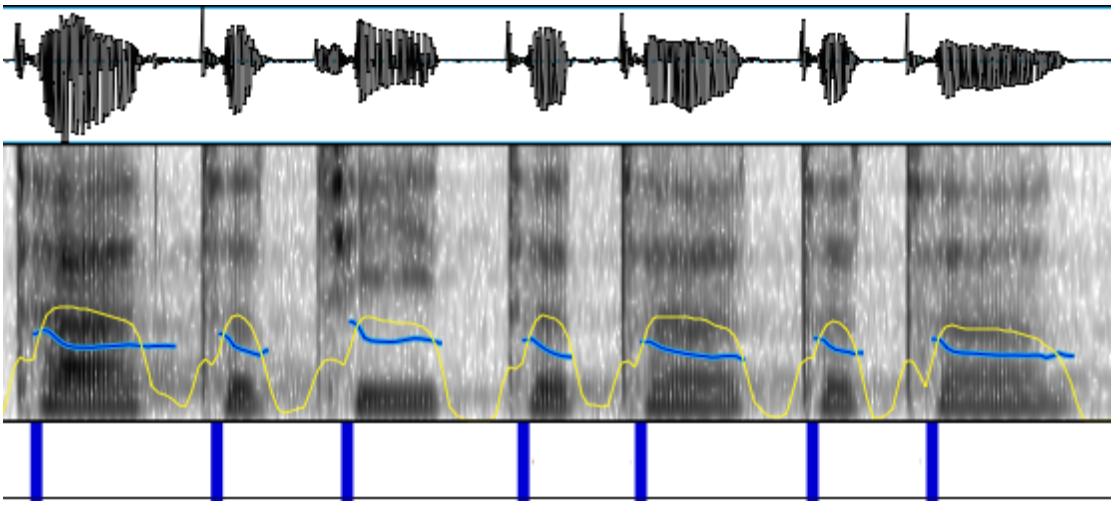

Figure 6 {long short} syllable string composition from chain D generation 5 string 4.

All but one string group configures into {long short} successions, yet the patterns were judged as 'not at all similar' by the participant. Given that we find more similar occurrences of string configurations in the chain D example than in the chain B example, why did both stimuli receive the same score? Or rather, why were the strings in chain D generation 5 perceived less similar than they actually are? One possible answer is that chain D includes vowel variation. Indeed: The mean results for question 2 of chain D are below average (3.5) and actually the lowest of all chains (chain A: 5.25, chain B: 4.125, chain C: 5.375).

Furthermore, mind the fact that generation 0 of chains C and D are identical and yet the strings of chain C were rated the highest on average with regard to similarity while the strings of chain D were rated the lowest in similarity. It seems clear that vowel variation must have had an impact on the perception of similarity of strings. But let us still be clear that not all string configurations of the generations of chains C and D are identical. Both chains did apparently have the potential to develop at least one string with identically long PCIs; the second string for chain C and the fifth for chain D. However, while the rest of chain C regularly configures into {long short long long short long} syllable successions, the rest of chain D configures into {long short long short long short} alterations. Each chain developed a very regular configuration of sound patterns where the

string compositions are very similar to themselves. Still: due to vowel variation chain D was perceived to hold the least similar sound patterns while chain C was perceived to hold the most similar sound patterns.

Question 3 regarded the musicality of sound patterns. The mean of all chains ranged between 4 and 4.75 and seems thus rather uninteresting at first sight. More interesting were the answers by specific participants. Among the lowest scores was participant 6's perception of generation 5 from chain B. However, the participant's stimulus shows clear peaks on the KDE plots. As the part about the measurement of entropy will show, the peaks approximate into multiples of each other. Typically, a succession of sound durations that are multiples of each other is a basic feature of music where durations are classified into e.g. fourths and eighths. So, either participant 6 did not hear that the durations of PCIs were distributed similarly to the relation of the duration of musical notes, or the participant does not see rhythmicity as important as other parameters in order to perceive the sound patterns as very musical. Apparently, what is perceived as musical or not is a subjective matter, and participants may very well vary in this regard. However, there is still a general trend that the musicality of sound patterns slightly increases on average among all chains with minor fluctuations. This would suggest that there is also a correlation between the KDE plot development and the increase of musicality; only that the question still stands to which parameters of musicality the participants were sensitive to. It is very well plausible that the question 'how musical is X' is simply too vague, since musicality can at least be divided into rhythm and harmony, which perhaps deserve their own question each in future research.

The fourth question concerned how familiar the participants were with the perceived sound patterns. The general trend of scores given is clearly fluctuating, and we find no observable correlations between specifics here. Interestingly, chain D shows the lowest mean scores (2.875) while chain C shows the highest. This could again be due to vowel randomization, or participants found the string configuration {long short long long short long} from chain C to be more familiar than the string configuration {long short long short long short} from chain D.

Lastly, question 5 asked participants how 'language-like' they found the sound patterns to be. Among the participants who gave the lowest scores are those who heard the voice synthesizer as their stimulus. It would seem that hearing unstructured strings of syllables generated by a computer had an impact on the judgment of these participants. Interestingly, vowel variation does not seem to have influenced the judgment of generations 1, since the first participants of chains C and D did not differ in this regard. Nonetheless, vowel variation does seem to have an overall impact on judgments about the sound patterns being 'language-like', because chain D was generally considered to be more 'language-like' than chain C, and we also find the highest scores being given by chain D participants. Further lowest scores were given by participant 4 of chain A and participant 6 of chain B. It could be that this is due to pauses between syllables we find in their stimuli, as can be seen below in figure 7.

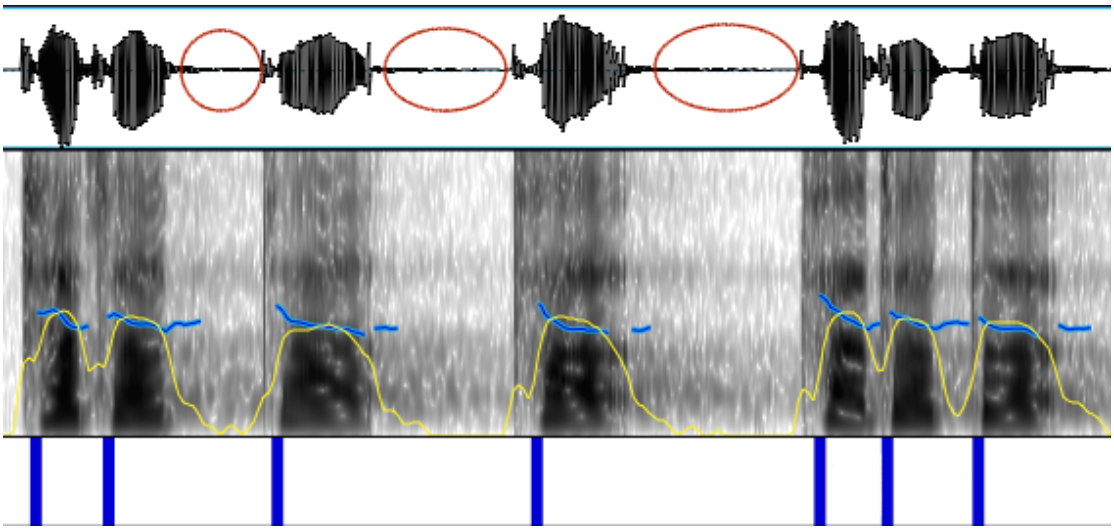

Figure 7 Occurrences of pauses (red circles in the oscillogram) in chain A generation 3 string 1

Pauses as these are rather unusual in casual speech where sound segments are typically connected. However, pauses were also present in chain A generation 4, and it was, contrary to the cases before, judged as extremely 'language-like'. Interestingly, later generations copied fewer pauses as they were reduced in duration and more and more syllables were connected. In chain A, only string 3

maintained a single long pause until the end of the iteration chain, always in close proximity to a very long PCI, as can be seen below in figure 8. Note that this is not the case with chain B, where the occurrence of pauses resulted in the emergence of a stable third duration category, whereas chain A lost the majority of very long PCI durations.

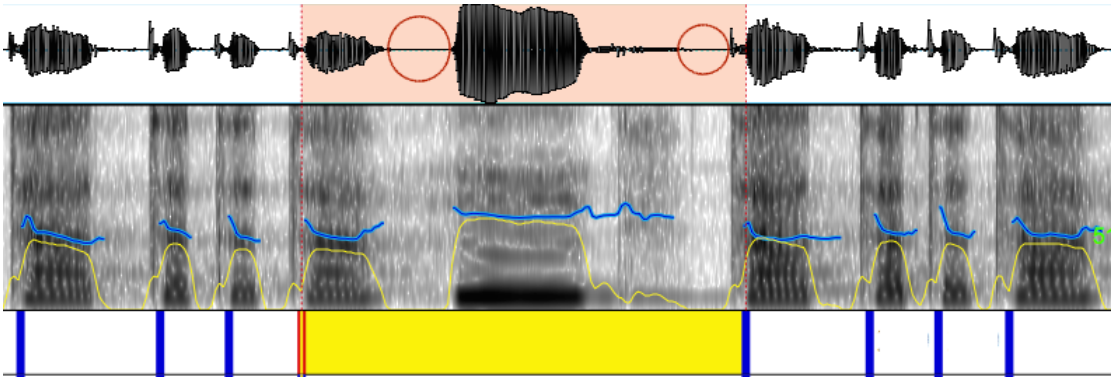

Figure 8 Two pauses (red circles in the oscillogram) within a long PCI in chain A generation 8 string 3

6.6.3. Vowel Variation

Before we turn to the measure of entropy, we should analyze the development of vowel variation in chain D in more detail. Figure 9 below shows the increase and decrease of /ta/ and /ti/ occurrences throughout all generations.

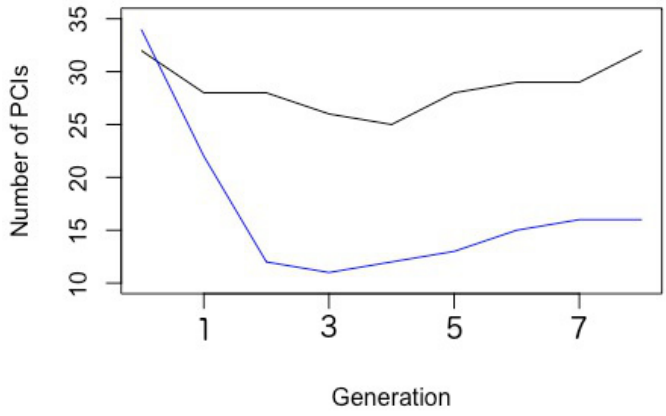

Figure 9 Number of PCIs in chain D that contain either the syllable /ta/ (black) or the syllable /ti/ (blue)

What becomes clear is that not only did participants omit syllables during imitation but they tended to rather omit /ti/ syllables than /ta/ syllables. Beginning with generation 0, the distribution of syllables containing the low or the high vowel is almost equal. Subsequently, the number of /ta/ occurrences slightly decreases and rises again after generation 4. As such, the number of PCIs that contain the /a/ vowel remains rather stable throughout generations. Contrastingly, the count of syllables with high vowel /i/ however drops significantly already during generations 0, 1 and 2. After that, it slowly rises but remains only half as frequent as occurrences of /ta/.

Even more interesting is the KDE development of chain D's PCI durations. To also account for vowel variation and make the differences visible, the KDE was calculated for PCI durations of syllables with either /a/ or /i/ as vowel. The results are quite striking and may possibly explain why chain D fails to develop clear durational categories similar to the other chains.

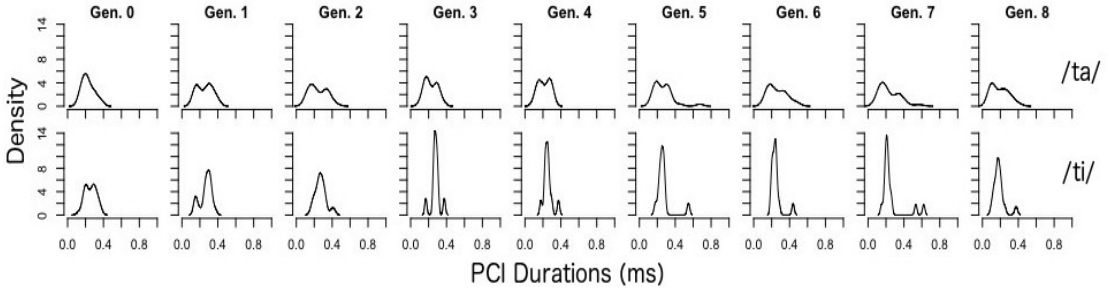

Figure 10 KDE plots showing the Density developments of PCI durations of /ta/ and /ti/ syllables throughout generations 0 to 8 for chain D

As is visible from figure 10, two small peaks emerge and remain stable for the PCI durations of /ta/ syllables. Similarly, two peaks are also visible for /ti/syllables, but then the durations cluster toward one peak already after generation 1. Note that the small peaks in /ti/ syllables after generation 1 are in fact single occurrences of these PCI durations and are thus statistically irrelevant. This is so not only because their density is significantly less than for the high peaks, but also due to their low stability during the diffusion of behavior. For example, in generation 3 of /ti / syllables the leftmost small peak

before the 0.2 ms boundary approximates slowly toward the high peak on its right in generation 4 and 5, then disappears in generation 6, resurfaces slightly in generation 7, and ultimately disappears again in generation 8.

Furthermore, the peaks of /ta/ are not as steep as for /ti/. This means that the durations of /ti/ syllables are closer together and vary less than for /ta/ syllables. Most interestingly, the peaks of /ti/ syllables appear predominantly between the peaks of /ta/ syllables on the x-axis. This might explain why the KDE plots in figure 3 did not show any stable development of density peaks for chain D: the PCI durations of either /ta/ or /ti/ syllables never assimilated towards each other. These findings would let assume that the disruption of KDE development is not solely due to an increase in phonological information to be memorized by the participants, but possibly also due to the qualitative difference between the two vowels differing in motor-control effort during articulation.

6.6.4. Measure of Entropy

Hypothesis (i-b) refers to a prediction that the speech signals become more ordered throughout generations. 'Ordered' in this sense means that a measure of entropy will show less and less chaotic distribution of PCI duration categories. Calculations for this are done with R and the package called "Entropy" by Jean Hausser and Korbinian Strimmer (2014). This package allows for an estimation of Shannon Entropy (Shannon & Weaver 1949) with a single line of code. Basically, the Shannon Entropy is applied in the field of Information Theory. As such it is not to be confused with thermodynamic entropy.

Shannon Entropy relates to the predictability of information units such as an array of letters in a string. If a string of letters contains only the same letter then the string will have an entropy value of 0. If the sequence of units is less predictable then the value for entropy increases. However, in order to calculate a measure of entropy it is necessary to group the continuous data (milliseconds) into discrete groups, since the Shannon entropy deals with discrete units. Similarly to Ravignani et al.'s study (2016: 6), a kmeans cluster algorithm is applied to group the data into three discrete categories (k = 3). The kmeans

cluster algorithm also maps each tercile to the duration it belongs to. Thus, every duration enters the calculation for the measure of entropy as belonging to one of the three discrete categories, e.g.: 'short', 'medium' and 'long' (Ravignani et al. 2016: 6).

The decision to group the data into three categories rather than, say two or four, is firstly to replicate the statistical analysis of Ravignani et al. as closely as possible. Secondly, a k = 3 grouping approximates to the KDE clusters from figure 3 more closely than other k values. This means that grouping the data set into two or four duration clusters is not as representative as when grouped into three clusters since the members of all groups are the most similar when k = 3. As such, this kmeans clustering algorithm further strengthens the statistical validity of the PCI duration density peaks from figure 3, as well as hypothesis (i-a) for the first two cluster groups of chains A, B and C, as can be seen in the table below:

Kmeans of chain A:	0.19	0.41	1.21
Kmeans of chain B:	0.15	0.33	0.70
Kmeans of chain C:	0.16	0.29	0.36
Kmeans of chain D:	0.17	0.29	0.41

Table 1 Kmeans clusters for PCI durations of all chains

This table clearly shows how the first two kmeans clusters of a particular chain are approximate multiples of each other, be it either 15 times 2 or 20 times 2. Only for chain B do we find the third cluster to be an equal multiple of the second cluster, while chain A develops a third kmeans cluster which is actually a triple of the second cluster. Additionally, it can be mentioned that the third kmeans cluster categories of chain A and B reach higher values than other categories generally due to the occurrence of pauses in the speech signal. This is clearly visible in chain B's KDE plot, where such a third duration category partly emerges during generation 3 and appears to remain stable during generations 6, 7 and 8.

The next procedure was to measure entropy for all chains and see how this measure of entropy would develop throughout generations. Note, that chain D exhibited quite different KDE developments than the other chains. Furthermore, measure of entropy was also calculated with the exclusion of chain D. Both graphs can be observed in figure 11 below.

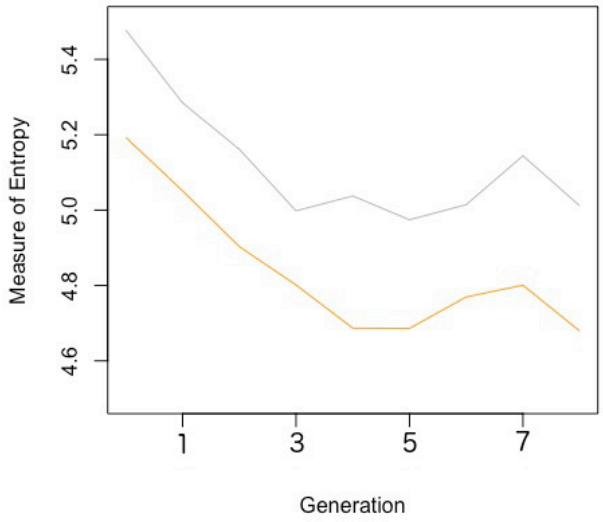

Figure 11 Measure of entropy for all chains (grey) and for all chains except chain D (orange)

Quite strikingly, the measure of entropy appears to be declining throughout generations. Admittedly, there are increases observable as well, however the general trend seems to be strengthening hypothesis (i-b) rather than weakening it. Note that an exclusion of chain D decreases the overall measure of entropy in all generations. To check whether this is indeed caused by chain D's chaotic PCI duration distribution or rather due to the fact that the measurement algorithm is fed with less data observations, further measurements were calculated with the exclusion of other chains.

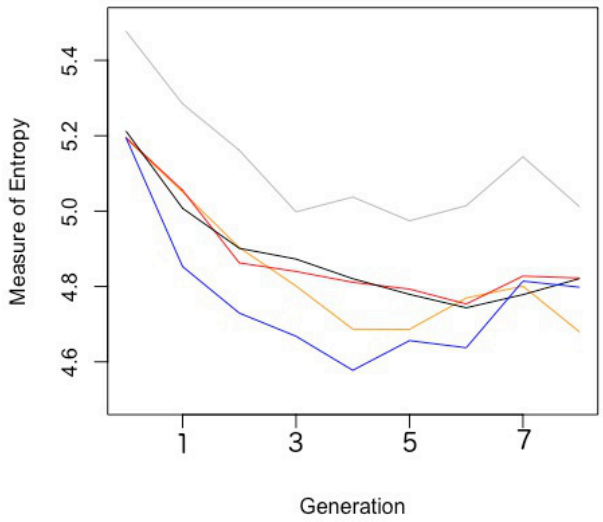

Figure 12 Measure of entropy for all chains (grey), all chains except chain A (black), all chains except chain B (blue), all chains except chain C (red), and all chains except chain D (orange).

What is observable in figure 12 is that the only graph showing a lower measure of entropy throughout all generations is the one excluding chain B. Remember that chain B is the only chain where a third category emerged since generation 6, so an exclusion of chain B's data might result in a decline in measure of entropy for these reasons. Note that an exclusion of chain D shows nonetheless a lesser measure of entropy than depicted by the black and red graphs. Additionally, an exclusion of chain D exhibits the lowest measure of entropy observable in the last generation.

6.6.5. Imitation Inaccuracy

One might argue that hypothesis (ii) is supported by the findings on syllable omission in figure 2, and by the fact that syllable omission decreases in later generations. However, this is a rather poor method and given the information from the KDE plots, a different method would be desirable. What follows are three different models to do this. All of these models are based on calculating the difference between PCI durations, but we want to show that there is more than one option to do so. First, model 1 will show a simplistic and introductory

attempt to corroborate (ii), which we will ultimately drop. Second, model 2 will convey an improved method where we introduce mapping conditions. Third, model 3 will serve as a competing method to model 2, and we will discuss benefits and drawbacks for choosing either of them before we ultimately argue for a synthesis of both to be achieved in future projects. Lastly, note that to corroborate (ii) we want the differences between PCIs to decrease throughout generations in the graphs. Therefore due to convenience, we regard (ii) as a decrease in imitation inaccuracy (II) instead of an increase in imitation accuracy.

6.6.5.1 Model 1

One possible way to account for the decline of II is to calculate the difference between the means of PCI durations of subsequent generations (the square root of {[n] – [n+1]} squared). If the difference equals zero due to the fact that the mean of Gn is identical to the mean of Gn+1, then II amounts to null and we would assume a perfect imitation performance. As such, hypothesis (ii) would be supported if the difference between the means would be declining throughout generations. Figure 13 below shows the II for all chains, where the II is calculated as the difference between the mean of all PCI durations of Gn of all chains and the mean of all PCI durations of Gn+1 of all chains.

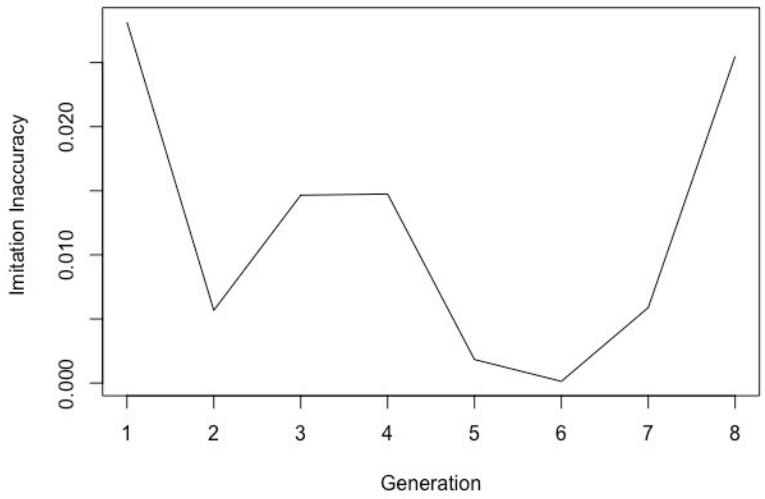

Figure 13 Imitation Inaccuracy of all chains as the difference between means of PCI durations of generation n and generation n+1

At first sight, it would seem that (ii) is not supported by the data. However, given the fact that these calculations do not include information on the omission of syllables, they are not satisfactory to support (ii). What is more, these calculations do not capture the facts that the experiment was split into four separate iteration chains and that participants imitated 6 unrelated syllable strings during the experiment. Thus this model appears rather inaccurate since the difference of means of all strings of all chains does not necessarily map between two strings that are related by virtue of the imitation task. In order to increase this model and make it more exact, a blunt application of the differences of means must be dropped and a number of conditions need to be introduced for sharper calculations. Put simply, instead of first computing the mean and then calculating the difference, it makes more sense to first calculate the difference between two related units. Thus, before we even start to calculate the difference, we first need to decide which two units are related so that we can map them together.

6.6.5.2. Model 2

The three conditions (1), (2) and (3) below serve as a device to map two related PCI durations together for which we can assume that the PCI duration of Gn+1 is a copy of the PCI duration of Gn. These conditions demand that the difference between individual PCI durations is only to be calculated when

(1) the two PCI durations of both Gn and Gn+1 occur in the same chain and syllable string
(2) the two PCI durations occur in similar positions within their strings
(3) the two PCI durations are similar enough in length to assume one is the copy of the other

A fourth condition (4) demands that when syllables are left out, omitted PCIs count as 'empty' PCIs with zero duration. This mapping procedure was done with Excel, and after the differences between mapped units were calculated the mean of the differences for each Gn including all chains was computed. The II calculation gives the following results visible in figure 14 below.

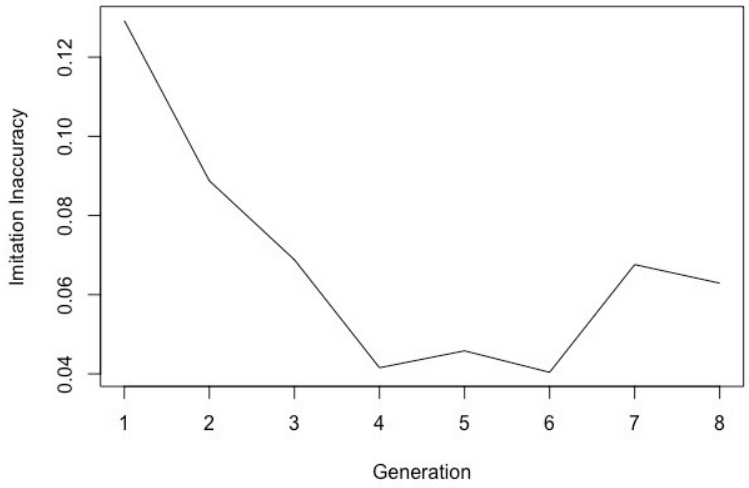

Figure 14 Imitation Inaccuracy of all chains as the mean of differences between mapped PCI durations of generation n and generation n+1, including string correlation and syllable omission.

The decline of II corroborates hypothesis (ii). However, there still are instances where the II increases, which can be seen in generations 4 to 5 and 6 to 7. Be that as it may, the general trend rather conforms to (ii) than disproving it. Most strikingly, the graph in figure 14 closely resembles the grey graph in figure 12 where all chains are included.

At this point it must be noted that conditions (2) and (3) are somewhat troublesome. For one, under very few circumstances can we be sure that participants indeed intended any specific PCI they have produced to be an imitation of any particular PCI they have heard. As such, our model serves only as an approximation of what is assumable to count as a copy given conditions (2) and (3). A further problem is that these two conditions can be prone to error depending on the individual interpretation of the person who does the mapping. This makes the procedure harder to replicate because different interpretations and decisions on what counts as a copy might well lead to different results.

6.6.5.3 Model 3

Since the issue about the replicability of the experiment is indeed a serious one, an alternative should be found. As it stands, condition (1) appears to be a solid contribution to our model and can be kept. However, a different method to calculate the difference between the duration of single PCI units seems necessary. The method applied by Ravignani et al. (2016: 6) is to compute the so-called Levenshtein Edit Distance (LED) which is "the total cost of the minimal cost set of substitutions, insertions or deletions". As such, the difference between PCIs is computed so that the result reaches the lowest outcome. Typically, the LED is computed between strings of symbols such as orthographic words composed by the alphabet. For example, when the LED is to be calculated between the words *rim* and *room*, the <i> is exchanged for an <o> and another <o> is inserted. This amounts to a LED of 2 because every operation (substitution, insertion, deletion) has a cost of 1 (Levenshtein 1966). The minimal cost to edit *rim* to *room* given these three possible operations is thereby 2.

Of course, since milliseconds of PCI durations can be considered to be continuous data, the LED needs to be modified so that it can apply to it. Hence, similarly to Ravignani et al.'s procedure (2016: 6) we calculate the absolute difference between PCI durations so that the total sum of differences reaches the lowest possible outcome. We still make sure that our calculations satisfy condition (1) and that the minimal II is to be calculated only between related strings. For this we use R to return only the minimal absolute difference between string S of generation n and string S of generation n+1. The template for the R code is the following:

```
sapply(string S of Gn+1, function(x) min(abs(string S of Gn – x)))
```

This code repeatedly calculates the absolute difference between PCI durations until a minimal outcome is reached. The results for the sum of differences of all chains are visible in the graph in figure 15 below.

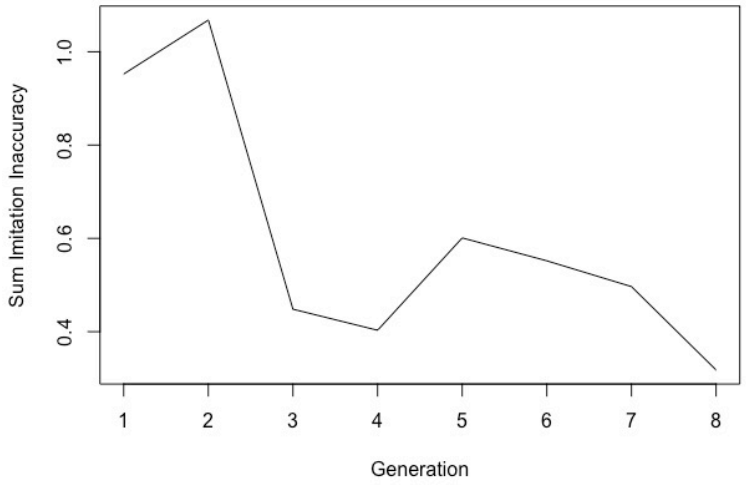

Figure 15 Imitation Inaccuracy of all chains as the minimal sum of differences between PCI durations of generation n and generation n+1

Figure 15 shows how II drastically drops from the second to the third and fourth generation. Afterwards, it slowly rises but continues to decrease throughout the last generations. As such, hypothesis (ii) is corroborated and replicability is ensured.

It might also be interesting to look at the II development for each chain individually.

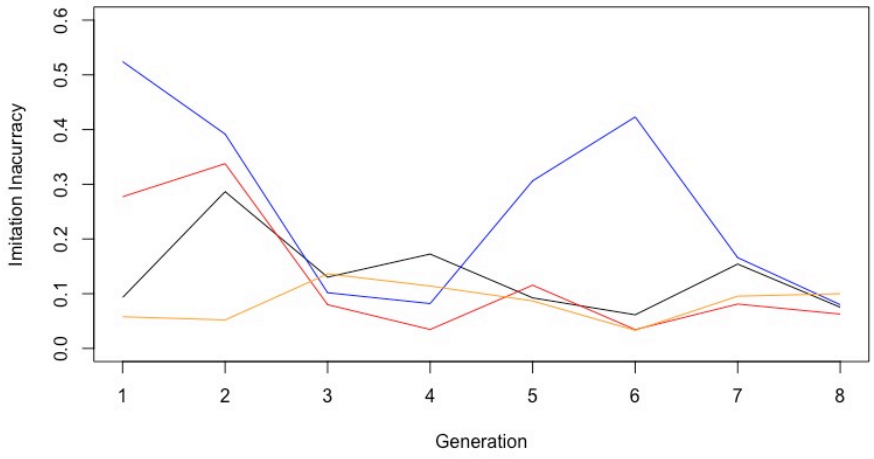

Figure 16 Imitation Inaccuracy of chains A (black), B (blue), C (red) and D (orange) as the minimal mean of differences

Figure 16 shows the development of II for each chain. The most inaccurate performances with regard to imitation appear to be in chain B, while chain D seems to contain the most accurate imitation performances. However, all chains reach a similar II towards the last two generations. Additionally, all chains contain some amount of fluctuation throughout generations with the exception of chain D having the least amount of fluctuation.

Albeit this model corroborates hypothesis (ii) while simultaneously ensuring replicability, it is still worth discussing to what extent the Levenstein Edit Distance is a sound method to calculate II. Firstly, the LED is typically used on categorical data properties, whereas the data properties here are continuous. For example, if the LED between the discrete observations of strings {1 2 3 4} and {1 1 2 2} are calculated, then the minimal edit distance is 3. But if these two strings are regarded as continuous observations, then the minimal edit distance is 4. The crucial difference here is that while any operation has an edit cost of 1 if the strings are discrete, this is not the case if strings contain observations that are continuous between which the actual difference is calculated.

Secondly, the LED as applied by Ravignani et al. (2016: 6) seems to strictly deal with paradigmatic values of a string, e.g. interval duration in milliseconds, and does not take into account syntagmatic aspects, such as the sequencing and order of interval durations in a particular string. As such, the calculations could be sometimes prone to represent the imitation process of participants incorrectly. For instance, since the LED demands the minimal distance between interval durations of Gn and Gn+1, it is possible that an LED is computed between two syntagmatically unrelated intervals, e.g. the first interval of Gn and the last interval of Gn+1. In this case the paradigmatically computed LED model would wrongly represent the imitation process, since it is unlikely that a participants intention to copy the first PCI in a string would result in placing it at the ultimate string position during production.

Although it is beyond the scope of this thesis, a possible solution could be the following. Given that the weakness of model 2 is the low replicability rate due to

individual interpretation of mapping; and the weakness of model 3 is the strict application of minimal LED calculation, a better model should try to mitigate the weaknesses of model 2 and 3. Such a thing could be well achievable if model 2 was fully automatized and would not yield to idiosyncratic variation of interpretation during mapping. As such, the conditions brought forward would inform and dictate the mapping choices of a program and not of a person, which in turn would increase replicability.

6.7. Feet Intervals

One may want to ask whether similar developments can be observed on a higher constituent level; that is: feet. A prosodic foot can be determined as consisting of at least one Designated Terminal Element (DTE) (Giegrich 1985: 8). The DTE is typically a phonetically prominent unit such as a stressed syllable, and recent linguistic theory highlights the fact that prominence of the DTE is a factor relative to the prominence of its phonetic environment. As such, the DTE is always in a prominence relation with neighboring constituents. Furthermore, it seems that Giegrich's theory of feet is in accord with our theory of PCs, since he defines a foot as "a unit that stretches from the onset of a stressed syllable to the next such onset." (1985: 268)

More importantly, if the symbol {S} stands for 'strong prominence' or stressed syllable, and {w} for 'weak prominence' or unstressed syllable, and {S} and {w} always represent a prominence relation to each other, then a foot is always a configuration {Sw} or {Sww} (Giegrich 1985: 15). Firstly, note that this model does not permit {wS} structures. Rather, unstressed syllables are connected as prosodic enclitics to precedent {S} nodes, even though they might belong lexically to an iambic word (Giegrich 1985: 15, 271, 272). Secondly, trisyllabic {Sww} configurations are actually hierarchically structured as {{Sw}w}. This means that the {Sw} configuration acts in this case as a single unit which dominates the ultimate {w} syllable. Thirdly, monosyllabic feet count as {Sw} configurations as well, where the {w} unit is typically a notation of either a lengthening of {S} or a pause. This syllabically empty {w} slot can then be filled by weak syllables if necessary (Giegrich 1985: 15, 17).

Of course, it is somewhat questionable whether or to what extent this theory is applicable to the data. This is due to the fact that the data does not contain anything which can be called a word, at least not in German. However, the minimal {Sw} model reduces options, and thus minimizes idiosyncratic variation in interpretation: an {SwwSw} string will always be segmented into {Sww}{Sw} and not into {Sw}{wSw}.

6.7.1. Phonetic Parameters for Prominence-Stress

Next it is necessary to include parameters that inform whether a syllable carries stress or not. There exist three factors which serve as informative cues for prominence-stress: loudness, duration and pitch (Couper-Kuhlen 1989: 21). We assume that the hierarchical importance of these is pitch variation first, duration second, and loudness last (Couper-Kuhlen 1989: 24). However, any satisfactory annotation should always consider all three of them together, since there is no consensus on the correspondence between one particular phonetic parameter and stress (Lehiste 1970: 110). The annotation was done in PRAAT, where all three parameters are observable so that judgments whether a syllable is to be tagged as stressed or unstressed are not conditioned primarily by auditory perception. After syllables were annotated for stress prominence, feet were constructed according to the minimal metrical stress theory presented above.

However, one possible alternative should be mentioned: the application of the Iambic-Trochaic Law (ITL) to the feet construction process (Hyde 2011: 1052, 1054). The strong interpretation of the ITL suggests that if the prominence relation of units such as syllables is realized by loudness, then these units group into trochaic {wS} configurations. If, on the other hand the prominence relation of units such as syllables is realized by duration, then these units group into iambic {Sw} configurations. As such the ITL could inform and dictate the feet construction process by allowing for both {Sw} as well as {wS} structures, depending on the type of stress cue informing the prominence relation of a specific syllable group. However, the ITL itself is controversial (2011: 1074,

1075) and its strong interpretation is deemed as “unsustainable” (2011: 1055). Hence we will maintain our minimal feet construction system at present.

6.7.2. Results

First we decided to permit only {Sw} or {Sww} structures, and then we discussed when a syllable can be annotated as {S} or {w}. Now after annotating the data for feet-structure we repeat the statistic procedure regarding the KDE. Note, that annotation of Feet Intervals (FIs) for all G0 are more or less dubious, since the voice synthesizer did not include loudness and pitch variation as additional stress parameters besides duration. The annotation of these signals functions as an assumption on how feet structure could have been perceived only by virtue of durational differences.

Similar to the KDE plots for the PCI durations, the following graphs in figure 17 represent the KDEs for the FI durations.

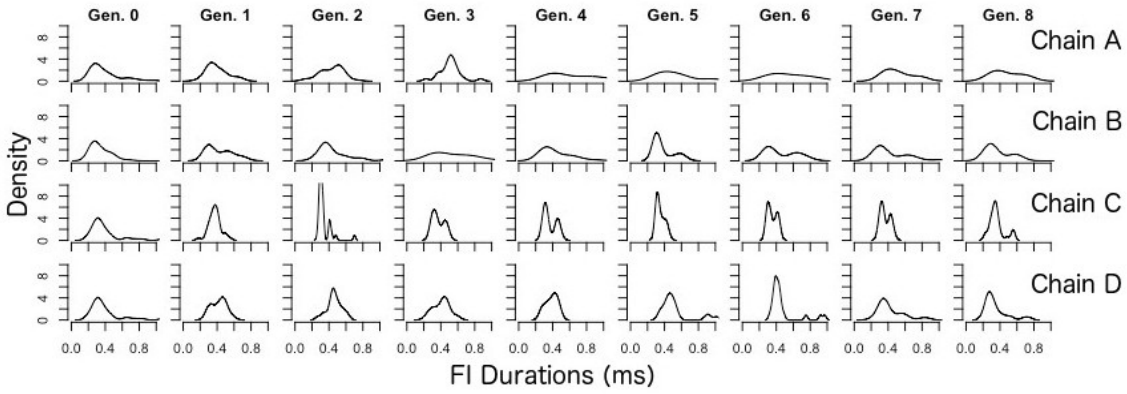

Figure 17 KDE plots showing the Density developments of FI durations throughout generations 0 to 8 for chains A, B, C and D.

Apparently, there does not seem to be a clear emergence of categories for FI durations as for PCI durations. In this regard, only chain B appears to have a stable development throughout generations and high density peaks are found in chain D, but even more so in chain C. However, all chains seem to share one FI duration category that lies approximately between 0.3 and 0.4. This is also confirmed by the kmeans clusters:

Kmeans of chain A:	0.34	0.61	1.09
Kmeans of chain B:	0.32	0.62	1.26
Kmeans of chain C:	0.31	0.44	0.89
Kmeans of chain D:	0.32	0.48	0.92

Table 2 Kmeans clusters for FI durations of all chains

What is interesting is that we do not find this trend for the PCIs in table 1. It is very likely that this is due to the fact that German is a stress-timed language, and that isochronic effects in the prosody of such languages apply generally to higher-order constituents such as feet instead of syllables. Further, these 0.30 FI durations consist of monosyllabic as well as multisyllabic feet, which would suggest that participants transformed the syllable durations and configurations as to conform to isochronic FI durations.

Lastly, we take a look at the development of FI durations with regard to measurement of entropy. This is shown by the graphs in figure 18 below.

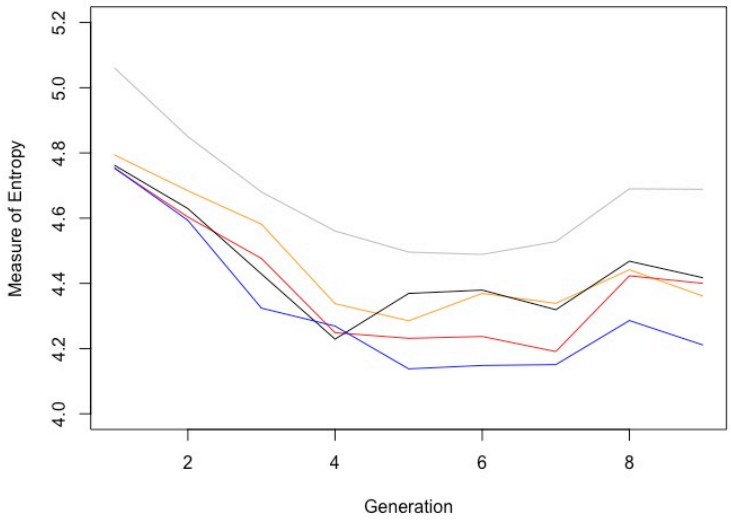

Figure 18 Measure of entropy for the FI durations in all chains (grey), all chains except chain A (black), all chains except chain B (blue), all chains except chain C (red), and all chains except chain D (orange).

The development of the measure of entropy for the FI durations is similar to the measure of entropy development for PCI durations. Typically, the measure of entropy is decreasing throughout generations. Interestingly, all graphs show a rise from antepenultimate to penultimate generation, followed by a final decline from penultimate to ultimate generation. Most prominently, the lowest measure of entropy is reached when chain B is excluded from the calculation.

7. Discussion

In the following we want to discuss our results in the light of the theories we have presented. First we discuss the position of our results inside our language theory. Then we discuss our results in relation to the evolution of language. In the end, we argue that the processing biases we deem responsible for the emergence of rhythmic structure may assumingly be regarded as mental remnants or fossils of protolanguage. This means we suggest that this cognitive trait to process phonetic information in this way is phylogenetically old; probably older than semantic and syntactic components of the linguistic system.

To start, we want to position participants' linguistic behavior inside our language theory. In the experiment, participants simply followed the task instructions and imitated speech strings they had heard. However, imitation was never perfect since the task was difficult and cognitively demanding. Nonetheless, the resulting diffusion of behavior was never random but became increasingly structured and easier to reproduce. It is very likely that these developments are due to cognitive processing biases that relate to categorization functions of language and cognition (Ravignani et al. 2016: 3,4). As such, we find the emergence of structural categories inside the prosodic layer of phonology. These structural categories become apparent in the durations of syllables, where all participants seem to prefer that syllable durations be relatable to a specific duration type. It seems that categorization of duration types into either short or long syllables meet cognitive biases that support perception, memorization and reproduction.

Furthermore, adding vowel variation seems to interfere with this development since PCI durations of /ta/ and /ti/ syllables never approximate towards each

other in chain D. A first assumption is that this is due to a higher amount of information to be memorized during imitation. A second more precise explanation regards the articulatory features of the phonemes in question. Since /a/ demands jaw lowering and requires farther tongue movement from /t/ to /a/ than /t/ to /i/, we predicted that /ta/ syllables could either undergo shortening and schwaization, or be kept long and undemoted due to these special jaw and tongue movement conditions. Our results show that both processes occur and that these processes are not observable for /ti/. However, note that presently no statistical distribution of schwa cases are given in the data, although instances of schwa were found during prominence stress analysis. Further research would be necessary regarding this issue, especially since "syllables in German cannot end in a lax vowel" with the exception of schwa (Wiese 1986: 701). Thus, a possible research question would be the distribution of tense [a], lax [ʌ] and [ə] occurrences for /ta/ syllables in our data. Note that this would not need to be limited to vowel duration but could include formant analysis in addition. One could also enquire whether the vowel in /ti/ realizes either as tense [iː] or lax [ɪ] as well, though the KDE plots show that this syllable type did not develop two duration categories as did /ta/ type syllables.

Analyzing the development inside higher-order constituents shows that syllable durations concatenate into feet with predominantly isochronic durations. It appears somewhat plausible that this might be due to the fact that all participants were native speakers of German; a stress-timed language where syllables tend to be compressed or stretched so as to compose into isochronic feet durations. However, this could also be due to the fact that during their lifetimes participants were exposed to music that generally conveys isochronic time patterns as well. We can propose two arguments favoring the linguistic explanation. Firstly, no musicians were allowed in the experiment, and we might assume that non-musicians would perform in less isochronic fashion than participants who were trained to hold a steady beat. Secondly, in contrast to the FI durations, the PCI durations did not convey this isochronic structure but showed more variation in this regard; a predictable asymmetry if we assume interferences resulting from Germanic prosody. Both these arguments could be

empirically tested in future studies. For the first, the experiment could be repeated with only trained musicians as participants. For the second, the experiment could be repeated with only speakers of a syllable-timed language, such as Italian or Spanish.

To some extent the experiment is an example that the conceptions of nature and nurture need not be two opposing notions but are rather interrelated and interdependent. For one, we can observe that cultural transmission in the form of replication and imitation created specific structures that corroborated our hypotheses (i) and (ii). For another, the emergent structures appear to be constrained by natural cognitive processing biases. We might assume that whatever is responsible for these biases could be found in FLB, more precisely as a component of SM. As such, they could explain why certain structures emerge while others do not. For example, one emergent structural regularity was that syllables were composed into groups, either in the form of isochronous feet or prominently recurring PCI duration type configurations such as {long long short short}. It seems that grouping units together into hierarchically higher units supports memorization, also referred to as 'chunking' where the act of remembering and accessing a string of elements is made easier since the whole string of elements is being reinterpreted as a single element (Fitch 2010: 100). Note however that our experiment merely showed that such biases exist, not whether they originate in domain-general or domain-specific modules of the mind; if the mind is to be conceived as a multi-modular entity in the first place.

Lastly, we want to discuss to what extent our results have implications for the evolution of language. Since our experiment represents a form of micro-cultural evolution, it makes sense to start with glossogeny. For one, our results show how micro-cultural change in a lab setting is influenced by psychological biases on processing, which may also be present in the actual glossogenetic development of historical languages. Indeed, Fitch notes that the stability of hierarchical structures may have some important implications for this form of evolution (2010: 101).

Furthermore, these mechanics that are responsible for the emergence of such structures may be additionally responsible for their replicating success in Ritt's sense (2004). If we take syllables to be memetic replicators or complexes of memetic replicators, then it follows that the structures they convey will yield to cognitive selection pressures. Following Ritt's interpretation, the replication process during imitation would function so that firstly a syllable string is perceived as a behavioral expression. Then this behavioral expression is being checked by SM and its competence constituents for aligning syllables on a metrical grid. Here we could assume that there exist parametric variations depending on the language and that our data conveys effects specific to Germanic constraint rules that influence the metrical configuration of syllable strings on the prosodic layer. Initially in the diffusion chains, participants in early generations perceived behavioral expressions that did not conform to their competence constituents. This resulted in a cognitive reinterpretation during prosodic processing, destabilizing the replication of the behavioral expressions. This led additionally to a transformation of the behavioral expression's structure that increasingly corresponded with participants' competence constituents in SM so that stable structural regularities emerged in later generations that were easier to imitate.

It is not clear what implications our results have with regard to ontogeny. It is rather that findings from early language acquisition may let us assume that the cognitive constraints on prosody and sensitivity towards rhythmic regularities we observed in the results are attributable to very early stages in ontogenetic development. An example we mentioned was the seeming primacy of syllables during the babbling stage of infants. Interestingly, more recent studies applying fetal biomagnetometry suggest that infants use prosodic cues such as rhythm to discriminate between languages that differ in rhythmic organization already before birth (Minai et al. 2017). As such, it seems not only likely that the development of competence constituents relating to rhythmic organization in language must emerge incredibly early, but that sensitivity to such prosodic cues exists already in utero.

Concerning phylogeny, it seems that "complex vocal control and vocal learning via imitation" is indeed a cognitive specialization of humans shared only with few other species such as whales and birds (Fitch 2010: 467; Fitch 2005: 200, 201). Fitch's Phonological Continuity Hypothesis states that phonology must have evolved prior to syntax when compared with regard to computational complexity, and that this less complex system can be found in varying degrees in other species' communication systems (Fitch 2018: 70, 71). In the case of singing birds it is not only that birdsong resembles hierarchical structures similar to human prosody, but that the acquisition of a song shares similarities to language acquisition: both nature and nurture factors are necessary for development, and the imperfect sub-song of young birds resembles the babbling stage of human infants (Hauser et al. 2002: 1572). It appears that certain cognitive traits for prosodic processing that prefer rhythmic organization are shared among vertebrates. One might suggest that this connection possibly represents a case of convergent evolution, but Fitch warns to not make any premature propositions "without a better understanding of such processing in more basal vertebrates" (Fitch 2018: 72).

Our experiment presented which mechanics underlie imitation processes for humans and how they may be responsible for the emergence of structure. Since the diffusion of behavior was influenced by memory limitations, we could regard them as forms of disruption. In Jackendoff's sense (2002), it would follow that the cognitive biases surfacing during these disruptions could constitute cognitive remnants that were present in PL. More specifically, these cognitive remnants may indeed be cases that represent fossils of proto-prosodic processing, especially of a prosodic PL in Fitch's sense (2010). Given that prosodic competences develop incredibly early in human ontogeny seems to further support these assumptions. Additionally, the discovery of mirror-neurons allows for natural explanations of infants' imitation behavior in the human species based upon neurophysiological mechanisms (Rizzolatti et al. 2001). What is more, Fitch argues that sexual selection and kin selection appear to be plausible forces for explaining how these traits of prosodic PL could have developed in phylogeny (2010: 490, 491).

Finally, it should be mentioned that our results do not necessarily prefer one PL model to the other since both models require a phonological system to function. It rather seems that both lexical and prosodic PL models could find synthesis if one takes Miyagawa's Integration Hypothesis (IH) seriously (2017). The IH suggests that language ultimately derived from two types of progenitor communication systems: lexical and expressive. The former might have originated from primate alarm calls and resembles the semantic layer of language, while the latter finds its similarities in birdsong and is responsible for the emergence of phonology and syntax. Miyagawa's main claim is that both these systems evolved in parallel gradually through phylogeny, and that the integration of these two systems led to the abrupt emergence of ML as we know it today.

Summary, Conclusion & Outlook

To summarize, we stated at the beginning of the thesis that the major issues and problems in evolution of language science are those concerning language theory and evidence. We addressed these issues by firstly attempting to define the language system with its functional and formal properties. Then we discussed to what degree this system derives from natural and cultural processes, specifically in the light of the POS and UG. Secondly, we briefly surveyed crucial aspects of recent evolution theory. Thirdly, we merged together the notions of language and evolution, and presented a systematic methodology for language evolution research. Fourthly, we offered possible scenarios of language evolution that make use of the concept of protolanguage. Fifthly, we reported on recent empirical language evolution studies that apply the concept of iterated learning to psychological experimentations known as diffusion chains. Sixthly, we replicated and readjusted such an experiment and intended to show the existence of cognitive processing biases during a vocal imitation task. Finally, we discussed our results and argued that these processing biases surfacing during memory limitation disruptions could assumingly constitute cognitive remnants or fossils of protolanguage.

To conclude, there are number of things we have learned throughout this thesis. For one, the human linguistic system is more complex than it appears on surface and the origins of language are not easily discovered. Nonetheless, the science of language evolution must be able to explain how this complex cognitive specialization developed in phylogeny as well as how it is developing in ontogeny and glossogeny. Modern language evolution research can apply PL models as hypothetical stages in the phylogeny of linguistic systems, and there seem to be instances where remnants of PL surface in the structure of ML. What is more, recent empirical methods to test hypotheses on the emergence of structure are known as diffusion chain experiments based on IL. Our experiment has shown that human perception and memorization of phonetic input is biased, namely that rhythmically organized prosodic structures emerge in the data and are preferred during imitation. We had good reasons to assume that this processing preference is a candidate fossil of PL. First, the preferred structures emerged due to disruptions caused by memory limitations and remained stable during micro-glossogenetic transmission. Secondly, sensitivity for prosodic structure develops very early in human ontogeny, and imitative behavior during early developmental stages can be related to the existence of mirror neurons. Thirdly, this cognitive trait is shared to some extent with other species and appears to be a plausible evolutionary result of sexual and kin selection.

It seems the evolution of language will have a bright outlook as a scientific field of study in the future. This is not only due to a better understanding of linguistic systems and a growth of evidence but also due to an increase in interdisciplinarity. Experimental research on IL via diffusion chains enjoys increasing attention, and we proposed that especially our method allows for a systematic advancement in stimulus and experiment design for future projects. Since empirical evidence that supports any hypotheses on the phylogeny of language evolution are still hard to come by, it is even more important that we are making progress with regard to ontogeny and glossogeny of language evolution.

References

Bartlett, Frederic C. 1932. *On Remembering*. Oxford: Macmillan.

Berwick, C. Robert; Pietroski, Paul; Yankama, Beracah; Chomsky, Noam. 2011. "Poverty of the stimulus revisited". *Cognitive Science* 35, 1207-1242.

Bickerton, Derek. 1990. *Language and Species*. Chicago: University of Chicago Press.

Bickerton, Derek. 2007. "Language evolution: a brief guide for linguists". *Lingua* 117, 510-526.

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

Boersma, Paul; Weenik, David. 2018. *Praat: doing phonetics by computer*. Version 6.0.39, retrieved 3 April 2018 from http://www.praat.org/

Bolhuis, Johan; Tattersal, Ian; Chomsky, Noam; Berwick, Robert. 2014. "How could language have evolved?". *PLoS Biol* 12(8), 1-6.

Brighton, Henry; Smith, Kenny; Kirby, Simon. 2005. "Language as an evolutionary system". *Physics of Life Reviews* 2, 177-226.

Carstairs-McCarthy, Andrew. 1999. *The Origins of Complex Language*. Oxford: University Press.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Chomsky, Noam. 1980. "On binding". *Linguistic Inquiry* 11(1), 1-46.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, Noam. 1986. *Knowledge of language: its nature, origin and use*. New York: Praeger.

Chomsky, Noam. 1988. *Language and Problems of Knowledge*: *The Managua Lectures.* Cambridge, MA: MIT Press.

Chomsky, Noam. 1995. *The Minimalist Program.* Cambridge, MA: MIT Press.

Chomsky, Noam. 2002. *On Nature and Language.* Cambridge, MA: MIT Press.

Chomsky, Noam. 2014. "Minimal recursion: exploring the prospects". In Roeper, Tom; Speas Margaret (eds.). *Studies in Theoretical Psycholinguistics 43. Recursion: Complexity in Cognition.* Cham: Springer, 1-15.

Couper-Kuhlen, Elizabeth. 1986. *An Introduction to English Prosody.* Tübingen: Max Niemeyer Verlag.

Couper-Kuhlen, Elizabeth. 1993. *English Speech Rhythm.* Amsterdam: John Benjamins.

Darwin, Charles. 1859. *On the Origin of Species.* London: John Murray.

Darwin, Charles. 1871. *The Descent of Man, and Selection in Relation to Sex*. London: John Murray.

Dawkins, Richard. 1976. *The Selfish Gene.* Oxford: University Press.

Dutoit, T.; Pagel, V.; Pierret, N.; Bataille, F.; Vreken, O. van der. 1996. "The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes". *Proc. ICSLP'96 Philadelphia* 3, 1393-1396.

Evans, N.; Levinson, S. 2009. "The myth of language universals: language diversity and its importance for cognitive science". *Behavioral and Brain Sciences* 32, 429-492.

Evans, V. 2016. "What do Brexit and Universal Grammar have in common?".

Psychology Today. 14 September 2016. https://www.psychologytoday.com/blog/language-in-the-mind/201609/what-do-brexit-and-universal-grammar-have-in-common (8. January 2017)
Fitch, Tecumseh. 2005. "The evolution of language: a comparative review". *Biology and Philosophy* 20, 193-230.
Fitch, Tecumseh. 2010. *The Evolution of Language.* Cambridge: University Press.
Fitch, Tecumseh; de Boer, Bart; Mathur, Neil; Ghazanfar, Asif. 2016. "Monkey vocal-tracts are speech ready". *Science Advances* 2, 1-7.
Fitch, Tecumseh. 2018. "What animals can teach us about human language: the phonological continuity hypothesis." *Current Opinion in Behavioral Sciences* 21, 68-75.
Fujita, Koji. 2014 "Recursive merge and human language evolution". In Roeper, Tom; Speas Margaret (eds.). *Studies in Theoretical Psycholinguistics 43. Recursion: Complexity in Cognition.* Cham: Springer, 243-264.
Goldsmith, John A. 1979. *Autosegmental Phonology.* New York: Garland Press.
Goldberg, Adele E. 2003. "Constructions: a new theoretical approach to language". *TRENDS in Cognitive Sciences* 7(5), 219-224.
Griegrich, Heinz J. 1985. *Metrical Phonology and Phonological Structure.* Cambridge: University Press.
Han, Chung-hye; Musolino, Julien; Lidz, Jeffrey. 2016. "Endogenous sources of variation in language acquisition". *PNAS* 113(4), 942-947.
Hauser, M. D.; Chomsky, N.; Fitch, W. T. 2002. "The faculty of language: what is it, who has it, and how did it evolve?". *Science* 298, 1569-1579.
Hausser, Jean; Strimmer, Korbinian. 2014. "Estimation of entropy, mutual information and related quantities". Version 1.2.1. https://cran.r-project.org/web/packages/entropy/entropy.pdf.
Hengeveld, Kees; Mackenzie, Lachlan. 2008. *Functional Discourse Grammar: A typologically-based theory of language structure.* Oxford: University Press.
Hockett, Charles F. 1960. "The origin of speech". *Scientific American* 203, 88-111.
Horner, Victoria; Whiten, Andrew; Flynn, Emma; de Waal, Frans B. M. 2006. "Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children". *PNAS* 103(37), 13878-13883.
Hornstein, Norbert; Nunes, Jairo; Grohmann, Kleanthes. 2005. *Understanding Minimalism*. Cambridge: University Press.
Hurford, J. R. 1990. "Nativist and functional explanations in language acquisition". In Roca, I.M. (ed.). *Logical issues in language acquisition*. Dordrecht: Foris, 85-136.
Hyde, Brett. 2011. "The iambic-trochaic law". In van Oostendorp, Marc; Ewen, Colin; Hume, Beth; Rice, Keren (eds.). *A Companion to Phonology*. Oxford: Blackwell, 1052-1077.
Ibbotson, Paul; Tomasello, Michael. 2016a. "Evidence rebuts Noam Chomsky's theory of language learning". *Scientific American*. 7 September 2016. https://www.scientificamerican.com/article/evidence-rebuts-chomsky-s-theory-of-language-learning/ (8. January 2017).
Ibbotson, Paul; Tomasello, Michael. 2016b. "Evidence rebuts Noam Chomsky's theory of language learning". (from *Scientific American*) *Salon.* 11 September 2016. http://www.salon.com/2016/09/10/what-will-

universal-grammar-evidence-rebuts-chomskys-theory-of-language-learning_partner/ (8. January)
Jackendoff, Ray. 2002. *Foundations of Language*. Oxford: University Press.
Kirby, Simon; Cornish, Hannah; Smith, Kenny. 2008. "Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language". *PNAS* 105(31), 10681-10686.
Kirby, Simon; Griffiths, Tom; Smith, Kenny. 2014. "Iterated learning and the evolution of language". *Current Opinion in Neurobiology* 28, 108-114.
Knowlton, B. J.; Squire, L. R. 1994. "The information acquired during artificial grammar learning". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(1), 79-91.
Ko, Kwan Hyung. 2016. "Origins of human intelligence: the chain of tool-making and brain evolution". *Anthropological Notebooks* 22(1), 5-22.
Kousta, Stavroula-Thaleia. 2009. "Innateness". In Chapman, Siobhan; Routledge, Christopher (eds.). *Key Ideas in Linguistics and the Philosophy of Language*. Edinburgh: University Press, 96-97.
Labov, William. 1972. *Sociolinguistic Patterns.* Pennsylvania: University Press.
Larsen-Freeman, Diane; Cameron, Lynne. 2008. *Complex Systems and Applied Linguistics.* Oxford: University Press.
Laurence, Stephen; Margolis, Stephen. 2001. "The poverty of the stimulus argument". *Brit. J. Phil. Sci.* 52, 217-276.
Lehiste, Ilse. 1970. *Suprasegmentals.* Cambridge, MA: MIT Press.
Levenshtein, V. I. 1966. "Binary codes capable of correcting deletions, insertions, and reversals". *Sov. Physi. Doklady* 10, 707-710.
Lidz, Jeffrey; Han, Chung-hye; Musolino, Julien. 2016. "Endogenous content". *PNAS* 113(20), E2765.
Lidz, Jeffrey. 2016. "Chomsky's theory of language learning dead? Not so fast…". *Scientific American.* 25 November 2016. https://blogs.scientificamerican.com/guest-blog/chomsky-s-theory-of-language-learning-dead-not-so-fast/ (8. January 2017)
Liberman, Mark; Prince, Alan. 1977. "On Stress and linguistic rhythm". *Linguistic Inquiry* 8, 249-336.
Liljencrants, J.; Lindblom, B. 1972. "Numerical simulations of vowel quality systems. *Language* 48, 839-862.
Lisker, L.; Abramson, A.S. 1964. "A cross-language study of voicing in initial stops: acoustical measurements". *Word* 20, 384-422.
Milanov, R.; Huotilainen, M.; Välimäki, V.; Esquef, P. A.; Tervaniemi, M. 2008. "Musical aptitude and second language pronunciation skills in school-aged children: neural and behavioral evidence". *Brain Research* 1194, 81-89.
Minai, Utako; Gustafson, Kathleen; Fiorentino, Robert; Jongman, Allard; Sereno, Joan. 2017. "Fetal rhythm-based language discrimination: a biomagnetometry study". *NeuroReport* 28(10), 561-564.
Miyagawa, Shigeru. 2017. "Integration hypothesis: a parallel model of language development in evolution". In Watanabe, Shigeru; Hofman, Michel A.; Shimizu, Toru (eds.). *Evolution of the Brain, Cognition, and Emotion in Vertebrates.* Cham: Springer, 225-247.
Nudel, Ron; Newbury, Dianne F. 2013. "FOXP2". *WIREs Cogn Sci* 4, 547-560.
Perfors, Amy; Tenenbaum, B. Joshua; Regier, Terry. 2006. "Poverty of the

stimulus? A rational approach". *Proceedings 28th Annual meeting of the Cognitive Science Society*. Vancouver, Canada.
Piantadosi, Steven T.; Kidd, Celeste. 2016. "Endogenous or exogenous? The data don't say". *PNAS* 113(20), E2764.
Pinker, S.; Bloom, P. 1990. "Natural language and natural selection". *Behavioral and Brain Sciences* 13(4), 707-784.
Prince, Alan; Smolensky, Paul. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell.
Pullum, G.; Scholz, B. 2002. "Empirical assessment of stimulus poverty arguments". *The Linguistic Review* 19, 9-50.
Ravignani, Andrea; Delgado, Tania, Kirby, Simon. 2016. "Musical evolution in the lab exhibits rhythmic universals". *Nature Human Behavior* 1, 1-6.
Rosch, Eleanor. 1978. "Principles of categorization". In Rosch, Eleanor; Lloyd, Barbara B. (eds.). *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum, 27-48.
Ritt, Nikolaus. 2004. *Selfish Sounds and Linguistic Evolution: A Darwinian Approach to Language Change.* Cambridge: University Press.
Rizzi, Luigi. 2016. "The concept of explanatory adequacy". In Roberts, Ian (ed.). *The Oxford Handbook of Universal Grammar*. New York: Oxford University Press.
Rizzolatti, R.; Fogassi, L.; Gallese, V. 2001. "Neurophysiological mechanisms underlying the understanding and imitation of action". *Nature Reviews Neuroscience* 2, 661-670.
RStudio Team. 2015. *RStudio: Integrated Development for R.* RStudio Inc., Boston, MA. http://www.rstudio.com/
Saussure, F. de. 1916. *Course in General Linguistics.* New York, NY: McGraw-Hill.
Schreuder, Maartje J.; Gilbers, D.; Quené, H. 2009. "Recursion in phonology". *Lingua* 119, 1243-1252.
Shannon, C.; Weaver, W. 1949. *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.
Taylor, R. John. 1989 [2009]. *Linguistic Categorization*. Oxford: University Press.
Thelen, E. 1991. "Motor aspects of emergent speech: a dynamic approach". In Krasnegor, N. A.; Rumbaugh, D. M.; Schiefelbusch, R. L.; Studdert-Kennedy, M. (eds.). *Biological and Behavioral Determinants of Language Development.* New York: Lawrence Erlbaum, 339-362.
Torres, Hector. 2009 3rd May. "The 100 most influential books ever written". *Goodreads*.https://www.goodreads.com/list/show/36714.The_100_Most_Influential_Books_Ever_Written (1st March 2018).
Tinbergen, N. 1963. "On aims and methods of ethology". *Zeitschrift für Tierpsychologie* 20, 410-433.
Ullman, Michael T.; Corkin, Suzanne; Coppola, Marie; Hickok, Gregory; Growdon, John H.; Koroshetz, Walter J; Pinker, Steven; 1997. "A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system". *Journal of Cognitive Neuroscience* 9(2), 266-276.
Wiese, Richard. 1986. "Schwa and the structure of words in German". *Linguistics* 24, 697-724.
Verhoef, Tessa; Kirby, Simon; de Boer, Bart. 2014. "Emergence of combinatorial

structure and economy through iterated learning with continuous acoustic signals". *Journal of Phonetics* 43, 57-68.

Appendix

English Abstract

The aim of the thesis is to show that there exist cognitive processing biases in the prosodic domain of human phonology, and to propose that these biases might represent fossils of a linguistic system older than modern human language. The existence of these processing biases is exhibited in a phonetic diffusion chain experiment. In this experiment we demonstrate that participants are sensitive to structural regularities inside a phonetic stimulus they hear during an iterated learning task. Participants preferred stimuli that were rhythmically organized to those which were not. Preference in this sense means that the former structure type was more stable and easier to learn during transmission than the latter.

The first three sections of the thesis are a literature survey on the relevant methodology. We start by presenting functional and formal properties of language, and discuss their ontological status. After this, we briefly summarize key concepts of modern evolutionary theory. Then we present a systematic approach to language evolution research that distinguishes between three historical layers of explanation.

The last four sections deal with proposals for possible language evolution scenarios and experimentation on iterated learning. Section four presents two models of protolanguage as hypothetical stages of ancestral linguistic systems pre Homo sapiens. It is argued that protolanguages are structurally simpler forms of modern language and that remnants of the former can surface in the latter. Then we showcase and review empirical studies on the emergence of structure through iterated learning in section five. In section six we replicate and readjust their experiment design in our own phonetic diffusion chain experiment. We discuss the implications of our results for the evolution of language in section seven and propose that the cognitive biases responsible for the emergent prosodic structures could be valid candidates for remnants of protolanguage.

Key words: evolution, language, cognition, protolanguage, phonology, prosody, phonetics, rhythm, iterated learning, diffusion chain

German Abstract

Diese Masterarbeit zeigt, dass menschliche Kognition im Bereich der Phonologie bestimmte Präferenzen für prosodische Regularitäten aufweist und dass diese Präferenzen möglicherweise Relikte eines linguistischen Systems sind welches älter als moderne Sprache ist. Wir zeigen die Existenz dieser Präferenzen in einem phonetischen Diffusionskettenexperiment. Während iterativem Lernen bevorzugten TeilnehmerInnen Stimuli die rhythmisch organisiert waren mehr als Stimuli bei denen das nicht der Fall war. Diese Präferenz drückte sich während den Iterationen insofern aus, dass rhythmische Strukturen stabiler und leichter lernbar waren als unrhythmische Strukturen.

Die ersten drei Sektionen dieser Masterarbeit verschaffen einen Überblick über die relevante Methodologie. Zuerst werden die funktionalen und formalen Eigenschaften von Sprache präsentiert, und ihr ontologischer Status wird diskutiert. Danach werden wichtige Schlüsselkonzepte moderner Evolutionstheorie kurz zusammengefasst. Dann präsentieren wir eine systematische Methode zur Sprachevolutionsforschung welche zwischen drei historischen Erklärungsebenen unterscheidet.

Die letzten vier Sektionen behandeln Vorschläge für potentielle Sprachevolutionsszenarien und die experimentelle Forschung von iterativem Lernen. Sektion vier zeigt zwei Modelle von Protosprache als hypothetische Abschnitte von angestammten linguistischen Systemen vor Homo Sapiens. Protosprachen werden als strukturell einfachere Formen von moderner Sprache angesehen, und es wird behauptet dass Relikte von Protosprache in moderner Sprachstruktur sichtbar werden können. Danach begutachten wir empirische Studien über die Entstehung von Struktur durch iteratives Lernen in Sektion Fünf. In Sektion Sechs replizieren und adjustieren wir die Experimentausführung in einem eigenen phonetischen Diffusionskettenexperiment. Wir diskutieren die Implikationen unserer Resultate für Sprachevolutionsforschung in Sektion Sieben und schlagen vor, dass der Teil der Kognition der für die Entstehung der prosodischen Strukturen verantwortlich ist einen gültigen Kandidaten für ein Relikt von Protosprache darstellt.

Stichwörter: Evolution, Sprache, Kognition, Protosprache, Phonologie, Prosodie, Phonetik, Rhythmus, Iteratives Lernen, Diffusionskette

Curriculum Vitae

Personal Information

Name: Duro Rosic, BA
Address: Rokitanskygasse 14/14, Wien 1170
Mobile: +43 0650 3846451
E-mail: rosic.duro@gmail.com
Sex: Male | *Date of Birth:* 20.05.1990 | *Nationality:* Austria

Work Experience

- 2017.9.7. – 8.: Student helper for Diachronic Phonotactics Workshop 2017
- 2017.2. – 5.: Digital humanities researcher internship at ÖAW - ACDH
- 2016.11.21. – 24.: Student helper for ExAPP Conference 2016
- 2015.6. – ongoing: Freelance english trainer at Institut Connect, DELPHIN ILS, educom
- 2015.3. – ongoing: English tutor for the VHS 2.0 education assistance program
- 2015.2. – ongoing: Freelance english trainer at SPIDI
- 2010.7. – 2014.7.: Various customer service jobs
- 2009.1. – 2011.9.: Deputy coach for children's water-polo team
- 2008.11. – 2009.6.: Community service at ÖHTB

Education and Training

- 2017.5.12.: Student award for exceptional academic performance
- 2016.8.29. – 30.: International postgraduate course in Functional Discourse Grammar
- 2015.10 – 2016.10.: BEd English and Psychology/Philosophy
- 2015.2. – 2018: MA English Language and Linguistics at the University of Vienna
- 2014.9.2. – 12.12.: CELTA course at BFI Vienna, awarded with Certificate in English Language Teaching to Adults, Pass (Grade B)
- 2009 – 2014: BA English and American Studies at the University of Vienna, degree: Bachelor of Arts (November 2014)
- 2000 – 2008: Grammar school BRG18 Schopenhauerstraße, Wien 1180 degree: AHS Matura (June 3rd 2008)
- 1996 – 2000: Elementary school VS Schulgasse, Wien 1180