Received: 14 September 2021 | Revised: 13 December 2021 | Accepted: 15 December 2021

DOI: 10.1002/bies.202100216

BioEssays

PROBLEMS & PARADIGMS

Prospects & Overviews

Toward an evolutionary framework for language variation and change

Emmanuel D. Ladoukakis[1] | **Dimitris Michelioudakis**[2] | **Elena Anagnostopoulou**[3]

[1] Department of Biology, University of Crete, Iraklio, Greece

[2] Department of Linguistics, School of Philology, Aristotle University of Thessaloniki, Thessaloniki, Greece

[3] Department of Philology, Division of Linguistics, University of Crete, Rethymnon, Greece

Correspondence
Emmanuel D. Ladoukakis, Department of Biology, University of Crete, 70013 Iraklio, Greece.
Email: ladoukakis@uoc.gr

Abstract

In this paper, we identify the parallels and the differences between language and life as evolvable systems in pursuit of a framework that will investigate language change from the perspective of a general theory of evolution. Despite the consensus that languages change similarly to species, as reflected in the construction of language trees, the field has mainly applied biological techniques to specific problems of historical linguistics and has not systematically engaged in disentangling the basic concepts (population, reproductive unit, inheritance, etc.) and the core processes underlying evolutionary theory, namely mutation, selection, drift, and migration, as applied to language. We develop such a proposal. Treating language as an evolvable system places previous studies in a novel perspective, as it offers an elegant unifying framework that can accommodate current knowledge, utilize the rich theoretical framework of evolutionary biology, and synthesize many independent strands of inquiry, initiating a whole new research program.

KEYWORDS
evolutionary framework, evolvable system, language change, language diversification, language genotype, language phenotype

INTRODUCTION

According to the theory of evolution, life started from a common ancestor and has been diversified into myriads of different forms. Living systems are characterized by certain features that inevitably lead them to evolve. First, they consist of populations of individuals that replicate, resulting in a succession of generations over time. Second, replication entails inheritance of traits, that is, characteristics are transmitted, at least partially, from parents to offspring by replicating and inheriting genetic material (i.e., DNA). Third, replication inevitably creates variation of traits as the probability of mistakes during replication does not equal zero.

Based on this description, it is expected that any system meeting the relevant prerequisites will evolve in a way similar to living organisms. We will call these systems "evolvable systems" in agreement with the most common use of the concept of evolvability, which is defined as the ability of a system to generate heritable phenotypic variation[1] (p. 15). Life is the best-studied evolvable system, though it has recently been shown in Ecology that communities of organisms constitute evolvable systems as well.[2]

Languages, like species, have been diversified in space and time as human populations spread all over the globe. Since the 19th century, historical linguistics has established that systematic similarities between languages often reflect common ancestry and that the longer they have diverged from their most recent ancestor, the more differences we observe between them.[3] For example, it is well-established that French, Italian, and Hindi all have the reconstructed language *Proto-Indo-European* as their common ancestor,[4] but French and Italian have more properties in common than Hindi because they derive from Latin, while Hindi derives from Sanskrit (see ref.[5] among many others).

Since languages and species evolve in a "curiously parallel" way,[6] according to Darwin, it is reasonable to hypothesize that language fulfills the three prerequisites of evolvable systems: (a) populations

of replicating units, (b) inheritance of characters, and (c) variation of characters as a result of imperfect replication. Even though it has sometimes been assumed that language indeed meets these prerequisites,[7–10] a careful overview of the existing literature (see also[11]) reveals the lack of a generally agreed upon framework based on a comprehensive, systematic, and precise assessment of all relevant issues involved. It is therefore not trivial to agree on the properties that make language an evolvable system with its unique profile despite sporadic efforts to develop such proposals in different frameworks.[9,12–14]

The purpose of this paper is to argue that language possesses the three key characteristics of an evolvable system and thus evolutionary theory can also account for language change. We will first introduce some background on language as an object of scientific inquiry and the concepts of grammar, language acquisition, and language change (Box 1). We will then make a proposal regarding the correspondences between the units of evolution, replication, and information, as well as the fundamental processes of evolution, namely mutation, selection, drift, and migration, in language and life (Table 1). Finally, we will address some crucial differences between biological and linguistic evolution and discuss their methodological implications.

BACKGROUND: GRAMMAR AS KNOWLEDGE, LANGUAGE ACQUISITION, AND LANGUAGE CHANGE

Following a major trend in modern linguistic theory we approach language not as a sociopolitical and cultural phenomenon but as a cognitive mechanism (see Box 2). There are two major approaches toward language and language change, usage-based theories versus grammar-/acquisition-based theories.[12] According to usage-based approaches, language is a network of "stored pairings of form and function"[15] and speakers "retrieve expressions from their stored linguistic experience."[16] Language acquisition is a frequency based probabilistic process assisted by non-language-specific cognitive biases.[16,17] According to grammar-based theories, a major object of study of linguistics is the concept of "grammar," that is, the knowledge native speakers have that allows them to generate the well-formed sentences of their language.[18,19] This notion of grammar as a discrete combinatorial generative system should not be confused with descriptive and prescriptive grammars of particular languages as taught in school. The system of knowledge speakers possess is highly abstract consisting of a finite set of lexical elements and rules that may generate an infinite set of novel well-formed sentences and rule out the ill-formed sentences of their native language. We side with the latter view based on arguments presented in Yang[20] (and references therein). He argues on the basis of computational and quantitative studies that there is a categorical distinction between rules and exceptions. Language learning is a search for productive generalizations tolerating a small only amount of exceptions rather than listing everything in lexical storage.

Box 1. Clarifying certain terminological issues in linguistics

In this paper, we have used certain terms in a way that diverges from common uses in linguistics. We provide a list here.

1. There are two concepts of evolution in language. On the one hand, linguists commonly employ the term "language evolution" in order to refer to the process that gave rise to language as a property of humans.[80,81] On the other hand, one can use the term "evolution" in order to describe the change in the properties of language systems diachronically (see e.g., refs. [82,83]). Linguists commonly refer to this second process by the term "language change." In this paper, we use the term "language evolution" only to refer to language change.
2. We have proposed that the counterpart of mutation is innovation. Some linguists have called the question of generation of novel variation in language populations "the actuation problem" (e.g., ref. [84]). In the more standard use of the term however[85] (p. 102) the actuation problem refers to changes in a structural feature in a particular language at a particular time, as opposed to other languages and other times where the same change does not take place. In that more standard use, the actuation problem subsumes all four processes of language evolution.
3. The term "drift" has not been used in linguistics in the sense we describe here, with the notable exception of.[51,52] Since Sapir,[86] drift in historical linguistics has been understood as a cluster of changes moving to a specific direction, the opposite of what we adopt here. In fact, according to Sapir[86] (p. 127) drift has a direction which "may be inferred […] from the past history of the language"; in other words language change produces similar patterns after the split of lineages independently of areal factors and due to common ancestry[87] (p. 1)[13] (p. 9). Yanovich[88] develops a model according to which genetic drift sets the conditions for Sapir's drift to arise, and illustrates this in relation to semantic change in the domain of Germanic modals. Alternatively, the parallelism in drift between descendant languages might in fact reflect the workings of language universals and correspond to homoplasies in biology. This is ultimately a testable question.

On the view adopted here, a language consists of two main components: (i) a mental lexicon containing its abstract features (e.g., tense, grammatical person/number/gender, case) and morphemes (the minimal sound-meaning associations), and (ii) a computational system with the set of rules that combine morphemes to create words

TABLE 1 Analogies of concepts and terms between biology and linguistics (cf.[10] on parallels between biological and cultural evolution)

Biology	Linguistics	Biology definition	Linguistics definition	Unifying concepts
Individual	Language–individual	A single independent organism.	Each speaker level idiolect.	The atom of a population.
Population of individuals	Population of language–individuals	A number of individuals of the same species living in a common space at a common time.	A number of idiolects of the same language spoken by individuals living in a common space at a common time.	The unit of evolution.
Species	Languages	A group of living organisms consisting of individuals capable of potentially exchanging genetic material.	A group of idiolects produced by similar grammars, such that they would potentially provide input to monolingual speakers.	Populations of individuals not anchored to a specific space and time.[a]
Subspecies	Dialects	Distinct subgroups of individuals within a species/language which have sufficient differentiation from other subgroups but not enough to be characterized as distinct species/languages.[b]		
Mutation	Innovation	The processes leading to the introduction of de novo variation in the population.		
Selection	Selection	Non-random increase or decrease of the frequency of the variants in a population.		
Random genetic drift	Drift[c]	Stochastic increase or decrease of the frequency of the variants in a population.		
Gene flow (migration)	Language contact	Exchange of genetic/grammatical information between populations.		
Genotype	I-language/grammar	The total genetic information of an individual.	A system of units of information that generates exactly those combinations of words that form grammatical sentences in a given language.	
Phenotype	E-language/idiolect	The set of observable characteristics of an individual resulting from the interaction of its genotype with the environment.	The set of observable characteristics of an individual resulting from the interaction of its grammar with the environment.	
Gene	Characteristics relativized to the level of analysis (lexicon, phonology, morphology, syntax, semantics)	The unit of information for the construction of living organisms.	The unit of information which qualifies as a primitive at different levels of analysis.	The unit of information.

[a] It should be taken into account that there is a difference in time scales of species as opposed to languages. A dog in ancient Greece would be the same species as a dog today. This does not hold for ancient Greek which cannot be seen as the same language as modern Greek as defined here. If a child had input from a native speaker speaking ancient Greek and a native speaker of modern Greek (s)he would probably not become a monolingual individual.
[b] It is notoriously difficult in linguistics to define the limits between languages and dialects and there are socio-political considerations also playing an important role. We are abstracting away from this here assuming an idealized picture of this division.
[c] See Box 1.

and words to create complex sentences. Words and sentences are pronounced and interpreted by the speakers. The grammatical system consists of three components, each having its own primitive units and rules; phonology (the sound system of language), morpho-syntax (the rules producing words and sentences), and semantics (the assignment of meanings to words and sentences in context). Speakers are not conscious of their knowledge of grammar. Their native language is acquired without explicit teaching, in the critical period, approximately the first 6 years of life,[21] on the basis of positive evidence only (the data they hear, without systematic corrections) and in uniform ways across languages and circumstances[22] (p. 3). If acquiring language was a conscious process, speakers would be able to learn their native language at any age, the way they learn other cultural skills (e.g., chess playing, math, or writing). According to Chomsky, language acquisition is possible because human beings are genetically endowed with a system of linguistic knowledge, called Universal Grammar, which is activated through experience.[18,23,24] This initial state enables acquirers to analyze their input data and make specific choices that

Box 2. Addressing some common misconceptions about language

Theoretical and experimental linguistics has provided extensive evidence against certain common misconceptions concerning language:

1. Language is not simply a list of words.
2. Language is not taught.
3. Language is not learned via imitation. It is acquired by children in the critical period on the basis of input from the parental generation and hence it is vertically transmitted.
4. The enrichment of the vocabulary as well as writing and learning foreign languages, which are straightforward cases of cultural transmission, happen later in life and does not alter the core of one's native language system (the grammar).
5. There is cross-disciplinary evidence that language is a unique and defining capacity of the human brain and not merely a socio-cultural phenomenon.[81,89] One property that has been argued to distinguish language from other communication systems in animals is recursion, that is, the capacity to produce infinite strings via the repeated application of the rule combining lexical elements and constituents.[90] Therefore, it is arguably an oversimplification to think of language change as a typical instance of cultural evolution (pace[10,91]).

Finally, it should be kept in mind that "language acquisition," which is the technical term describing the process underlying vertical transmission in language, should not be confused with the term "acquired characteristics," which is borrowed from biology and describes traits that enter the idiolects later in life and are responsible for horizontal transfers.

lead them to construct their native grammar. The sets of options out of which choices are made have been called "parameters."

Each generation constructs its grammar in an attempt to reconstruct the parental grammar, without having direct access to the grammatical system itself but rather to its surface manifestations in the speech the acquirers are exposed to. Language change is gradual grammar change across generations mediated by language acquisition[25–27] (p. 720 in ref.,[51] for references to specific case studies). It happens when the transmission from the parent generation to the descendant generation is flawed with respect to certain properties.[28–30] According to the children-based approach, variation in the adult speech plays a crucial role since it shapes the input to the children. The alternative usage-based framework according to which linguistic structures evolve solely via language use[12,31] focuses on the first part of this process, while the children-based approach crucially also includes the transition from one grammar to another.

LANGUAGE AS AN EVOLVABLE SYSTEM: CORE CONCEPTS

The first set of correspondences between life and language as evolvable systems concerns the units of evolution, reproduction, replication, and information (Table 1). Evolutionary theory defines evolution as the process of change of heritable characteristics over time. In biology, researchers focus on the frequency change of inherited features over time. Therefore, the unit of evolution is a population of individuals (independent living organisms) belonging to the same species, since frequencies can be only assessed within populations. Extending this to language, we can understand "language populations" as groups of idiolects spoken in the same place at the same time by people having the same native language. An idiolect is an individual's knowledge of language when put to use in actual speech[32] (cf. also Chomsky's distinction between I(nternalized) and E(xternalized)-language, in ref.[19]). The idiolect is the "language-individual" of a population (see Table 1 and elsewhere in the text). Although all monolingual speakers of the same language community have the same native language, each speaker has their own idiolect because of the special circumstances of their upbringing, which are unique to them. Inspired by biology, which defines biological species as a number of individuals having the potential to produce fertile offspring,[33] we propose that two (or more) idiolects are part of the same native language if, by providing input to a descendant at the stage of language acquisition, they have the potential to yield a monolingual and not a bilingual (or multilingual) speaker. In biology, individuals are reproduced via their gametes, what is replicated is the genome and the unit of information is the gene. In language, idiolects are reproduced via the set of utterances that serve as cues/triggers in the acquirer's input and guide them to determine what is actually replicated, namely the grammar.[34–37] The grammar contains the units of information that are transmitted across generations. As such, the grammar counts as the genotype of a language, while idiolects correspond to its phenotype.[9]

A linguistic system consists of different submodules such as the lexicon, phonology, morphology, syntax, and semantics, each of which has its own units and abstract operations. This means that the units of information in each submodule are potentially distinct. It is not a trivial question to determine what exactly are the units that are transmitted and are amenable to change in each submodule. Chomskyan linguists argue that, at least as far as morphosyntax is concerned, "parameter" is the unit of linguistic heredity.[38–43] In practice, many linguists working with tree topologies on the basis of phonology and morphosyntax employ as taxonomic characters the typological features contained in databases such as WALS,[44] which code systematic patterns observed in languages of the world. The trees obtained are accurate to a certain extent, if we take as our standard the trees constructed on the basis of cognate words (words with the same etymology).[45,46]

A fundamental property of evolvable systems is the concept of inheritance. In biology, inheritance is transmission of genetic information from identifiable parents to their children. In language, inheritance is not a process taking place between humans as biological organisms,

as it is evident that children are not born speaking the native language of their parents. Rather, it is a process taking place between idiolects and can be characterized as the vertical transmission of traits from a population of parental idiolects to a population of the idiolects of the descendant generation. This happens during the critical period of language acquisition. This is indeed an instance of inheritance, as kids do not speak or create a language from scratch but infer their mother language from exposure to the idiolects of the previous generation. The grammar of each speaker is constructed on the basis of the input from the environment, which does not assign any privileged status to biological parents. All external stimuli serve as "parents" as long as they form part of the primary linguistic data the child is exposed to: the idiolects from sources such as extended family, daycare, media, other social groups, and institutions that form part of the child's everyday experience.[22]

THE FOUR PROCESSES UNDER WHICH LANGUAGES EVOLVE

Generation of variation

There are two fundamental steps in the evolution of any evolvable system. The first is the initial production of variation and the second is the dynamics of variation across generations which results in diversification. The first process is the generation of de novo variation, called mutation in biology. In language, de novo variation is generated by linguistic innovations.[25,47] Ringe and Eska[29] identify four sources of linguistic innovations: (a) contact with speakers of other languages that leads to borrowing of foreign words and in some cases may also lead to the borrowing of structures, (b) deliberate manipulation of language which can give rise to new words and can also affect language structure, (c) errors in the course of language acquisition which persist in adulthood and eventually spread in a speech community (pp. 37–44[29]; other approaches, e.g., ref.,[48] attribute innovations of this type to a process commonly known as reanalysis), and (d) a combination of language contact and learner errors arising when migrating individuals use the (foreign to them) dominant language imperfectly.

The transmission of language from generation to generation is not faithful, that is, children are likely to base their grammars on misperceptions[49] or reanalyses of the adult input. Such errors are a major source of language change when they are generalized across young speakers and persist into adulthood. According to Ringe and Eska[29] (p. 28) "a linguistic change has occurred when an innovation has spread and become accepted in a speech community." In evolution-theoretic terms, a linguistic change has occurred when an innovation has been fixed in a language population. One proposal that aims to account for the gradual process of language acquisition leading to the gradual replacement of older linguistic forms by new ones (with the graphs of frequency distributions having the form of S-shaped curves[25–27]) is developed by Yang.[20,50] He models language change in terms of an interaction between internal and external forces following insights from biological evolution.

Selection and drift

Once new variation is produced there are two potential outcomes. Either the new variant is lost in subsequent generations or it prevails in the language population (diffusion[25]). Until this happens the two forms co-exist. The fate of each new variant is determined both by stochastic processes (drift) and by deterministic ones (selection) (Table 1). If the new variants are selectively neutral, their frequencies follow a trajectory with chance fluctuations, which finally leads either to its establishment (frequency 1) or to its extinction (frequency 0). On the other hand, under pure selection the trajectory of each variant is predetermined: when it is beneficial it gets established (positive selection) and when it is detrimental it becomes extinct (negative selection). In most cases, drift and selection act simultaneously and the new variants have a certain probability of fixation or extinction which depends both on their selective advantage (or disadvantage) and on the population size, which determines the strength of drift (Box 3).

In linguistics, little attention has been paid to the potential of random change in the relative frequencies of variants across generations. This relates to the researchers' tendency to look for nonrandom patterns in search for causality. Notable exceptions are Newberry et al.[51] and Clark[52] who employ the term "drift" from biology, a practice we will also follow here (Box 1). They use data of word frequencies (tokens) from corpora and employ mathematical models from population genetics and computer simulations in order to demonstrate that some diachronic phenomena in English, such as the irreguralization or regularization of the past tense of some verbs, have unfolded randomly.

Finding nonrandom patterns in language change and attempting to identify the reason why a specific change took place is the usual way linguistic research proceeds. Selection produces nonrandom patterns. Newberry et al.[51] mathematically demonstrate the workings of selection with the past tense of six verbs in the history of English showing two variants, a regular and an irregular one in the corpus of historical American English. In two of these cases, selection favored the regular variant (wove → weaved and smelt → smelled) while in the other four cases selection favored the irregular variant (lighted → lit, waked → woke, sneaked → snuck, dived → dove). Note that it is not always possible to identify the cause of selection but after identifying a change as selective we can speculate about its causes. For the particular cases at hand, Newberry et al. propose that selection for regularization can be linked to economy or cognitive ease, while the more intriguing case of selection favoring irregularization might involve rhyming with existing irregular verbs, a process also supported by psychological studies.[53,54] For example, for the case of *dive-dove* in the early 20th century they suggest that the choice of the irregular form coincides with a significant increase in the use of the irregular form *drove* due to the invention of cars in that period. Another case of selection Newberry et al. identify concerns the evolution of syntactic verbal negation from the 16th to the 20th century which involves a pattern known in linguistics as Jerspersen's cycle[55] featuring three steps: preverbal negation (Old English "Ic **ne** secge"), followed by double negation (Middle English "I **ne** seye **not**"), which leads to postverbal negation (early Modern English "I say **not**"). To a linguist the discovery that Jespersen's

Box 3. Interactions between drift and selection

Realistically, we expect many more cases where drift and selection interact than pure drift and pure selection situations if we take into account insights from evolutionary biology. Newberry et al.[51] describe one linguistic change that starts with drift and continues with selection. This is the case with *do*-support, which arguably increased by chance via drift in interrogative sentences in early Modern English and then was rapidly generalized in negative contexts providing evidence for selection.

A more complex case concerns simultaneous interaction between drift and selection. Evolutionary theory suggests that neutral variants ("variant" being the counterpart of polymorphism in biology) are only subject to drift and are expected to be fixed with the same rate in both small and large populations. Selection, on the other hand, is reflected in the fixation rate of the variants relative to the size of populations. Selection is more efficient in larger populations, whereas drift prevails in smaller populations. If a variant is subject to negative selection it will be more easily removed in a large population, while the same variant has a higher probability to be fixed in small populations thanks to drift. On the other hand, if a variant is subject to positive selection, it has higher probability to be fixed in a large population and to be removed by chance in small populations. To illustrate this, we will use a case that is well known in linguistics, namely the observation that irregular inflection is more common in frequent verbs than in infrequent ones (see also[51]). Let us take as a starting point two irregular verbs, one infrequent and one frequent. The infrequent one represents a small population of tokens (a few tokens present in everyday life and as input to children), while the frequent one represents a large population of tokens (extensive use and therefore a large number of tokens in children's input). Note that, what matters for this case is populations of tokens and not populations of speakers or idiolects, as employed in studies such as,[76,77] what we call populations of individuals in Table 1. The two notions must be kept distinct but the workings of drift and selection relative to the size of populations arguably can be applied to both. Let us ask the question which one of the two irregular verbs has a higher chance to become regular. In order to do so it will have to change. Resistance to regularization ('faithfulness to the input', in optimality theoretic terms[92]) should be considered as an instance of negative selection because it will violate the faithfulness to the input of the child. Therefore, it is expected to be selectively removed. As a result, even though regularization is naively considered to be driven by constraints against markedness and thus a positive change that would be favored by the system, it turns out that it is actually subject to negative selection since it leads to a novel form. The net result of this is that, as drift predicts, regularization is more likely to happen in smaller populations, that is, infrequent forms, than in larger ones, that is, frequent ones. This is what we actually observe. It is well known to linguists that common words are much more likely to be irregular than uncommon ones. The usual explanation for this in linguistics is that common words are much more frequent in the primary linguistic data than uncommon ones and therefore the child is much more likely to introduce an innovation favoring a regular pattern in a word (s)he encounters for the first time or with very low frequency.

cycle is driven by selection comes as no surprise, as the directionality of this change has been identified for numerous unrelated languages and there are theoretical explanations for this.[56,57] As is known in statistics, drift is the null hypothesis. In order to postulate selection, we need to have evidence for rejecting the null hypothesis. This creates testable hypotheses that can be evaluated on the basis of methods such as the one proposed in ref.[51] Note that, even though the specifics of this method concerning binning decisions in linguistic corpora have been criticized by Karjus et al.,[58] the authors nevertheless conclude that the basic results of Newberry et al. hold and that distinguishing between stochastic and selective processes in the investigation of language change is worth pursuing.

Migration

The fourth evolutionary process is migration, that is, the exchange of variation among populations. In linguistics, this is a pervasive and multifaceted phenomenon which falls under the term "language contact"[59,60] and is perhaps best-studied among the processes we describe here. It is important to note that, unlike exchange of variation in biology which only happens at the population level and cannot happen among species, exchange of variation in language is not limited to the population level but can take place between any two languages regardless of their degree of relatedness (e.g., how recent a common ancestor they have and/or whether they are mutually intelligible).

When language contact leads to change, then either (a) features of the one language enter into the other by a process generally called "transfer/borrowing", or (b) some characteristics change via a process called "restructuring," whereby the result does not resemble any feature of the two languages that get in contact[61] (p. 415). Both outcomes are involved in the generation of de novo variation. Regarding transfer/borrowing, researchers have demonstrated that different elements and patterns have different degrees of borrowability. For example, contentful words (e.g., nouns and verbs) are easier to borrow than functional/grammatical ones (e.g., pronouns and articles) and words are easier to borrow than structural patterns in phonology, morphology, and syntax which require more intense contact.[59,62,63] Regarding restructuring, a striking and well-studied case of generation of de novo variation is *pidgin* formation which has arisen in situations where slaves or workers were transferred from home to different countries with a dominant language foreign to them (e.g., African

populations transferred to the Caribbean Islands). In these situations, the migrants of the first generation construct a communication code that makes use of the most unmarked (non-complex) features that are likely to be understood by both them and native speakers of the dominant language. In subsequent generations, many of these features are gradually reanalyzed according to general first language acquisition principles,[30,64] eventually leading to a new natural language which contains innovative forms and constructions that are not transfers from any of the languages in contact. This constitutes the genesis of new native languages, known as *creoles*.[65,66]

It is important to stress that these four processes operate in a short evolutionary timescale and shape the variation within populations (polymorphism). As different fixed variants are accumulated in alternative populations they lead to population-splits and, eventually, in long timescales, to speciation. The theory of phylogenetics treats species as monomorphic when studying their divergence. Similarly, linguistics treats languages as uniform when comparing properties of different languages, while it focuses on variation within populations (dialects, registers, etc.) when studying language change.

DIFFERENCES BETWEEN LIFE AND LANGUAGE

As we hope has become clear by now, there are many commonalities between life and language which originate from their nature as evolvable systems. The historical work that has been successfully reconstructing proto-languages with the aid of trees has demonstrated the robustness of vertical transmissions in language, just as in biology (see refs.[7,8]). Nevertheless, it is underestimated that there are also crucial differences between the two systems. These differences make each system unique and should be taken into account in order to determine how methods from biology can be adjusted to linguistics. These differences can be summarized as follows:

1. It is a definitional property of language that it connects sounds (in spoken languages)/signs (in sign languages) to meaning mediated by a computational system (morphology–syntax). This means that the grammar, the genotype of language, has several levels and each level has its own units and rules of combination. There is no such analogue in life where the genotype is a single level consisting of a sequence of nucleotides which encodes the genetic information.
2. As opposed to life where inheritance reduces to replication of parental genotypes, genotypes in language (i.e., the grammar) are necessarily affected by the phenotypes (the idiolects, called "speech" by Haider[9]) that the acquirers are exposed to, because the acquirers have no direct access to genotypes.
3. A significant difference between the two systems concerns the inheritance of acquired characteristics. In living organisms, the phenotype may change during the lifetime of an individual either due to developmental changes or due to its interaction with the environment, but these changes leave genetic information intact, which is passed on from generation to generation and is altered only through mutation. In language, alongside inherited traits obtained through the children naturally acquiring their native language in the first years of their life, an idiolect may also incorporate characteristics after the critical period due to external pressure from education and trends in the linguistic community related to language ideologies, attitudes, etc. In the next generation, children treat the idiolects that contain both inherited and acquired traits as their input and construct their grammar accordingly.[67,68]
4. Another important difference between the two evolving systems concerns parenthood. In living organisms parents can be one, two, or more individuals, always specific and traceable. In language, vertical transmission of information can be influenced by many potentially untraceable parental idiolects as the acquisition of one's native language is influenced not necessarily by physical parents but by the environment more broadly.
5. Furthermore, any two languages, no matter how genetically distant they are, can interact and influence one another, in contrast to biological species where the exchange of genetic information is severely restricted.
6. From a methodological point of view, an important difference between biology and linguistics regards the data. In biology the data that are currently used for phylogenetics are mostly genetic, which has two consequences: (a) their vast amount and (b) their independence from the environment which guarantees that they are exclusively vertically transmitted. In language, the data that have so far been used are morphological (in the biological sense), that is, close to the surface. More abstract data, that is, the counterpart of genetic data, are not generally agreed upon, their identification depends on theory internal reasoning and, in contrast to biology, it is far from evident that the abstract characters outnumber the more superficial ones (many linguists do not believe in abstract characters and even proponents of parameter-based frameworks typically assume that parameters are limited in number).

Each of these differences has implications for the way we approach language history as a set of evolutionary processes. First, the existence of several discrete levels of linguistic analysis can be seen as an asset for the study of language change because it opens up the possibility to construct independent trees based on the units of the different levels and compare them in order to control for their adequacy. This is not possible in biology.

Second, given the nature of inheritance and parenthood in language, biological phylogenetic methods which rely on the assumptions that (i) information is exclusively vertically transmitted and (ii) acquired characteristics are not inherited cannot be blindly applied to the study of language change. We need to make sure that we draw on data that are less susceptible to borrowing. This sort of work that has been undertaken for cognates and has resulted in the Swadesh list[69] but it is entirely missing in the work on morphosyntax. Since all characteristics of the grammar can be transferred areally or genealogically[70] with higher or lower probability (cf. scales of borrowability, stability, etc.[60]) it is imperative to develop more sophisticated models of the interaction between areal and genealogical transmission when we look at grammatical data (cf.[71] for an exploration of the problem with the aid of

computer simulations). Moreover, even though biology has developed methods to identify events of horizontal transmission (e.g., ref.[72]), these methods are not easily transferable to morphosyntactic characters due to the pervasive presence of homoplasies (Box 1).

At this point, it is necessary to clarify the concept of horizontal transmission taking place in biology and in language. In biology, horizontal transmission is the transfer of genetic information between two contemporary individuals who do not stand in a parent-offspring relationship. Such transfers are rare in multicellular organisms but quite common in unicellular ones, particularly in prokaryotes. In language, it is very common for idiolects, dialects, and languages to exchange material when speakers are in contact with each other. This affects the phenotype of the language, that is, the speech of the speakers, but not immediately its genotype, that is, the grammar. In order for borrowed elements to affect the grammar disrupting vertical/genealogical relationships, they need to become the input of acquirers of the next generation. This entails that contact-induced grammar change is always mediated by vertical transmission and it is thus incorrect to equate borrowing to horizontal transmission in biology, despite the fact that it disrupts vertical relationships. For this reason, we are using the term "areal transfer" and not "horizontal transfer" when we discuss the effects of language contact.

WHY STUDY THE TWO EVOLVABLE SYSTEMS IN CONCERT?

Studying biology and linguistics as two subcases of the general concept "evolvable system" has implications for the research in both domains. Under a unified framework, new questions arise which can be explored by importing and adapting methods and tools from one field to the other.[7,8] The most salient domain where we can find affinities between historical linguistics and evolutionary biology is phylogenetic research, where linguists nowadays adopt the quantitative and statistical methods of modern evolutionary research in an attempt to reconstruct the history of several language families.

Most such attempts have mainly capitalized on cognate data, a famous example being the work by Gray and Atkinson on the origin of Indo-European.[73] Despite the skepticism expressed in some recent research concerning the usefulness of morphosyntactic taxonomic characters[45] there are two arguments in favor of extending this line of quantitative research to morphosyntax. First, methods based on cognate data saturate at a shallow historical depth, whereas morphosyntactic data may help us probe more ancient splits.[70] Second, as already mentioned, trees based on morphosyntactic data can serve as controls for trees resulting from cognate data and vice versa.[74] In order to work with morphosyntactic data more effectively, we need to identify characters prone to areal transfer and formulate prior assumptions regarding the conditions under which a certain distribution of characters should be attributed to contact, in order to factor out their effects when computing exclusively genealogical relationships.

Another domain where the methods from evolutionary biology can open up new horizons in historical linguistics is the study of the emergence of variation and its dynamics within and across populations on the basis of mutation, drift, and selection. There are some sporadic examples of how this type of research should unfold. We have already mentioned the study of intralinguistic variation and its spread patterns by Newberry et al.,[51] who convincingly argue that certain processes in the history of English should be attributed to drift while others result from selection and yet others from the combination of the two (Box 3). Another example is provided by studies,[75,76] which correlate language evolution with the size of the speakers' population. There are many testable predictions that this framework gives rise to and can be explored in the future. For example, a positive correlation is expected between the size of the populations and the amount of neutral variants in different domains (phonemes, lexical items, syntactic structures, etc.) in monolingual societies. Even part of what is common ground in historical linguistics, for example, the positive correlation observed between language isolation and linguistic divergence follows from the framework under discussion. Every time different groups are isolated within a language population for geographical or social reasons, differences start to gradually accumulate, eventually resulting in the emergence of language divisions which correspond to different varieties of the language (e.g., dialects). When continued, this process leads to language splits (see ref.[77] for ecological factors influencing language isolation). This corresponds to the conditions determining speciation in biology. In both systems, isolation is the key factor.

CONCLUSIONS

In conclusion, the similarities between different types of evolvable systems, as summarized in Table 1 for life and language, are enough to guarantee that there are testable hypotheses arising in one discipline that can be applied to the other. The exact methodological tools that will be employed need to be adjusted to the specific characteristics of the system under investigation. For the cases at hand, the two major differences that need to be taken into account concern (a) the nature of the data and (b) the interaction between genealogical and areal transfer. Both disciplines can benefit from approaching their questions from the broader perspective of evolvable systems. Apart from the benefits for linguistics outlined so far, there are also potential gains for biology. For example, the evolutionary models in biology assume noninheritance of acquired characteristics. Recent advances in epigenetics have raised the possibility of inheritance of acquired characteristics and this has even led to putting into question the adequacy of the current evolutionary framework.[78,79] Historical linguistics teaches us that inheritance of acquired characteristics, which is common in language, is an admissible property of evolvable systems and interacts with genealogical/vertical transmission in interesting ways.

ACKNOWLEDGMENTS

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: HFRI-FM17-44) and by the Special Account for Research Funds of University of Crete (KA 10213). The authors would like to

thank Drs. Maria Margarita Makri and Pavlos Pavlidis, and the Professors E. Zouros and I. Karakassis for their valuable comments on the manuscript.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

Data sharing not applicable—no new data generated.

REFERENCES

1. Wagner, A. (2011). *The origins of evolutionary innovations: A theory of transformative change in living systems*. OUP.
2. Vellend, M. (2010). Conceptual synthesis in community ecology. *Quarterly Review of Biology*, *85*(2), 183–206. doi: https://doi.org/10.1086/652373
3. Bopp, F. (1816). *Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der griechischen, lateinischen, persischen und germanischen Sprache: Nebst Episoden des Ramajan und Mahabharat und einigen Abschnitten aus den Vedas*. Andreä.
4. Brugmann, K., & Delbrück, B. (1967). *Grundriss der vergleichenden Grammatik der indogermanischen Sprachen*.
5. Beekes, R. S. (2011). *Comparative Indo-European linguistics: An introduction*. John Benjamins Publishing.
6. Darwin, C. (1872). The descent of man, and selection in relation to sex *(Vol. 2)*.. D. Appleton.
7. Atkinson, Q. D., & Gray, R. D. (2005). Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic Biology*, *54*(4), 513–526. https://doi.org/10.1080/10635150590950317
8. Bromham, L. (2017). Curiously the same: Swapping tools between linguistics and evolutionary biology. *Biology & Philosophy*, *32*(6), 855–886. https://doi.org/10.1007/s10539-017-9594-y
9. Haider, H. (2021). Grammar change: A case of Darwinian cognitive evolution. *Evolutionary Linguistic Theory*, *3*(1), 6–55.
10. Mace, R., & Holden, C. J. (2005). A phylogenetic approach to cultural evolution. *Trends in Ecology & Evolution*, *20*(3), 116–121. https://doi.org/10.1016/j.tree.2004.12.002
11. Bowern, C. (2018). Computational phylogenetics. *Annual Review of Linguistics*, *4*, 281–296.
12. Croft, W. (2008). Evolutionary linguistics. *Annual Review of Anthropology*, *37*, 219–234.
13. List, J.-M., Pathmanathan, J. S., Lopez, P., & Bapteste, E. (2016). Unity and disunity in evolutionary sciences: Process-based analogies open common research avenues for biology and linguistics. *Biology Direct*, *11*(1), 1–17.
14. Mufwene, S. S. (2001). *The ecology of language evolution*. Cambridge University Press.
15. Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219.
16. Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1-2), 61–82.
17. Tomasello, M. (2009). *Constructing a language*. Harvard University Press.
18. Chomsky, N. (1981). *Lectures on government and binding*. Foris.
19. Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.
20. Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.
21. Lenneberg, E. H. (1967). The biological foundations of language. *Hospital Practice*, *2*(12), 59–67.
22. Guasti, M. T. (2017). *Language acquisition: The growth of grammar*. MIT Press.
23. Chomsky, N. (1979). The logical structure of linguistic theory. *Synthese*, *40*(2), 317–352.
24. Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*(1), 1–22.
25. Kroch, A. (2000). Syntactic change'. In M. Baltin, & C. Collins (Eds.), *The handbook of contemporary syntactic* (pp. 629–739). Blackwell.
26. Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, *1*(3), 199–244.
27. Lightfoot, D. (1997). Catastrophic change and learning theory. *Lingua*, *100*(1-4), 171–192. doi: https://doi.org/10.1016/S0024-3841(93)00030-C
28. Hale, M. (2007). *Historical linguistics: Theory and method*. Wiley-Blackwell.
29. Ringe, D., & Eska, J. F. (2013). *Historical linguistics: Toward a twenty-first century reintegration*. Cambridge University Press.
30. Roberts, I. (2007). *Diachronic syntax*. Oxford University Press.
31. Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Education.
32. Bloch, B. (1948). A set of postulates for phonemic analysis. *Language*, *24*(1), 3–46.
33. Mayr, E. (1940). Speciation phenomena in birds. *The American Naturalist*, *74*(752), 249–278.
34. Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry*, *29*(1), 1–36.
35. Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(3), 407–454.
36. Lightfoot, D. (2006). *How new languages emerge*. Cambridge University Press.
37. Roeper, T. (2011). What frequency can do and what it can't. In *Frequency effects in language acquisition*, I. Gülzow, & N. Gagarina (Eds.), (pp. 23–48). De Gruyter Mouton.
38. Galves, C., Cyrino, S., Lopes, R., Sandalo, F., & Avelar, J. (2012). *Parameter theory and linguistic change* (Vol. 2). OUP.
39. Roberts, I. (2012). A programme for comparative research. *Parameter Theory and Linguistic Change*, *2*, 320.
40. Baker, M. C. (1996). *The polysynthesis parameter*. Oxford University Press.
41. Baker, M. C. (2003). Linguistic differences and language design. *Trends in Cognitive Sciences*, *7*(8), 349–353. https://doi.org/10.1016/S1364-6613(03)00157-8
42. Biberauer, T., & Roberts, I. (2016). Parameter typology from a diachronic perspective. *Theoretical approaches to linguistic variation*, *234*, 259.
43. Kayne, R. S. (2013). Comparative syntax. *Lingua*, *130*, 132–151. https://doi.org/10.1016/j.lingua.2012.10.008
44. Dryer, M. S. (2013). Order of relative clause and noun. In M. S. Dryer, & M. Haspelmath (Eds.), *The world atlas of language structures online*.
45. Greenhill, S. J., Wu, C. H., Hua, X., Dunn, M., Levinson, S. C., & Gray, R. D. (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(42), E8822–E8829. https://doi.org/10.1073/pnas.1700388114
46. Dediu, D. (2011). A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proceedings of the Royal Society B-Biological Sciences*, *278*(1704), 474–479. https://doi.org/10.1098/rspb.2010.1595
47. Lightfoot, D. (1999). *The development of language: Acquisition, change, and evolution*. Wiley-Blackwell.
48. Hale, M. (1998). Diachronic syntax. *Syntax*, *1*(1), 1–18.
49. Ohala, J. J. (1993). The phonetics of sound change. *Historical Linguistics: Problems and Perspectives*, *237*, 278.
50. Yang, C. D. (2000). Internal and external forces in language change. *Language Variation and Change*, *12*(3), 231–250.
51. Newberry, M. G., Ahern, C. A., Clark, R., & Plotkin, J. B. (2017). Detecting evolutionary forces in language change. *Nature*, *551*(7679), 223–226. https://doi.org/10.1038/nature24455

52. Clark, R. (2020). Drift, finite populations, and language change. In *Syntactic architecture and its consequences I* A. Bárány, T. Biberauer, J. Douglas, & S. Vikner (Eds.), (p. 3).
53. Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*(1), 1–56.
54. Ullman, M. T. (1999). Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. *Language and Cognitive Processes*, *14*(1), 47–67.
55. Jespersen, O. (1917). *Negation in English and other languages* (Vol. 1). AF Høst.
56. Horn, L. (1989). *A natural history of negation*, University of Chicago Press.
57. Ladusaw, W. A. (1993). Negation, indefinites, and the Jespersen Cycle. Paper presented at the *Annual Meeting of the Berkeley Linguistics Society*, Semantic Universals, pp. 437–446.
58. Karjus, A., Blythe, R. A., Kirby, S., & Smith, K. (2020). Challenges in detecting evolutionary forces in language change using diachronic corpora. *Glossa*, *5*(1), 45.
59. Thomason, S. G. (2001). *Language contact*. Citeseer.
60. Thomason, S. G., & Kaufman, T. (1992). *Language contact, creolization, and genetic linguistics*. University of California Press.
61. Walkden, G. L. (2017). *The actuation problem*. Cambridge University Press.
62. Matras, Y. (2008). The borrowability of structural categories. In *Grammatical borrowing in cross-linguistic perspective*, Y. Matras, & J. Sakel (Eds.), (pp. 31–74). De Gruyter Mouton.
63. Muysken, P. (2013). Two linguistic systems in contact: Grammar, phonology, and lexicon. In *The handbook of bilingualism and multilingualism*, T. K. Bhatia, & W. C. Ritchie (Eds.), (pp. 193–216). Wiley-Blackwell.
64. Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences*, *7*(2), 173–188.
65. Aboh, E., & DeGraff, M. (2017). A null theory of creole formation based on Universal Grammar. In *The Oxford handbook of universal grammar*, I. Roberts (Ed.), (401–458).
66. Ansaldo, U., & Meyerhoff, M. (2020). *The Routledge handbook of Pidgin and Creole languages*. Routledge.
67. Labov, W. (1989). The child as linguistic historian. *Language Variation and Change*, *1*(1), 85–97.
68. Labov, W. (2013). Preface: The acquisition of sociolinguistic variation. *Linguistics*, *51*(2), 247–250.
69. Anderson, S. R., Anderson, S. R., & Lightfoot, D. W. (2002). *The language organ: Linguistics as cognitive physiology*. Cambridge University Press.
70. Nichols, J. (1992). *Linguistic diversity in space and time.* University of Chicago Press.
71. Currie, T. E., Greenhill, S. J., & Mace, R. (2010). Is horizontal transmission really a problem for phylogenetic comparative methods? A simulation study using continuous cultural traits. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *365*(1559), 3903–3912. https://doi.org/10.1098/rstb.2010.0014
72. Sarris, P. F., Ladoukakis, E. D., Panopoulos, N. J., & Scoulica, E. V. (2014). A phage tail-derived element with wide distribution among both prokaryotic domains: A comparative genomic and phylogenetic study. *Genome Biology and Evolution*, *6*(7), 1739–1747.
73. Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*(6965), 435–439. https://doi.org/10.1038/nature02029
74. Longobardi, G., Buch, A., Ceolin, A., Ecay, A., Guardiano, C., Irimia, M., Michelioudakis, D., Radkevich, N. V., & Jaeger, G. (2016). Correlated evolution or not? phylogenetic linguistics with syntactic, cognacy and phonetic data. Paper presented at *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11)*.
75. Bromham, L., Hua, X., Fitzpatrick, T. G., & Greenhill, S. J. (2015). Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(7), 2097–2102. https://doi.org/10.1073/pnas.1419704112
76. Greenhill, S. J., Hua, X., Welsh, C. F., Schneemann, H., & Bromham, L. (2018). Population size and the rate of language evolution: A test across Indo-European, Austronesian, and Bantu languages. *Frontiers in Psychology*, *9*, 576. doi: ARTN 576 10.3389/fpsyg.2018.00576
77. Hua, X., Greenhill, S. J., Cardillo, M., Schneemann, H., & Bromham, L. (2019). The ecological drivers of variation in global language diversity. *Nature Communications*, *10*, 2047. doi: ARTN 2047 10.1038/s41467-019-09842-2
78. Skinner, M. K. (2015). Environmental epigenetics and a unified theory of the molecular aspects of evolution: A neo-Lamarckian concept that facilitates neo-Darwinian evolution. *Genome Biology and Evolution*, *7*(5), 1296–1302.
79. Richards, C. L., Bossdorf, O., & Pigliucci, M. (2010). What role does heritable epigenetic variation play in phenotypic evolution? *BioScience*, *60*(3), 232–237.
80. Bickerton, D. (2007). Language origins: perspectives on evolution. *Journal of Linguistics*, *43*(1), 259–264. https://doi.org/10.1017/s0022226706274589.
81. Fitch, W. T. (2010). *The Evolution of Language (Approaches to the Evolution of Language)*. Cambridge University Press.
82. Bromham, L., Hua, X., Algy, C., & Meakins, F. (2020). Language endangerment: a multidimensional analysis of risk factors. *Journal of Language Evolution*, *5*(1), 75–91. https://doi.org/10.1093/jole/lzaa002.
83. Greenhill, S. J., Atkinson, Q. D., Meade, A., & Gray, R. D. (2010). The shape and tempo of language evolution. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1693), 2443–2450. https://doi.org/10.1098/rspb.2010.0051.
84. Baker, A., Archangeli, D., & Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change*, *23*(3), 347–374. https://doi.org/10.1017/s0954394511000135.
85. Weinreich, U., Labov, W., & Herzog, M. (1968). *Empirical foundations for a theory of language change*, Vol. 58, University of Texas Press Austin.
86. Sapir, E. (1921). *Language: An introduction to the study of speech*, Harcourt, Brace New York.
87. Aikhenvald, A. Y., & Dixon, R. M. W. (2007). *Grammars in Contact: A cross-linguistic typology*, Oxford: Oxford University Press.
88. Yanovich, I. (2016). *Genetic drift explains Sapir's "drift" in semantic change. The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)* 2016, https://evolang.org/neworleans/pdf/EVOLANG_11_paper_24.pdf.
89. Friederici, A. D. (2017). *Language in Our Brain: The Origins of a Uniquely Human Capacity, The MIT Press*. https://direct.mit.edu/books/book/3653/Language-in-Our-BrainThe-Origins-of-a-Uniquely.
90. 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*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569.
91. Greenhill, S. J., Currie, T. E., Gray, R. D. (2009). Does horizontal transmission invalidate cultural phylogenies?. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1665), 2299–2306. https://doi.org/10.1098/rspb.2008.1944.
92. Prince, A., Smolensky, P. (1997). Optimality: From Neural Networks to Universal Grammar. *Science*, *275*(5306), 1604–1610. https://doi.org/10.1126/science.275.5306.1604.

How to cite this article: Ladoukakis, E. D., Michelioudakis, D., & Anagnostopoulou, E. (2021). Toward an evolutionary framework for language variation and change. *BioEssays*, e2100216. https://doi.org/10.1002/bies.202100216