Phonemic clicks and the mapping asymmetry: How language emerged and speech developed

M.A.C. (Riny) Huybregts

Utrecht Institute of Linguistics, Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 7 September 2016
Received in revised form 26 January 2017
Accepted 27 January 2017
Available online xxx

ABSTRACT

Language existed before human populations became separated (all descendant populations have language) but language did not emerge until long after these population divergences occurred (behavioral modernity only showed then). Distinguishing capacity for language from externalized language resolves the apparent paradox, eliminates the need of proto-language, and rules out monogenesis. Speech emerged only after the capacity for language became (sufficiently) fixated in the species. This accords well with a fundamental property of human language. Rules mapping to meaning rely on structural properties only, while rules mapping to sound are (also) sensitive to linear order, reflecting properties of sensorimotor modalities. The asymmetry suggests (i) primacy of internal language over speech/sign, and (ii) evolution of capacity of language preceding externalized language. Click phonemes with their unique genealogical, genetic and geographical distribution may be relevant here. All biologically Khoisan groups speak clicks languages, which are spoken by biologically Khoisan groups only. Separation followed possession of internal language but preceded externalized language. Clicks were recruited for externalization in San populations only after deepest separation.

© 2017 Published by Elsevier Ltd.

Contents

1. Introduction ................................................................. 00
2. Some properties of clicks .................................................. 00
   2.1. Click sounds: how complex are they? .......................... 00
   2.2. Click phonemes outside Khoisan ............................... 00
3. A puzzle for evolution of language ..................................... 00
4. The language phenotype ................................................ 00
5. Asymmetric mappings to the interfaces ................................ 00
   5.1. The structural position of verb-second ...................... 00
   5.2. Anaphoric binding .............................................. 00
   5.3. Bound variable interpretation ................................ 00
   5.4. Linear order: not even wrong! ............................... 00
6. Capacity for language and the use of it .............................. 00
7. Dissolving the paradox ................................................ 00
8. Some consequences .................................................... 00
   8.1. Asymmetry of interface mappings to meaning and sound 00
   8.2. Externalized language and phonemic clicks .................. 00
   8.3. Vocal tract morphology in San .............................. 00
   8.4. Emergence of clicks in externalized language ............. 00
   8.5. The illusion of Proto-World .................................. 00
   8.6. A failed human experiment with a twist ................... 00
   8.7. Genomic results suggest late externalization of language 00

E-mail address: M.A.C.Huybregts@uu.nl

http://dx.doi.org/10.1016/j.neubiorev.2017.01.041
0149-7634/© 2017 Published by Elsevier Ltd.

1. Introduction

In his book, Chance and Necessity, a landmark contribution to the natural philosophy of modern biology, the French biochemist and Nobel laureate Jacques Monod plausibly argued that “when behavior implies elements acquired through experience, they are acquired according to a program, and that program is innate – that is to say, genetically determined” (Monod, 1972; p. 152). He then continues, “Thus, in all likelihood, is the process to be understood whereby the child acquires language.” So if the core of human language is part of human biology, evidence for evolution of language must ultimately be biolinguistic, linking properties of the language phenotype to aspects of the neurobiology of language and underlying genomics. Since, unfortunately, such evidence is only sparse, fragmentary and suggestive, evolution of language is a hard problem to solve. There we attempt to connect some suggestive results from various disciplines, in particular genomics, palaeoanthropology, archaeology, and linguistics in an effort to account for the gap that must have existed between the emergence of internal language and its externalization in spoken or sign language. In particular, the role of click sounds in evolution of language will be discussed. Language design is characterized by a deep asymmetry of the mappings from syntax to meaning and sound. The question addressed here is what clicks have to say on the asymmetry of the mappings, their nature, origin and use.

Phonemic clicks are not exactly fashionable outside of Africa. And inside Africa they are in a relevant sense restricted to Khoisan languages only, where they are immensely popular. Consequently, their skewed geographical and genealogical distribution should raise some questions with implications for the emergence of human language, in particular (a) the combinatorial operation Merge preceding speech in evolution of language, and (b) the timing of Merge relative to speech in human evolution. Merge is the basic operation of language that provides for an unbound array of hierarchically structured expressions with systematic interpretations at the interfaces with other internal systems, the sensorimotor system SM for externalization (perception and production) and the conceptual-intentional system CI for structured thought and planning of action (Chomsky, 2010, 2011, 2013b). Some conceptual and empirical arguments involving clicks will be presented that show that speech, externalized language, must have developed later than the basic principle of language that yielded the distinctively human and linguistic property of discrete infinity.

Click consonant systems seem very stable and are easily learnable by infants. Therefore the question naturally pops up. Why are they so isolated? Restricted almost to Khoisan speaking populations only! In fact, all Koisian languages have click consonants. More precisely, all biologically Khoisan groups speak click languages (Güldemann, 2007; Güldemann and Stoneking, 2008), and apart from some exceptional cases that are likely to receive independent explanations (borrowing or language shift), click languages are spoken by biologically Khoisan groups only. Results from a recent study of the genetic prehistory of southern Africa (Pickrell et al., 2012) support the hypothesis of an ancient link between southern African Khoisan (northwestern and southeastern Kalarahi groups, who separated only within the last 30,000 years) and eastern African Hadza and Sandawe. Interestingly, there may be an answer to the question raised above which directly supports the asymmetry of the mappings to the language interfaces with sound and meaning, and that furthermore suggests that language evolved primarily as a system of thought (Chomsky, 2011, 2013b, 2015; Berwick and Chomsky, 2011, 2016).

Vocalization abilities are prerequisites for externalized language (speech), but are not themselves part of language – they say nothing whatsoever about the internal representation of phonological systems. What is relevant here is that by clicks we invariably mean specific vocalizations (cf. § 2.1) with phonemic function, linguistic (not paralinguistic) vocalizations that are wired to internal phonological representations and satisfy conditions imposed by universal phonology. Articulation of click phonemes was just part of the externalization of language that occurred only after Merge took place. The sensorimotor system had been essentially in place for a long time by then so that apart from the wiring of articulated/perceived sounds to internal representations of phonological systems (developing the capacity for mentally representing phonological structure) not much further evolution may have been necessary.

2. Some properties of clicks

2.1. Click sounds: how complex are they?

Click consonants are complex obstruents that are articulated with a double closure in the oral cavity. The pocket of air thus enclosed between the more forward and more backward occlusions is rarefied by a sucking action of the tongue causing a velar ingressive airstream. The forward closure is released subsequently, sharply producing the distinctive “pop” sound. The sound produced by the rapid release of the anterior occlusion is accompanied by pulmonic airstreams that follow the release of the posterior closure (Ladefoged and Traill, 1984, 1994; Ladefoged and Maddieson, 1996). Clicks are complex speech sounds, and click phoneme systems may be quite elaborate, but notwithstanding all this, their inventory sizes can be exceptionally large in many Khoisan languages.

However, as argued by Bradfield (2014), sizes can be reduced dramatically to more normal proportions by analyzing clicks as clusters of pure click phonemes (lingual obstruents) and accompaniment phonemes (specified for laryngeal and manner features of articulation). This way the click inventory size of 1Xöö, which may be as large as 83 (Traill, 1985) or even 115 (namely 5 lingual obstruents × 23 unitary accompaniments if all of these are retained), can be significantly reduced to 5 pure clicks plus 23 phonemic accompaniments, which each occur independently in the language. Bradfield shows how numbers can be further reduced if clustering is analyzed concurrently rather than sequentially. If his analysis is accepted, 1Xöö will have only 13 click phonemes (five pure clicks and eight accompaniments) with simple clustering yielding apparent complexity. Generally, treating clicks as phonemes concurrent with phonemic accompaniments will radically reduce inventory size.

Reducing complexity to clustering has a few consequences as Bradfield observes. Significantly smaller click inventory sizes make Khoisan languages much less “exotic” (though there are still the modest numbers of pure clicks that continue to characterize Khoisan languages). Furthermore, clustering analysis provides a
better explanation of the acquisition and stability of these languages. Finally, it challenges the proposal original with Atkinson (2011) that phoneme inventory size is progressively reduced with increasing distance from Africa.

Click sounds are also special in having genealogically, genetically and areally restricted properties unlike other exotic speech sounds like e.g. doubly-articulated labio-velars (e.g. g b, k p), which frequently occur in the majority of languages in West and Central-African (Niger-Congo, Ubangiian, Chadic, Central Sudanic) as well as in New Guinean languages (ladefoged and Maddieson, 1996). Some or all of these properties may find an explanation, partially at least, in the vocal tract morphology of Khoisan speakers, who seem to have short smooth palates that may facilitate click production (Traill, 1985). This group difference may have relevance for the evolution of externalized language, speech, and the positioning of phonemic clicks in evolutionary space and time.

2.2. Click phonemes outside Khoisan

The few exceptions include, first, some Bantu languages (Kavango Bantu, Sotho, Ye, and the Nguni languages, e.g. Xhosa, Zulu, and Ndebele), which have borrowed their clicks from neighbouring Khoisan languages. Second, Sandawe. Next, Hadza, and finally Dahalo, a Cushitic language with only dental clicks. The “functional load” of clicks in Hadza and Dahala is pretty low (just a few clicks and a relatively low frequency of lexical items with click phonemes). Although we argue for click retention in Hadza below, both these languages may have borrowed clicks from neighbouring Khoisan groups when these languages were still spoken in East-Africa before the Bantu expansion invaded these regions (Ehret, 2013). Similarly for Sandawe. But here it may be the case that Sandawe is in fact related to Khoe (“Central Khoisan”) genealogically (Güldemann and Elderkin, 2010). Sandawe and Khoen are genetically related but diverged before the Bantu expansion (Tishkoff et al., 2007). Alternatively, Hadza may have retained clicks after they separated from San early in human history (Knight et al., 2003; Pickrell et al., 2012), and Dahalo could be a case of language shift with click retention (Güldemann, 2007; Sands and Güldemann, 2009).

Within Africa clicks are therefore uniquely associated with Khoisan, at least in the evolutionary time frame of our interest. But there is more. Strikingly, there are no click languages outside of Africa. The one exception here, Damin, an invented but extinct ritual speech register of Lardi, has just a few phonemic nasal clicks, probably introduced from the paralinguistic use of click sounds. Assuming there are about 5000 non-African languages, “outlier” Damin would be the single exception to an otherwise exclusive association of phonemic clicks with the Khoisan language family, or rather – as improved knowledge (Güldemann, 2007; Güldemann and Stoneking, 2008; Voszen, 2013) would have it – with a Sprachbund of loosely related isolates Tuu (“Southern Khoisan”), Ju and Hôa (“Northern Khoisan”), and Khoe (“Central Khoisan”). Rather than accept Damin, the only attested “language” with click sounds outside of Africa, as exemplifying the naturalness of click-using languages, we prefer to consider it an exotic contingency, and dismiss the ≤ 0.0002 chance occurrence of a non-African invented “click language” with an exclusively ritual use. Absent singleton Damin, the overarching question seems to be why phonemic clicks are completely absent outside of Africa, but are prominently and idiosyncratically present in a genealogically related cluster of languages spoken by genetically closely related populations in geographically constrained regions of southern Africa. This question has received no satisfactory answer yet. Evidently, no answer is acceptable that prioritizes chance over necessity. It is reasonable to conclude, therefore, that apart from a few “exceptional” cases of borrowing and/or language shift, the relation between click languages and (pre-historic) Khoisan populations is one of correspondence.

Finally, there are still a few “quirky” loose ends to be tied up: Dama and Khwe groups. Dama (or Ṛũ-Khoe, i.e. “Black People”) are presumably a remnant group of hunter-gatherers having adopted a Khoa language, viz. Khoekhoe. And Khwe/Masarwa are Negroid hunter-gatherer groups that speak other Khoe languages. These are all marked cases of “San” speaking “Khoi” languages that do not affect the generalization.

In more recent post Bantu expansion time (less than 1500 years ago) favourable spatiotemporal conditions for language contact may have facilitated borrowing from Khoisan into Bantu (through language contact, sometimes through hlonipha, a special register of avoidance speech by Khoisan brides to Bantu communities). Reborrowing from Bantu into Khoisan may also have been a possibility (Nguni Bantu into |Xegwi, or, perhaps language shift with click retention in the case of Dahalo (Güldemann, 2007; Sands and Güldemann, 2009). However, these more recent “areal” features of clicks do not invalidate the genealogically Khoisan primacy of clicks discussed here. In more remote pre Bantu expansion time, there may have been several separation events within the first human lineage that diverged before externalization of language (speech) set in or became successful. In these conditions language contact may have facilitated spread of clicks. But this would have been an areal phenomenon restricted to genetically Khoisan and descendants from the first humans to diverge from the rest of us.

3. A puzzle for evolution of language

Let us turn to some relevant genomic and archaeological evidence that will be used later to support the asymmetry of the interface mappings: the generative procedure that builds hierarchically structured and maps it to the conceptual-intentional interface (CI) in a computationally efficient manner vs. the operations of the sensorimotor interface (SM), which process linearized structures that frequently pose problems for parsing. First, a large recent genomic study among Kho and San groups reveals that these groups are descendants of the earliest diversification event in human evolutionary history (Schlebusch et al., 2012), confirming earlier results (Vigilant et al., 1989; Knight et al., 2003; Tishkoff et al., 2007; Veeramah et al., 2012). There is firm evidence that the earliest divergence of human lineages, ancestral Kho-San and Niger-Kordofanian populations, occurred somewhere around 125 kya, or 121–128 kya across calibration times (Gronau et al., 2011). Second, taking symbolic behavior as a proxy for communicative language use, fixation of externalized language (speech/sign) must have emerged around, say 100 kya, but presumably some time before Blombos humans showed their presence 80 kya (Henshilwood et al., 2002; d’Errico et al., 2005; Mourre et al., 2010). At the latest, speech, that is communicative language behavior, must have been in place before the human dispersal Out of Africa took place around 60,000 years ago (Henn et al., 2012).

But this yields an apparent paradox, anticipated by theoretical analysis of the human language phenotype (Chomsky, 2011, 2013a, 2015) when applied to evolution of language (Chomsky 2010, 2013b; Berwick and Chomsky, 2011, 2016; Bolhuis et al., 2014; Hauser et al., 2014). Since the first diverging ancestral populations of anatomically modern humans gave rise to current populations capable of speech, possession of language must have preceded genetic divergence. On the other hand, behaviorally modern humans with a capacity for speech (accepting symbolic behavior as a proxy for communicative language) did not appear until much later. I.e. language both preceded and followed the first human lineage split. Diversification occurred long before the emergence of symbolic
behavior, a proxy for language, but, at the same time, possession of language must have emerged in the ancestral population some time before the divergence event: A puzzle.

4. The language phenotype

Human language is a biological system that allows for the generation of a digitally infinite array of hierarchically structured expressions that receive definite and systematic interpretations at the sensorimotor and conceptual-intentional interfaces (Basic Principle). In order to account for the novel and biologically isolate position of human language, its recent and sudden emergence, presumably without much (external) selectional pressure, and the apparent absence of any significant mutation since, the recursive operation yielding discrete infinity is likely to have been the result of a few simple mutations, essentially subject only to general laws of nature. As a matter of logic, the Basic Principle must have been in place some time before the problem of using it in externalized language could be seriously addressed. What is externalized is the hierarchical expressions that are internally generated by a simple combinatorial operation Merge, conforming to principles of minimal computation. Without Merge there is no externalized language (but, significantly, the converse case is false). Merge must have preceded speech in evolution of language. Speech is therefore an ancillary property of language, reflecting the demands that a sensorimotor system imposes on hierarchically structured language. In order to get externalized, hierarchical structure must be “flattened” (linearized) to pass through a sensorimotor filter that only allows for linear order. It follows that since linearization applies to hierarchical structure, possession of internal language must have preceded externalization. As a consequence, the computations that map syntactic structure to the interfaces show a deep asymmetry: the narrow syntax mapping to CI is blind to linear order and relies on structural properties only while the mapping to SM is also sensitive to linear properties. Externalization is therefore secondary both in linguistic function and in evolutionary time.

Interestingly, prioritization of the basic property of language over externalized use of language may solve an apparent paradox explaining why clicks used phonemically occur exclusively in languages spoken in southern Africa. Ancestral San were the first diverging human lineage to split off from the rest after the onset of the mutational events that led to merge-based language but prior to adopting phonemic clicks as an element of externalization of language. This way clicks came to be associated uniquely and exclusively with a genetically and geographically well-defined group of populations. The unique distribution of clicks, genealogically, geographically and genetically, thus provides empirical support for the idea, fully complying with the Strong Minimalist Thesis, that internally generated language preceded externalized language (Chomsky, 2010).

The paradox observed above thus simply disappears when the ambiguity in the use of the word “language” is eliminated. Externalized language, speech, must be sharply distinguished from internalized language, that is structured thought and the result of a computational operation, called Merge, satisfying the basic property of language which provides a discrete infinity array of hierarchically structured expressions with systematic interpretations at the cognitive and sensorimotor interfaces (Chomsky 2011, 2013b; Everaert et al., 2015). Evolutionary analysis can therefore be reasonably restricted to this phenotypic property of natural language, bridging the ancestral non-syntactic and modern syntactic states of the language capacity, and its articulation, a process that is ancillary to the core internal processes of language structure and semantic interpretation (Chomsky 2010, 2011, 2013b; Berwick and Chomsky, 2011, 2016). Furthermore, evidence is accumulating (Somel et al., 2013; Prüfer et al., 2014; Pääbo, 2014) that just a few but critical brain-related developmental genes (e.g. changes in regulatory elements around FOXP2, CNTNAP2, MEF2A) could have made the difference between us and Neanderthals, causing cortical reorganization and rewiring of the brain that gave rise to the basic property of syntactic language “within the short time space between the human–Neanderthal split and the emergence of modern humans” (Somel et al., 2013). In addition, recent genomic studies seem to support the position that Neanderthal/Denisovan were a different species with different cognitive capacities (Vernot et al., 2016; Mendez et al., 2016; Sankararaman et al., 2016).

Finally, human brain growth results in a markedly globular shape, related to the volume of the frontal cortex, and differs from brain growth in Neandertals, who developed a larger occipital bun (Guzn et al., 2010). Apparently, Neandertals expanded their visual processing systems at the expense of the frontal cortex. Plausibly Neandertals did not share the cognitive style of anatomically modern humans, being visually, rather than frontal-cortex oriented, and therefore probably lacking symbolism and modern language (Pearce et al., 2013).

5. Asymmetric mappings to the interfaces

It’s important therefore to reach a better understanding of what is meant by language as a computational system that is part of human biology and the use of that system for communication in externalized speech (Chomsky, 1965). The Basic Principle of language somehow employs a combinatorial operation, Merge, that is recursively applied and generates a discrete infinity of linguistic objects with systematic interpretations at the semantic and sensorimotor interfaces. It is generally understood that the operation Merge is uniquely linguistic and uniquely human (Hauser et al., 2002).

Combinatorial Merge is a binary operation that takes two syntactic objects, X and Y, and constructs from them a single new syntactic object Z (Chomsky, 2013a,b). X,Y may be lexical atoms drawn from the lexicon or already constructed objects. Neither X nor Y is modified, and no arrangement, linear order or otherwise, is imposed on them by the operation. Merge (X,Y) thus forms the set (X, Y), containing just X and Y, and allows only two logical possibilities for combining X and Y and constructing a new object from them. Either X and Y are distinct and neither is contained in the other, e.g. reads and a book. Applying Merge to these elements yields a new linguistic object {read, a book}, say the phrase “read a book.” The combining operation here is called “External Merge.” Or else, either X or Y is contained in the other, e.g. which book and John read which book. Application of Merge to John read which book, a structure already constructed, and which book, a term internal to that structure, will construct the new object which book, John read which book”, say the phrase “which book John read which book?” The combining operation in this case is called “Internal Merge.” Roughly speaking, and abstracting away from effects of linear order and labeling, these two application modes of simplest Merge reformatulate defining properties of earlier Phrase-Structure Grammar (PSG) and Transformational-Generative Grammar (TGG), achieving unification of these systems under binary Merge (Chomsky 2013a).

Simplest Merge conserves previously constructed linguistic objects and therefore there will be two copies of which book. When mapped to sound the structures get linearized for externalization as read a book and (Guess) which book John read, with only one copy externalized in the latter. Pronouncing only one copy follows from another general efficiency condition: minimize phonological computation, but leads to filler-gap problems that pose significant complications for use of language, particularly parsing and per-
5.1. The structural position of verb-second

Consider some properties of verb-second, a rule of grammar that operates in many languages and is illustrated below for Dutch. Verb-second applies to root clauses but hardly ever places the finite verb in a strictly second position in the sentence. In the complementizer-initial dependent clauses of (1) the finite verb is in sentence-final position. In the corresponding root clauses of (2) the finite verb forms are found in sixth and third position, but not in second position. Nevertheless, the root’s name is not a misnomer. It correctly captures linguistic intuition and characterizes a structural aspect of knowledge of language that is linguistically more realistic than the linear properties of the sequence of wordlike elements that actually meets the eye.

\begin{verbatim}
(1) a. . . . dat [die vazen uit de oudheid] verkleuren
that those vases from the antiquity fade away
(“that those vases from antiquity fade away”)

b. . . . dat die vazen [uit de oudheid dateren]
that those vases from the antiquity date
(“that those vases date from antiquity”)

(2) a. [die vazen uit de oudheid] verkleuren
[cf. “die verkleuren vazen uit de oudheid”]

b. [die vazen] dateren uit de oudheid
[cf. “die dateren vazen uit de oudheid”]
\end{verbatim}

Strictly speaking, the finite verb is never in a linearly second position, but always in a structurally second position. Apparently, the rule positioning the verb depends on linguistic structure and is blind to linear order. Here, the string “uit de oudheid” acts as a hierarchically constructed unit, building a noun phrase with “die vazen” in (1a), viz. “die vazen uit de oudheid,” but building a verb phrase with “dateren” in (1b), viz. “uit de oudheid dateren.” The results are still more complex hierarchically structured constituents. In each case the finite verb immediately follows the first constituent, a structural construct, rather than the first word, a linear construct. This asymmetry of structural properties (relevant for the semantic interfaces) and linear properties (relevant for the sensorimotor interfaces) is a universal property of human language, structural conditions not being restricted to just this case but extending to every construction and to every language.

Verb-second illustrated here for Dutch is not just limited to Germanic languages (den Besten, 1983) but is a trait of various languages from different language families. The rule manifests itself in a slightly different form in languages that are widely separated in geographical and genealogical terms. In Icelandic (an Indo-European language), verb-second applies to tensed verbs as in Dutch or German; in the Australian language Warlpiri (Pama-Nyungan family), it applies to tensed auxiliaries (Legate, 2008), roughly as in English, and in the southern African language Khoekhoe (Khoisan), it applies to just an agreement factor (Hagman, 1977; Huybregts, 2003). Evidently, the element shared in all of these languages, a tense element associated with agreement features, must be there for a principled reason, presumably for typing propositional structure, a precondition for semantic processing at the CI interface. If it turns out to be correct, this result, the enforced labeling of syntactically derived phrases, is yet another aspect of the computational nature of human language.

Labeling of syntactic structure is required by the semantic interface as is elegantly demonstrated by classic examples like Flying planes can be dangerous or Visiting relatives can be a nuisance (Chomsky, 1957, 1965). The latter can be taken to mean “To visit relatives can be a nuisance” or “Relatives paying a visit can be a nuisance.” The element “visiting” must be labeled as a verbal noun in Visiting relatives is a nuisance (“to visit relatives”) but as a verbal participle in Visiting relatives are a nuisance (“relatives who are visiting”). The verbal noun “visiting” projects its label to the phrase “visiting relatives” in the former case, the noun “relatives” does so in the latter. This way the different projections required for compositional semantics (CI) automatically account for the different agreement patterns (“is” vs. “are”) in overt morphology (SM).

Syntax provides the hierarchical structure and labels that are necessary for compositional meaning and sufficient for morphological spell-out.

The labeling algorithm applies deterministically. The head-complement structure {reads, a book} exemplifies the general case {X, YP}, where minimal search detects the label of X before it sees the label of Y in YP. So “reads a book” has the basic structure {reads, {a book}} = {V, NP} and is labelable as a verbal phrase. In contrast to these endocentric structures, matters are different for exocentric constructions {XP, YP} like subject-predicate constructions or small clauses, which result from every instance of internal merge (small clauses) or some application of external merge (subject-predicate constructions). Since labeling must be deterministic, and unambiguous labeling is undefined for symmetric structures {XP, YP}, these linguistic objects are unlabelable unless each of X and Y happens to carry “identical” prominent labels, identity resulting from agreement. A simple example of such a construction is the sentence The man reads a book, which has the structure {NP, {V, NP}}. Simplifying matters considerably but preserving the basic insights, we can say that in the sentence structure {NP, VP} – e.g. “The man reads a book” – both NP and VP share the same prominent agreement feature [agr: 3.sg], and the labeling algorithm, constrained by minimal search, happens to detect the same agreement feature φ twice, once on the head N of NP, and once on the head V of VP. The structure is therefore labelable as a kind of agreement phrase, e.g. < φ, φ >. Labeling of exocentric {XP, YP}, a construct of internal merge, where each of XP and YP shares the same prominent labeling feature as a result of a previous agreement operation, is called “labeling by criterial agreement.”

Root clauses with Verb-Second receive analogous analyses as exocentric constructions that result from internal merge and must be labeled to meet conditions of full interpretation. In particular, roots must be typed for illocutionary force (interrogative, declarative, exclamative, etc.), and enabled to interface with discourse structure. Since root clauses lack an overt complementizer, hence a distinctive head, typing of the clause remains problematic unless further derivation makes it analyzable as an exocentric structure with criterial agreement.

There is an obvious way to proceed. Choice of complementizer C generally dictates choice of tense element T, thus constituting a C-T system. The tense element T covaries with the complementizer C selecting it. More specifically, root complementizers select verbs inflected for tense-agreement (Dutch), tense/aspect/mood clitics (Warlpiri), or subject agreement affixes (Khoekhoe). If complementizer C is null, C requires application of a head-raising operation that merges a tense element with C (Verb-Second): a finite verb in Dutch (den Besten, 1983), or an auxiliary (a complementizer clitic) in Warlpiri (Legate, 2008). Restricting discussion to Dutch, since a finite verb in complementizer position yields a Yes-No question interpretation, a further basic operation is required, merging NP (internal to CP) and CP. The resulting exocentric structure {NP, {C,
the contrast between verb-second type languages (Dutch or German) and residual verb-second type languages (English) finds a correlate in another parametric difference, viz. the licensing of a null complementizer in English root and dependent clauses that is missing in Dutch or German. In the latter a null complementizer must be filled with a tensed verb both in root and dependent clauses. Compare German (3a) with root and embedded verb-second to its English equivalent (3b) with null complementizers.

(3) a. Dieses Buch gelesen ist [ich nicht hat [er jemals]]
   b. This book [C [I don’t believe [C [he has ever read]]]]

Explaining the correlation and the contrast between German and English type languages is beyond the scope of this article but may plausibly be related to a bounding property of the C-T system, specifically “inherence/dephasing” of C by T in modern terminology (Chomsky, 2015), that has morpho-syntactic reflexes (e.g. complementizer agreement) in the dialects of these languages.

Summarizing, the answer to the What question (What characterizes the Verb-second phenomenon?) is that root sentences have a tensed verb in a structurally second position. The need of satisfying labeling requirements for full interpretation explains the Why question (Why is application of the rule that yields verb-second obligatory?). And, finally, the How question (How does the operation proceed to give the correct results?) is answered by internal merge of XP to CP in order to feed critical agreement for labeling. Language design is therefore overwhelmingly computational rather than communicative, with the structural factors of semantic systems taking primacy over the linear properties of sensorimotor systems. The core of language, its Basic Principle, is a biologically isolated system that can be used for communication but cannot be identified with it. In verb-second constructions, internal merge thus derives two effects. First, it creates the structural conditions that lead to a “second position” linearization of the tensed verb required by SM, and second, it creates the structural conditions for successful labeling required by CI. The conclusion to be drawn from all this is that language is essentially designed as an instrument for internal thought, externalized speech being an ancillary aspect of it with properties required by sensorimotor conditions (Chomsky, 2010, 2013a,b, 2017).

5.2. Anaphoric binding

Another case illustrating this asymmetry of the mappings to the interfaces is the reflexive sentence (4a). Here the reflexive pronoun himself must have who as its antecedent even though Bill is closer to it. This is unexpected and surprising on simple grounds of “analogy” and the fact that linear distance is a computationally simpler property than hierarchical structure. In both (4a) and (4b) exactly the same sequence of words “want to help” intervenes between Bill and himself, but apparently, only in (4a) is there a structural factor blocking Bill from binding the reflexive. What is the difference between these two cases that are superficially similar? In the case of (4a) the distance between Bill and himself is minimal only when measured in linear terms, but in structural terms who turns out to be closer to the reflexive element than Bill is. Why is this?

(4)

a. Who did Bill want [to help himself]?
   b. Did Bill want [to help himself]?

Free application of merge will take two linguistic objects already constructed, the clausal structure (5a) and who, the latter a term of the former, and constructs from these a new linguistic object, which will give rise to (5c). Internal who of (5a) is thus “raised” to the root of the sentence. This operation of internal merge conserves structure already derived. This is a consequence of simplest merge. Since simplest merge will not add anything and will not modify or order the terms it applies to, application of internal Merge to who and the clause containing it yields a new construct that contains both terms as its constituting elements. The copy of who left behind falls out from the structure preserving properties of simple merge applying to structures, not strings. Finally, linearizing this arrangement will lead to the structure shown in (5c) prior to its externalization.

(5)

a. Did Bill want [who to help himself]?
   b. Did Bill want [John to help himself]
   c. Who [did Bill want [who to help himself]]

Without need for any further stipulations, simple merge automatically yields the structure (5c) with the two copies of who, each making its own contribution to the semantics. The “displaced” who is interpreted as an operator in sentence-initial position and receives another interpretation elsewhere as the agentic subject of the embedded clause, the variable bound by displaced who. For comparison see the derivation of (5b), which lacks displacement. The “duality of semantics” simply follows from the discontinuous operator-variable factor (a result of internal merge) and the contiguous thematic function factor (a result of external merge). The representation is perfect for semantic interpretation at the CI interface but its externalization at the SM interface, Who did Bill want to help himself?, with the lowest copy deleted (a consequence of another efficiency condition of minimal computation) creates filler-gap problems for communicative efficiency (Berwick and Chomsky, 2011, 2016; Chomsky, 2013b; this issue). This illustrates another aspect of the asymmetry of the syntactic mappings to the interfaces. Language is “designed” as a system for constructing complex thought, not as a system for efficient communication. Computational conditions, not communicative conditions, constrain and shape linguistic form. As is evident, given the derivation proposed, the problematic case of (4a) is simply reduced to the unproblematic case of (5b), once we adopt an analysis with rules applying to structural positions and constrained by minimal structural distance. Minimal structural distance rather than minimal linear distance is what differentiates between (4a) and (4b) and unifies (4a) with (5b). Anaphoric binding of the reflexive pronoun in (4a) proceeds in precisely the same way as in (5b) given the structure (5c) in accord with the binding theory (Chomsky, 1981).

There is further empirical support for an analysis based on structural dependencies that is not expected on an analysis that is based on rigid dependencies of string elements. Why is contraction to wanna (Chomsky, 1976; Lightfoot, 1976; Chomsky and Lasnik, 1977; Everaert et al., 2015) legitimate in (6a) but blocked in (6b)?

(6)

a. Did Bill wanna help himself
   b. * Who did Bill wanna help himself (of 5c)
   c. * Did Bill wanna John help himself (of 5b)

In cases like these, contraction is blocked if and only if binding of the reflexive pronoun to the matrix subject is ruled out. Contraction and reflexivization are correlated phenomena that are causally related through an application of simplest Merge that leaves behind a copy of the internally merged element. The copy gives rise to an intervention effect which blocks contraction and additionally explains strict locality of reflexive binding. These examples again illustrate the computational nature of language: structure depen-
idence of rules that are governed by efficiency principles of minimal search (binding) and minimal computation (copy-deletion). The ill-formed contraction of (6b) becomes reducible to the case of (6c). In each of these cases there is a structural element in a position that is structurally higher than the tense particle “to” but lower than the verb “want” (Everaert et al., 2015). So there is an intervening structural factor in force, i.e. the lexical element “John” in (5b) and the invisible copy of “who” in (5c), blocking contraction (6b,c) and reflexive binding outside the embedded clause (4a). A standard case of “reducing complex visible to simple invisibles,” citing Nobel laureate in physics Jean Baptiste Perrin. The linear order is derivative and merely reflects hierarchical structure: roughly, element X asymmetrically c-commanding element Y is linearized as X preceding Y (Kayne, 1994).

Once again, language ignores the computationally simpler property of linear precedence and relies on the more complex property of hierarchical structure, a result that has often been seriously misunderstood (Chomsky, 2011; Berwick et al., 2011; and the references cited therein). Specifically, binding obeys structural conditions (answering the What question), and employs the copy results of internal merge (answering the How question), allowing minimal search to finally explain the strict locality effect (answering the Why question). In every natural language it is conditions on structural position and structural distance, not linear position or linear distance, that dictate rule applicability.

5.3. Bound variable interpretation

As a final illustration, consider bound variable interpretation. Sentences like The idea that every son will admire his father worried Batman vs. The idea that every son will admire Batman worried his father illustrate a universal principle that variables (“this”) must be bound to a quantifying expression (“every son”) in a structurally higher position rather than to an element in a linearly preceding position (Reinhart, 1981; Everaert et al., 2015). Batman may be the antecedent of his in both sentences but every son can bind the pronoun in the former sentence only. Only in this case is every son structurally higher than his father. The conclusion, once again, is that structural conditions prevail over linear precedence. Linear order is irrelevant (in both sentences the quantifying expression precedes the pronoun).

There is additional supporting evidence from Malagasy, a VOS language spoken in Madagascar (Reinhart, 1981). Consider the sentences of (7) with intended coreference between pronoun and Rakoto.

![Image of sentence (7)]

The pronominal precedes the antecedent in (7a) but follows it in (7b). In terms of linear precedence the well-formedness conditions for Malagasy and English are reversed: the pronoun preceding its antecedent yields ungrammaticality in Malagasy but not in English; conversely, the pronoun following its antecedent leads to ungrammaticality in Malagasy but not in English. However, pronominal binding in Malagasy and English can be unified in terms of hierarchical structure, technically c(onsituent)-command. A pronoun cannot occupy a structurally higher position than the antecedent. In both languages, the pronominal is a constituent of a verb phrase that does not include the antecedent in (7a), and therefore is not in a hierarchically higher structural position than its antecedent: coreference is not blocked. Also, in both languages, the antecedent is a constituent of a verb phrase that does not include the pronominal in (7b), and occupies a hierarchically lower structural position than the pronominal: coreference is blocked. Malagasy and English uniformly conform to binding theory (Chomsky, 1981) in precisely the same way. In brief, reference to linear order (SM) only leads to contradictions or loss of linguistic generalizations while reliance on hierarchical structure (CI) yields consistency and linguistic unification.

It should be borne in mind that we have not been discussing here pronominal reference per se but only binding conditions on anaphora and bound variable interpretation (binding of a pronominal by a quantifier expression). In both these cases c-command is strictly relevant. It is true that the reference problem of anaphora has been shown to be Non-deterministic Polynomial-hard (NP-hard), and can therefore not receive a solution in deterministic polynomial time (Ristad, 1993). But the problem of bound variable interpretation and validation of the conditions of binding theory both involve c-command and can be solved in deterministic polynomial time. In fact, the relevant c-command condition reduces the search space logarithmically, from exponential size to just linear size (Berwick and Waxler, 1982).

5.4. Linear order: not even wrong!

Viewing language not as an unbounded array of hierarchically structured expressions but as a set of linearly ordered strings of wordlike elements is highly problematic. One further problem is that mere transitional probabilities between these elements do not suffice for even weakly characterizing language. Arguments have been presented involving unbounded dependencies that prove that natural language is neither finite state (Chomsky, 1956) nor context-free (Huybregt, 1984) but mildly context-sensitive (Stabler, 2013). Furthermore, probabilistic models of language that take language to be linearly ordered stringsets leave all of the Why-questions unanswered. Consider the following general case. Let (A)BCD and BCD/E but not ABCDE be grammatical strings in the language. Short of listing, no bigrams or any other device for characterizing locally testable string sets precisely will be able to account for these data. But even listing would be impossible if there is no non-arbitrary limit on the linear distance “X” between the relevant items “A” and “B” in A-X-BCD. A finite state model would be able to distinguish “A-X-BCD”/“X-BCDE” from “A-X-BCDE,” but it can do so only by brute force, essentially by incorporating two finite state grammars, one for each subset A-X-BCD and X-BCDE.

It would solve the How-question by ignoring the Why-question. Natural language provides ample examples of this kind. Substitute “John can teach” for “BCD”, “French” for “A” and “Chinese” for “E”. Finally, let “X” range over any well-formed indefinitely extendible substring that can be followed by “John can teach,” e.g. “I think (that) they believe (that) . . . ” The result is the difference between, on the one hand, French I think they believe that John can teach (a well-formed filler-gap construction) or I think they believe that John can teach Chinese (a well-formed sentence without a gap), and, on the other hand, the ill-formed sentence French I think they believe that John can teach Chinese (the gap left by filler “French” is plugged here by “Chinese”). In order to appropriately characterize language we need make use of a recursive operation that applies to linguistic elements, primitive or already constructed, and yields hierarchical phrase structure. Internal Merge, one way of applying the simplest such operation Merge, appears to be an indisputable factor in accounting for filler-gap phenomena like these. Merge, both internal and external, constitutes an essential element of the Basic Principle of language that yields discrete infinity, unbounded dependencies, reconstruction effects, and many more properties characteristic of human language.

These cases suffice to show that computations of the mind rely on the structural organization of phrases but are blind to the linear organization of words that are articulated and perceived by
input and output systems at the sensorimotor interface. This is a significant finding since the universally adopted computational procedure that depends on hierarchical structure is computationally much more complex than an alternative that relies on linear order. Nevertheless, linear order is not available to the systems of syntax and semantics. It is an ancillary feature of language, probably a reflex of properties of the sensorimotor system that requires it for externalization, and constrained by conditions imposed by sensorimotor modalities. Furthermore, it follows that language is primarily an instrument for the expression of internal thought. Essentially, language is neither speech nor sign (but it can be externalized in these different modalities) nor communication (which is only one of its many possible uses).

6. Capacity for language and the use of it

Language is a computational system internal to the human mind that can (but need not) be used for communication. In principle, language systems could or could not be computationally or communicatively efficient. There are therefore four different cases to consider. The cases where computational efficiency and communicative efficiency converge can be effectively put aside since they will not tell us anything revealing. However, the conflicting cases are interesting.

- To first illustrate the uninformative converging cases of efficiency we may consider bound variable interpretation. Finding a quantifying expression that can serve as a legitimate binder for a pronounal variable is communicatively useful but a hard computational problem to solve without use of hierarchical structure, as shown by the contrasting sentences The idea that every son will admire his father worried Batman vs. The idea that every son will admire Batman worried his father. The quantifier phrase every son can bind the pronoun his only when it is in a structurally higher position than his father. Here the first sentence. Since the search space for finding pairs of quantifiers and bound variables is logarithmically reduced (Berwick and Waxler, 1982) by this supremacy restriction (technically c-command), itself a consequence of the basic combinatorial operation Merge (Epstein, 1999), the constraint is at the same time both computationally motivated and communicatively desirable.

- On the other hand, systems that exemplify convergence of communicative and computational inefficiency go unattested as may be expected. An example may be an artificially designed language requiring a verb-second linear position in the string for sentences with an even number of words but a pre-final position for the verb in sentences with an uneven number of words: John Slept, The slept man, That climbs bay mountains, every boy his likes father. Such systems would be communicatively demanding (co-opting additional numerical complexity) as well as computationally inefficient (violating structural simplicity and constituency restrictions). Therefore, both computational and communicative efficiency would rule out such systems as possible human languages.

- The two residual cases of conflicting evidence are more informative. Here computational efficiency prevails over ease of communication. Examples are the ubiquitous instances of structural ambiguity, problems of parsing and processing language (filler-gap problems, garden-path sentences, blindspots, self-embedding, etc.), syntactic islands and similar constructions, and more generally every single case showing asymmetry of interface mappings. Problems for communication arise when rules are allowed to apply freely but lead to inefficient use of language, e.g. self-embedding or structural ambiguity, or alternatively block derivations that would otherwise have communicative use, e.g.

"Who did you wonder when arrived? or "Mary married a lawyer who is richer than her sister did (intended meaning “Mary married a richer lawyer than her sister did”). On the other hand, cases where communicative ease prevails over efficiency of computation do not seem to exist (Musso et al., 2003).

The result that syntactic islands, which otherwise would be fine thoughts with appropriate communicative uses, are not derivable from merge-based computation does not argue against the conclusion that language is optimally designed for the expression of complex thought. Mapping to the semantic interface conforms to simple merge and minimal computation, syntactic islands being excluded strictly for reasons of computational efficiency and labeling. Communication is sacrificed in favor of simplicity of computation.

It might be suggested that case and agreement systems in language facilitate communication ("Who is doing what to whom?") and complicate computation but that suggestion would be misguided as careful analysis shows. Case assignment in natural language has been argued on conceptual and empirical grounds to be a reflection of a successful agreement operation in grammar (Chomsky, 2001; Huybregts, 2010), but agreement itself is a necessary step in labeling exocentric constructions, roughly all instances of internal merge application. In fact Case-Agreement turns out to be an essential factor in a computational system based on Simple Merge and Minimal Search, and is only an apparent counterexample to computational efficiency prevailing over facilitation of communication. Case and agreement are neutral between communicative and computational simplicity, and reduce to convergence already discussed. To conclude, it looks as if cases where communicative efficiency wins over efficiency of computation simply do not exist. Core properties of natural language expressions seem to derive from underlying computational principles that correctly explain them. In contrast, thus far no structural property of the language phenotype seems to follow from communication theory.

7. Dissolving the paradox

The conclusion must be that language and speech must be sharply distinguished. Each has its own distinct properties that are sometimes mutually conflicting. At the same time, disambiguation of the informal term “language” leads to a coherent reformulation of the apparent paradox, which simple dissolves: Merge (i.e. internal language) preceded but externalized language (i.e. speech) followed the first population split. It follows that externalization is secondary both in function (i.e. “less than perfect” externalization of “perfect” internal language) and in evolutionary time (i.e. externalization follows the diversification of language-supporting lineages by several decades). With regard to the latter, the exclusive association of phonemic click systems with Khoisan languages provides further empirical support for the position that language is primarily a system for expressing thought (Chomsky, 2010, 2011, 2013b; Berwick and Chomsky, 2011; Tattersall, 2012).

However, the views of Chomsky and Tattersall differ in at least one crucial respect (Fig. 1). We need to be cautious here since Tattersall does not explicitly define the language phenotype. For Tattersall “symbolic thought” evolved as a language-external property, and language was “invented” subsequently to externalize it (i.e. strictly speaking there was no evolution of language). For Chomsky the merge-based capacity for language (UG) evolved and externalization may not have required much evolution (i.e. is strictly speaking ancillary to language and may even have been partially a cultural event). UG makes available hierarchically structured expressions that are interpreted at CI as structured thoughts. Consequently, for Chomsky, externalization...
capacity for symbolic thought | capacity for language (Basic Property) | Merge (CI interface) | step 1
---|---|---|---

invention of language | solving externalization problem | Speech (SM interface) | step 2

**Fig. 1.** Comparing Chomsky and Tattersall. Chomsky’s capacity for language is the Basic Property (UC) that yields discrete infinity of hierarchically structured thought. Externalization is an ancillary property. Tattersall’s capacity for symbolic thought is language-independent and evolved before language was “invented” to express thought. Structured thought is language-internal for Chomsky but language-external for Tattersall.

(but not merge-generated structured thought) is in a relevant sense language-external; for Tattersall, structured thought (but not linearized speech) is language-external. There is convergence on the primacy of thought over speech but divergence on how structured thought relates to language.

A possible and reasonable scenario. The Merge-relevant mutational event must have happened some time after anatomically modern humans appeared in the evolutionary record around 175–200 kya. Once the merge mutation was “drawn out of the realm of pure chance, the accident enter[ed] into the realm of necessity” (Monod, 1972), selective pressure (differential rate of reproduction) being exerted on populations, not on isolated individuals. The mutations that caused rewiring of the brain and yielded the basic principle of language took place in a “language-ready” brain in a small sub-Saharan population of hunter-gatherers some time prior to the first population separations around ~125 kya (Gronau et al., 2011). The mutation(s) then slowly spread in the population but *without reaching fixation before the first diversification event took place*. Next, after this population split into Khoe-San and Niger-Kordofanian lineages, the mutation which yielded Merge and had already started to spread in the ancestral populations, now became fixed in each lineage, some time around 100 kya before the Blombos proxies for language made their appearance.

Finally, the problem of externalizing internal language received diverse, complex, malleable and separate solutions after fixation of Merge in each lineage. According to Ambrose (1998), modern humans spread into separate regions from a restricted source, around 100 kya, but without significant population expansion, and then passed through severe bottlenecks with subsequent dramatic growth within dispersed populations that were genetically isolated from each other. In one of these regions of southern Africa or eastern Africa the ancestors of present-day San found a solution to linearizing their internal language that may have included extensive use of phonemic click sounds. From a number of different but conspiring reasons, physiological (weak bias for click articulation), demographical (small sized populations), geographical (restricted and isolated refugia), genetic (dispersed populations that were genetically isolated from each other), the areality and isolation of phonemic clicks may be explained.

Speech thus became an empirical reality some short time before anatomically and behaviorally modern Blombos humans appeared with the capacities for language and the use of it fully in place (Henshilwood et al., 2002; d’Errico et al., 2005).

We may distinguish four relevant scenarios for evolution of language and speech (Fig. 2).

(i) Both M and E prior to D1 (M, E < D1)
(ii) M prior to D1 and E subsequent to D1 but prior to D2a (M < D1; D1 < E < D2a)
(iii) M prior to D1 but E subsequent to D2a (M < D1; E > D2a)
(iv) Both M and E subsequent to D1 (M, E > D1)

**Fig. 2.** Relevant scenarios for evolution of language and speech. Legenda: “AMH” stands for anatomically modern human, “CMH” for cognitively modern human with a capacity for language; “M” represents the mutational events for recursive l-grammar, “E” represents subsequent developments that led to externalized language, specifically “E/Clicks” vs. “E-Clicks” denoting externalization with vs. without clicks; “D1” signifies the deep San/Yoruba divergence, and “D2a”/“D2b” the subsequent diversification events within Khoisan and non-Khoisan populations respectively.

Commentary: Assuming the mutation did not happen many times, independently in distinct human subpopulations, M must have preceded D1 since all descendants of separated populations have language. However, E could not have preceded D1 since possession of externalized language (speech) would have implied behavioral modernity, specifically symbolic behavior, absent from the archaeological record. A primary result is then that M preceded E; merge-based internal language must have preceded the first separation, which in turn preceded externalized language, explaining the asymmetry of the interface mappings. Secondly, phonetic clicks may have been recruited in the language systems of San populations as part of externalization between the first separation of human populations and the San/Hadza divergence (strong version). Alternatively, they may have emerged in San populations after separation from Hadza (weak version). In the latter scenario, Hadza clicks may have been borrowed from San. Either way phonetic clicks were restricted to human groups plausibly defined in genetic, geographical and linguistic terms and emerged after the first divergence of human populations.

Scenarios (i) and (iv) are problematic. Option (i) is ruled out by the unexpected absence of early behavioral modernity (increased symbolic behavior); option (iv) is ruled out by the exceedingly low probability of a single mutational event happening multiply and independently in different populations. Scenarios (ii) and (iii) are stronger and weaker variants respectively of the evolutionary trajectory suggested here. Regardless option (ii), the emergence of clicks between the San–Yoruba split and the subsequent San–internal diversifications fits in perfectly with the genetic results of Gronau et al. (2011) and Knight et al. (2003). Option (iii) is just a weaker version of option (ii): clicks may have been the result of later language change, though still restricted to a genetically defined set of populations with strong areal contacts. The stronger version holds that clicks were original with ancestral Khoisan; the weaker variant holds that clicks were restricted to descendant Khoisan populations with areal contacts. In both cases clicks emerged in the oldest human lineage after the first divergence in human population history. The stronger version would place linearized language, and particularly the emergence of clicks after the San divergence time of ~125 kya (Gronau et al., 2011) and before the ancient Hadzabe/San separation, somewhere between 112 kya and 60 kya (Knight et al., 2003). This is precisely the window of time in which there occurred
a regulatory change, unique to modern humans, in an intron of the FOXP2 gene, implicated in speech (Vargha-Khadem et al., 1995; Lai et al., 2001), which affects a binding site for the transcription factor FOXP3 (Marcic et al., 2013). According to the authors, the derived FOXP3 variant (uniquely human and absent from both Neanderthals and Denisovans) can reasonably be assumed to “regulate expression of FOXP2 in neurons.” Possibly, the change was “positively selected recently during the evolution of fully modern humans.” Selection would be for improved articulated speech. See below for further discussion.

8. Some consequences

8.1. Asymmetry of interface mappings to meaning and sound

The scenario suggested makes it clear why “symbolic behavior” emerged much later than the Khoe-San vs. Niger-Kordofanian divergence. Fixation of the Merge mutation and subsequent externalization required the loss of phonic clicks. It first, with merge-based language being structure-sensitive but blind to linear order (Chomsky, 2013b, 2015; Everaert et al., 2015), then the post-Merge externalization of hierarchically structured expressions explains the asymmetric properties of the mappings to the interfaces (e.g. perfect computation of semantic properties vs. parsing problems for externalized language). Also, the asymmetry of the interface mappings is supportive of the notion of language as primarily a system of thought (Chomsky, 2010, 2011, 2013b; Berwick and Chomsky, 2011, 2016). The evolutionary trajectory of hierarchical language preceding linear speech strongly points to this interpretation.

8.2. Externalized language and phonemic clicks

It accounts for how a partial “solution” to the externalization problem for language could have been selected that is radically different from all others. The problem here is how to connect a novel and biologically isolated computational system, which has evolved recently and saltationally, to an ancient sensorimotor system that had already been in place, virtually unchanged, for ages (Chomsky, 2010, 2013b; Berwick and Chomsky, 2011, 2016). There may be multiple solutions. The sensorimotor system must decide which one of the various parametric options it is going to select for externalizing the hierarchically structured expressions of internal language. The ancestral Khoe-San, a group of hunter-gatherers from southern Africa, and the first human lineage to diverge from the rest of us, “idiomsymmetrically” adopted non-pulmonic airstream mechanisms for the production of phonemic clicks. Clicks became naturally “isolated,” restricted genealogically to the languages spoken by the oldest human lineage branching off, a biologically coherent and isolated group, and therefore also restricted to a genetically defined human lineage (Schlebusch et al., 2012), confirming earlier results (Vigilant et al., 1989; Knight et al., 2003; Tishkoff et al., 2007; Veeramah et al., 2012). As a result, we derive a natural account of the skewed geographical, genealogical and genetic distribution of clicks and their speaker populations.

Despite these “isolated” properties and apparently complex nature of click sounds (articulation with multiple closures and dual airstream mechanisms, pulmonic and non-pulmonic), phonemic click systems have turned out to be impressingly stable elements in human languages. There may be language-internal reasons why this is the case. The lexicons of Khoe languages are often characterized by high frequencies of lexical items with click phonemes. In general, consonants are preferentially involved in lexical processing, identifying and distinguishing lexical items (Nespors et al., 2003; Mehler et al., 2006; Toro et al., 2008). The rich inventories of click accommodations provide for extensive systems of phonemic clicks, and therefore make these click consonants champions of lexical processing. Furthermore, the phonotactics of clicks (they occur only word- or stem initially) may be quite useful for segmenting words in processing speech. “Central Khoisan” Khoekho has an inventory of 15 click consonants but |g|ana|-gui, other Kho languages, have 52 phonemic clicks. San languages frequently have even more. “Northern Khoisan” Ju|hoan, a Ju language, possesses 48 click consonants, but -hôa has at least 68 clicks. Finally, “Southern Khoisan” (Eastern) Xôo, a Tuu language, has as many as 83 click phonemes (Traill, 1985). Once in place they will not be easily displaced. Still, there is loss of clicks (Traill and Vossen, 1997). Interestingly, however, clicks are lost primarily in the periphery under specific conditions of language contact and borrowing (Güldemann, 2007; Güldemann and Stoneking, 2008).

8.3. Vocal tract morphology in San

It has long been observed that Khoisan speakers generally tend to have short smooth palates that may facilitate click production, providing them with a weak anatomical bias for click articulation that is absent from all other human populations, which have a prominent alveolar ridge (Traill, 1985; Traill and Vossen, 1997; Ladefoed and Maddieson, 1996; Van Reenen and Allen, 1987 Winkler and Kirchgang, 1993). In his study of the phonetics of !Xôo clicks, Traill quotes an early study observing that this feature is widespread in San populations. The feature allegedly “reduces the amount of distortion of the tongue that is required in producing clicks,” predisposing speakers for their production and employment in human speech (Traumüller, 2003). Consequently, San vocal tract morphology may have exerted a weak bias, absent elsewhere, for articulating click sounds (Moiski and Dediu, 2015; Dediu and Moiski, 2016). In particular, the lack of a prominent alveolar ridge size facilitates their production.

The anatomical bias is plausible as suggested by a recent biomechanical model of click production (Moiski and Dediu, 2015; Dediu and Moiski, 2016). According to these studies, a short smooth palate requires “less articulatory effort to form click structure” (i.e. less deformation of the anterior tongue to create vacuum) and provides “better volume change characteristics” (presumably necessary for achieving efficient aero-acoustic effects). They suggest that human vocal tract morphology may exert a bias for articulating sound that might find expression at the population level. For our purposes, the exclusive expression of bias in San groups is not different from expression of other characteristic traits like e.g. ears without earlobes, peppercorn hair, high cheekbones, epicantthic folds, lighter skin, etc. The production bias on articulation of clicks, which has a real basis in palatal morphology of San populations, is “weak at best” but real and suggestive. It suggests that clicks were “recruited” for externalization in San groups after they had branched off from all other ancestral human populations.

If a smooth palate turns out to be an ancestral trait, retained only in San, the prominent alveolar ridge must have become derived in all non-San populations after divergence. In this case, if phonemic clicks were used in externalized language, speech, prior to divergence, then since biasing on clicks is only a weak effect of having a smooth palate, some evidence for clicks, or traces of them, should be present in the rest of the world’s languages. Also, we would have expected to find significantly earlier instances of increased behavioral modernity. So there is a weak biomechanical signal of why clicks are original with – and restricted to – Khoisan populations, who first used them in externalizing their internally generated language. Phonemic clicks were part of their solution of the “externalization problem” of how to connect an evolutionarily ancient sensorimotor system, limited to expressing linear structure, with
an evolutionarily novel and biologically isolate system for generating hierarchical structure.

On the other hand, if the smooth palate of San can be shown to be a derived trait, it must have become derived in San after ancestral San had branched off, providing them with a weak bias for click articulation. The exclusive position of clicks with Khoisan would then simply follow. However, in this case phonemic clicks might have emerged later in evolution of language, and would support only a weak version of our argument that emergence of Merge preceded the deep separation between San and Yoruba/Bantu and externalized language followed it.

8.4. Emergence of clicks in externalized language

However, there is a way of arguing for a more specific time frame for the emergence of phonemic clicks. Genomic evidence suggests that the deep genetic divergence among click-speaking Hadzabe (Tanzania) and Ju|’hoansi San (Namibia) is among the earliest divergences in human evolution, dating back to a time before the human exodus from Africa 60 kya, and possibly much earlier, with 112 kya as an upper bound (Knight et al., 2003). Therefore, unless it is assumed that clicks developed independently in Hadzabe and Ju|’hoansi San after their separation, they must have been an ancient trait of human language shared by the most recent common ancestors of Hadzabe and San since before they diverged. The result can be usefully applied in an effort to establish a more precise date for the emergence of click phonemes in externalized language.

Since externalized language necessarily postdates internally merge-generated language, and externalization started to pick up momentum after the first separation of human populations, phonemic clicks must have developed after San diverged from Yoruba and Bantu groups around 125,000 years ago (Gronau et al., 2011). Next, since having a smooth palate, a San-specific feature of vocal tract morphology (Traill, 1985; Traill and Vossen, 1997; Ladefoged and Maddieson, 1996; Van Reenen and Allen, 1987 Winkler and Kirchengast, 1993) provides a weak bias for articulation of clicks, clicks may have been ancestral with speakers of Khoisan languages only. Finally, since there is an ancient genetic link between the click-speaking Hadzabe (East Africa) and Ju|’hoansi San (southern Africa) whose most recent shared ancestry may have coincided with one of the earliest population divergences in human prehistory (e.g. Knight et al., 2003), click phonemes must have emerged uniquely in Khoisan speakers as an element in the externalization of their language after San separated from all other human populations but before ancestral Hadzabe diverged from all other Khoisan lineages, in particular Ju|’hoansi San. The latter population divergence dates back to at least 60,000 years ago, and possibly to 112,000 years ago (e.g. Knight et al., 2003). Therefore, if we accept these two genomic results and the line of reasoning involving them, phonemic clicks must have emerged anytime between 125,000 (the time of the deepest human population divergence following the merge mutational events) and 60,000 years ago (the time of the exodus out of Africa), and arguably between 125,000 and 90,000 years ago (shortly before behavioral modernity and Blombos/Klasies humans made their appearance in human prehistory).

In Knight et al. (2003) it is suggested that clicks are an ancient element of human language, lost in all other populations but independently retained in ancestral populations of extant Hadzabe and Ju|’hoansi. This result must now be qualified to mean that only the most recent common ancestors of Hadzabe and Ju|’hoansi recruited clicks, and only did so after the Merge mutation had spread in human populations, in particular, after San had diverged early from extant Yoruba/Bantu. Click sounds may have had an ancient link with populations ancestral to Khoisan but they were not part of Proto-World, the mother tongue allegedly spoken by the first modern humans. Linguistically functional clicks have never been lost elsewhere simply because they have not been recruited anywhere else, and they were never part of Proto-World because there never was Proto-World. Speech developed independently in human populations with the capacity of language (Merge) after the first deep divergences among human lineages had happened. Still, we may take the genomic results of Knight and others to mean that the early emergence of the ancestors of click-speaking Hadza and Ju|’hoansi is evidence in support of early use of clicks. Specifically, clicks may have been developed within the narrow time frame bounded by the deep separation events discussed in Gronau et al. (2011) and Knight et al. (2003).

8.5. The illusion of Proto-World

Another consequence is that Proto-World is ruled out. There cannot have been any unique Proto-World ancestral to all human languages (extinct or alive). If fully externalized language developed after population diversification there must have been several Proto-World externalizations that independently emerged in various populations at (slightly) different times. Moreover, these languages may have been radically different in morphology, phonology and lexical roots, highlighting the fact that language is essentially a system for structured thought, invariant at the core but variable in properties relating to the sensorimotor interface. This result follows from the logic of the argument that the onset of recursive language, Merge and the Basic Principle, preceded, and externalized language followed, the first human population splits. By the time we left Africa all human populations possessed externalized language. Every language postdating the Out-of-Africa event must have descended from these Proto-World languages with modification (that is through processes of language change, not language evolution).

Accepting a population size ranging between 10,000 and 50,000 humans at the relevant time frame 125,000–100,000 ya, with an average of 500 speakers per language, there would not have been more than 20–100 languages being spoken at any one time. Further, assuming some 5–10 languages per language family, no more than a couple of dozen Proto-World candidates would have found externalizations from which all living languages are descendants. This outcome is rather different from the suggestion (Gell-Mann and Ruhlen, 2011) that “all or almost all attested human languages share a common origin” even though “that origin need not necessarily refer all of the way back to the time when behaviorally modern humans emerged.” They suggest this is because of the small window of time in which language emerged and the subsequent “bottleneck” effects that resulted in the few residual languages spoken then being “ancestral to all or most attested languages.” In fact, quite independent of bottleneck effects all or most attested languages could not have shared a single ancestral source on our view since there never was such a language. Instead following the argument laid out here that “merge” preceded “speech” there must have been several contemporaneous but independent ancestral languages being spoken at the time by the different populations that had branched off the ancestral lineage before externalized language had begun to set in. Therefore all living or dead languages must have derived from some or all of these independent externalizations, quite apart from any later bottlenecks. An unchanged UG, the capacity for language, would be ancestral to all internal languages but there could not have been a single externalized language from which every spoken language has descended with morphological or morpho-syntactic modification.

8.6. A failed human experiment with a twist

The proposed scenario may explain why Jebel Irhoud (7160 kya) and Dar-es-Soltan (120 kya) hominids of North-Africa and the
Mugharet-es-Shkuel/Tabaq Qafzeh (100 kya) hominids of the Levant—though early anatomically modern humans—were not behaviorally modern, still using Mousterian technology (Tattersall, 2012). Plausibly, these populations could not be behaviorally modern because they lacked the capacity for language. They were far from being cognitively modern but “may represent a population that diverged early from other modern humans in Africa” and later interbred with Levantine Neanderthals (~100,000 ya, consistent with “[t]he finding of ‘African’ haplotypes as young as 100,000 years old in the Altai Neanderthal genome” (Kuhlwilm et al., 2016).

In contrast, even before the first divergence of human lineages took effect, contemporaneous sub-Saharan populations were already evolving to cognitively modern humans with the basic property of language in place. They were capable of structured thought, reasoning, planning, making inferences, and the like but possibly had not yet developed and used the capacity to externalize their internal language for communicative purposes. In brief, the relevant mutation event that yielded Merge was taking place in various sub-Saharan populations at a time when the North-African hominids were extant but it was not to reach North-Africa until much later, and then it may have been too late. These North-African and Levantine archaic moderns probably ended as failed human experiments but not without leaving us some indirect evidence suggestive of recent selection for improved articulated speech. The indirect evidence involves particular elements of gene flow from human moderns into archaics. Modern human introgressed DNA segments in Altai Neanderthals (Kuhlwilm et al., 2016) specifically exclude a uniquely human change in a functionally important regulatory stretch of FOXP2, viz. POU3F2. This regulatory segment is very close to another segment within FOXP2 that did introgress from modern humans into Altai Neanderthals, the likely descendants of Neanderthal populations in the Levant or Southern Arabia 120,000–100,000 years ago (Kuhlwilm et al., 2016). According to this study the human population responsible for the introgressed segments in the Levantine ancestors of later Altai Neanderthals must have branched off from other modern human populations “before or shortly after the split between the ancestors of San and other Africans,” which the authors assume “occurred approximately 200,000 years ago” (Kuhlwilm et al., 2016). It follows that not only may these humans have lacked the capacity of language (the Basic Property) that yields recursive language at CI, but also the capacity for externalization of language at SM may not have been sufficiently developed for them to produce articulated speech.

8.7. Genomic results suggest late externalization of language

Recursive hierarchical structure presumably was not part of any language phenotype in Neanderthal. If modern humans had shared hierarchical syntax with Neanderthal, we would expect similar selectional pressures for externalized language to have shown up. Furthermore, recent studies (Vernot et al., 2016; Sankararaman et al., 2016; Mendez et al., 2016) conclude that the modern human X chromosome has been swept clean of archaic Neanderthal and Denisovan sequence. This is as expected between distinct species that hybridize and would simply be the result of hybrid infertility. The implication is that modern humans and Neanderthals/Denisovans were most likely distinct species. These studies offer further corroborating evidence for ancient introgressed archaic sequence in modern humans. Various genomic regions in modern humans that are significantly depleted of archaic introgressed Neanderthal/Denisovan sequence were found to be enriched for genes linked to language and brain development (Vernot et al., 2016). In particular, a large such archaic ‘desert’ on chromosome 7 contains the FOXP2 gene, which is arguably implicated in speech and language (Vargha-Khadem et al., 1995), supporting the view (Berwick and Chomsky, 2016) that Neanderthal/Denisovan did not have human language or speech, and effectively dismissing contrary speculations and stipulations (Dediu and Levinson, 2013). Therefore, as argued here, the genomic events leading to recursive language must have taken place on the human lineage after divergence from a common ancestor shared with Neanderthal and prior to the earliest divergence among human lineages. Any selective sweep for speech would then be expected to have occurred subsequently only, i.e. in Homo sapiens but not in Neanderthal.

And, in fact, the POU3F2 change that enhances expression of FOXP2 in human brains was not part of the gene flow from modern humans into Neanderthal that occurred in the Levant or Southern Arabia 125,000–100,000 years ago (Kuhlwilm et al., 2016). Since the change suggests relevance for improved speech, it is tempting to relate fully externalized human language to a recent selective sweep for the uniquely human POU3F2 haplotype at FOXP2 (Maricic et al., 2013; Pääbo, 2014). From this perspective, therefore, selection for improved articulation (externalized language) would have been a late evolutionary development that took place after the faded Levantine adventure of anatomically modern humans ~100,000 years ago.

The POU3F2 stretch coding for a uniquely human transcription factor may have enhanced the expression of speech-relevant tissue regulated by the two ancestral amino acid substitutions in FOXP2 we share with Neanderthals and Denisovans (Enard et al., 2002; Krause et al., 2007; Reich et al., 2010). Since speech is affected by loss of functional copies of FOXP2 in humans (Vargha-Khadem et al., 1995), it can be reasonably inferred that POU3F2 is involved in evolution of modern language, specifically speech (Maricic et al., 2013; Pääbo, 2014). It is significant, therefore, that the introgressed human stretch of FOXP2 in Altai Neanderthal did not include the POU3F2 segment that mutated in humans. From this it may be tentatively concluded that the uniquely human POU3F2 substitution most likely occurred subsequent to the interbreeding of humans with archaics that led to gene flow from humans into Neanderthal. Interbreeding must have followed the divergence (~110,000 years ago) of Altai Neanderthal from El Sidrón and Vindija lineages (which did not receive introgressed sequence from humans) but must be dated prior to the migration of eastern Neanderthals to the Altai Mountains (Kuhlwilm et al., 2016). Since the Altai Neanderthal genome contains 100,000 years old modern human haplotypes but not the human POU3F2 mutation, the change must have occurred between 100,000 and 60,000 years ago before the exodus out of Africa of modern humans. An additional argument for the recency of the mutation may be that the ancestral allele occurs at frequencies of 10% in some African populations but is virtually absent from non-African populations (Maricic et al., 2013).

Summarizing, the more recent selective sweep at FOXP2 in humans supports the idea that spoken language developed after the mutational change that yielded the basic property of human language was sufficiently fixated in the human population. The selective sweep for improved speech was exclusively human (Maricic et al., 2013), and, in fact, late in human prehistory between 100,000–60,000 ya, after modern humans first interbred with archaics (Kuhlwilm et al., 2016). Externalization of language therefore followed internal language generated by recursive syntax. The basic operation, Merge, accounting for discrete infinity of human language, was already in place some time before fully externalized language was realized. The late use of externalized linear language subsequent to the emergence of internal hierarchical language neatly accounts for the asymmetry in the syntactic mappings to the language-external interfaces with the conceptual-intentional and sensorimotor systems. Finally, this outcome is also consistent with the exclusive recruitment of click phonemes in Khoisan populations some time after the deep San/Yoruba separation (Gronau et al., 2011). In particular, the exclusive properties
of click sounds in language provide a more detailed location in time for the asymmetry of language, hierarchical merge and linear externalization.

Since we argued that possession of internal language preceded separation of human populations, and separation in turn preceded externalization, it may be tempting to argue further that “home sign” in deaf isolates (Goldin-Meadow and Mylander, 1998; Goldin-Meadow 2003, 2005) and cases of “emergent sign” such as Nicaraguan Sign Language (Kejl. 2002; Senghas et al., 2005) or Al-Sayyid Bedouin Sign Language (Sandler et al., 2005) share significant properties with the emergence of speech in evolutionary time.

In all these cases modern human brains with the capacity of internal language fully in place are/were seeking ways of externalizing internally generated language in some sensorimotor modality, sign or speech (Huybregts et al., 2016). In particular, the problem of how to externalize internal language could profitably be addressed by humans already in possession of the basic property of recursive language.

8.8. Alternative scenarios

Alternative scenarios e.g. uniformly or globally adopted phonemic click inventories with subsequent loss of click sounds except in the case of Khoi-San populations restate the problem in different terms. So in a relevant sense phonemic click consonants are just limited to Khoi-San languages and occur nowhere else. In particular, they never made it out of Africa, and plausibly originated with Khoisan speakers exclusively after their separation from the rest of the human lineages.

A scenario that holds that analogical to human genetic and phenotypic diversity, phonemic diversity progressively declines with increasing distance from Africa (Atkinson, 2011) is similarly problematic, incorrectly predicting gradual loss along the routes of migration out of Africa. In particular, there has been no gradual loss of click phonemes when humans left Africa. Atkinson’s conclusion is that loss of genetic and phonemic diversity results from serial founder effects that independently support an African origin of both modern humans and modern human languages. Phonemic inventory size, a property positively correlated with size of populations, is progressively reduced with increasing distance from Africa as a result of “successive population bottlenecks during range expansion.” However, the identification of loss of phonemic diversity with loss of phonemes overlooks the problem that loss of genetic diversity is not the same as loss of genes. Humans did not lose genes on their Great Trek from Africa. The logic of this line of reasoning leaves something to be explained. A similar suggestion that phoneme inventory correlates with population size has been shown to be an artifact of population relationships or gene flow (Moran et al., 2013).

In fact, all humans have the same genes and have the same sensorimotor systems for articulation and perception of phonemes, in particular consonants. Genetic diversity is greatest in Africa. But the question is whether articulatory diversity for consonants is greatest in Africa. All humans have the same genes, but generally, for most human genes there is more allelic diversity in Africa than anywhere else. Do polymorphisms have an analogue in language? All languages do not have the same phonemes. The question whether for most phonemes there is more allophonic diversity in Africa than anywhere else, is not exactly parallel. This would have been a more precise analogue although still a failed one since individual genes are shared by all of us but individual phonemes are not. What is shared among humans is the use of the same sensorimotor system for perceiving and articulating speech sounds. This system is insensitive to any specific phonemic system and is not constrained by group differences.

Even though there is not the slightest evidence that clicks have been lost with increasing distance from Africa, it is nevertheless the case that phonemic clicks have not been recruited elsewhere. Absence of clicks is apparently total and absolute outside of Africa. This “catastrophic” result is hard to explain under the serial founder effect model but is easily accounted for if phonemic clicks were original with ancestral Khoi-San, and became part of solving the externalization puzzle for internal language in this first diverging lineage only. The click solution has turned out to be surprisingly stable. Ancestral Khoi-San kept it to themselves. No other lineage hit on the same solution. We don’t know exactly why but suggested that biasing on clicks was a weak effect of the lack of a prominent alveolar ridge, a feature characteristic of San. Although the “outlier” position of phonemic clicks works out fine for our argument – strict association of clicks with the first humans to diverge – the apparently natural use and easy acquisition of complex phonemic click systems poses the serious problem, still unsolved, of why these systems are non-existent outside the oldest human lineage(s). The complex articulation of click sounds in addition to presently unknown demographic and geographic conditions may have been factors that prevented importing this solution into other human lineages and their languages. Perhaps the relevant (environmental, cultural, or biomechanical) conditions that made selection of clicks an attractive partial solution for externalizing language were no longer in force at later times for other populations. Efficiency reasons of lexical processing caused phonemic clicks to be retained in lineages that had once adopted them in ancestral times. But they ceased to be likely candidate elements in later systems of externalization, possibly because of the marked complexity of their articulation or changes in vocal tract morphology.

9. Language and behavioral modernity

Since language does not fossilize we have to fall back on proxies in the study of evolution of language. But what is a proxy for language? That is not entirely clear (Botha, 2016). No one would seriously consider the protruding chin, uniquely human, a proxy for language even though this “predicts” Neanderthals to have been speechless, and “explains” that human infants start developing language when their chins start protruding. For a cognitive system like modern language only products of behavioral modernity (e.g. symbolic behavior, body decoration, advanced tool and hunting technology, musical instruments, bone and ivory artifacts, etc.) have generally been accepted as legitimate proxies for language. However, according to British archaeologist Paul Mellars “[e]vidence for the presence of complex ‘symbolic’ behavior among the late Neanderthal populations in Europe has now effectively collapsed” (Mellars, 2010, p. 20148). On these terms, since Neanderthals did not (consistently) show increased symbolic behavior (Mellars, 2005; Tattersall, 2012; Pääbo, 2014), they would not have had modern language.

In fact, population geneticist Cavalli-Sforza provides evidence from genomic and linguistic research, in particular some significant parallels between reconstructed language family trees and genealogical trees of molecular biology, that modern language abilities emerged about the time that anatomically modern humans developed in Africa (Cavalli-Sforza, 2000). Furthermore, one of the leading biochemists and molecular biologists at the time, the late Alan Wilson argued for a recent sub-Saharan origin of the human species, and went so far as to suggest that the human capacity for language resulted from a mutation in mitochondrial DNA (as opposed to nuclear DNA). The mutation event would have happened to a woman (“mitochondrial Eve”) in a group of anatomically modern humans about 200,000 years ago, causing some rewiring of the brain that resulted in a faculty for language that made dis-
create infinity possible (Wilson, 1989). The proposal was quickly dismissed (Thorner and Wolpoff, 2003), though for the wrong reasons, viz. natural selection for externalized language does not apply to neutral genetic mutations. However, in view of the recent and quite abrupt emergence of language, without external pressures from environmental factors or any significant adaptation, the Basic Property of language would be expected to be optimally designed, resulting from a simple mutation, and essentially determined just by laws of nature, specifically minimal computation (Chomsky, 2013b). It is not surprising therefore that while Wilson's particular suggestion has never been seriously entertained, leading evolutionary biologists like the late Stephen Jay Gould or Richard Dawkins have embraced Chomsky's proposal that recursive language was key to a few genetic mutations that made us all human (Gould, 1993; p. 321; Dawkins, 2015; pp. 382–4).

Accept for a moment that behavioral modernity, e.g. increased symbolic and technological complexity, reflects complexity of social organization and use of language. Put differently, language might have existed without any signs of behavioral modernity if social structure was primitive. On this account Neanderthal or even Homo heidelbergensis might have had modern language. A proposal similar to this is made by Powell et al. (2009), who single out demographic factors, e.g. regional population densities or complexity of social organization, to explain the "spatial and temporal structuring" of the first marks of modern behavior "without invoking increased cognitive capacity" that is biologically determined. The different spatiotemporal structuring of the western Eurasian Upper Paleolithic transition and the southern African Late Stone Age would then follow from similar regional population densities in early Upper Paleolithic Europe (45 kyA) and sub-Saharan Africa (90 kyA) when behaviorally modern traits first appeared. However, since there are no language fossils and absence of relevant proxies is no longer proof of absence of language, any forthcoming explanation of evolution of language will only be put at a further remove if these assumptions are proven correct. The question now left open is why Neanderthals have never reached the regional population densities or complexity of social structure that would have led to behavioral modernity if indeed they possessed language. This is really asking how demographic factors and language relate to each other. Complexity of social structure may have been a precondition on spoken language in that it provoked the externalization of internally merge-generated thought that led to modern behavior. Or, the other way around, externalized language, speech, may have contributed to reaching a complexity of social structure that became manifest in behavioral modernity. In the first case, the lack of behavioral modernity relates to absence of language, and therefore to reduced social complexity. In the latter case, the missing modern behavior explains the simpler social structure of Neanderthal and therefore the lack of language. Either way, the consequence is that modern language was not part of the Neanderthal phenotype.

We may accept the idea (Powell et al., 2009) that behavioral modernity and regional population densities are correlated, suggesting further that population density and speech are also correlated. Demographic factors like regional population size and complex social structure may likely have spurred on externalized speech for communicative purposes. Or the other way around, externalized language may have contributed to demographic factors that led to population density and further complexification of social structure. We argued earlier that externalization had to solve the hard problem of linking a novel cognitive system to an ancient sensorimotor system that was already in place long before language emerged, functioning independently of language, and based on different principles of organization, viz. linear order or similar arrangements rather than structural dependence. Solving this puzzle by inventing rules for linearizing structure would have taken some time. On the other hand, being in part a cultural invention, externalization may not have required much, or any, further evolution of language. Speech, externalized language, must then have developed after the Basic Property of language became fixed in humans. In particular, after the mutational events responsible for the rewiring of the brain that yielded Merge had spread sufficiently through the population.

From these assumptions we may therefore continue to argue that the first markers of modern symbolic behavior in the archaeological record signalled the emergence of externalized language. As observed by Cavalli-Sforza, “increased complexity and differentiation of Palaeolithic cultures accompanied increase in language complexity” (Cavalli-Sforza, 2000; p. 60). Both the timing of the first human population split around 125 kyA, necessarily shortly after Merge started to spread, and the symbolic/technological evidence of the Blombos horizon around 80 kyA, suggesting spoken language, would then be consistent with our proposal. Conversion of structural properties, generated by the Basic Principle of language, into properties of linear precedence, required by SM, needed some amount of time to yield fully externalized language for communication and other functions.

10. Conclusion

This line of arguing provides a speculative but coherent and plausible account of the emergence of language and speech. It is an attempt to connect the little we know about evolution of human language with current knowledge and results of palaeoanthropology, archaeology, molecular biology, and generative studies of the human language phenotype (Fig. 3).

If correct, it offers specific answers to a number of relevant problems. One problem is the time gap between the biological novelty of a Merge-supporting brain and one of its behavioral effects, namely externalized language, rephrasing Tattersall’s “invention of language” as the cultural stimulus that “released” the potential for symbolic thought (Tattersall, 2012). Since internal language (syntax) preceded externalized language and speech developed independently in several human lineages only after (multiple) population divergences, monogenesis of the world’s languages must be ruled out. There is a single, invariant and unchanged human capacity for language (UG) but there have been multiple “first” externalizations, spoken languages, from which later languages
descended with modification. Another problem is the linguistic singularity of phonemic click systems as the exclusive property of Khoisan language families whose ancestral speakers were involved in the deepest divergence of human lineages. Click phonemes may have emerged as part of externalized language in San populations after they separated from Yoruba/Bantu populations but before they themselves diverged in populations ancestral to eastern African Hadzabe and southern African San. Furthermore, articulation of clicks may have been facilitated by a unique feature of vocal tract morphology specific to speakers of Khoisan languages. A further problem is the failed experiment of human evolution, the anatomically modern humans of North-Africa and the Levant, who still used ancient Mousterian technology and did not show symbolic behavior. However, these “failed” moderns may have been responsible for the human gene flow into Levantine Neanderthals, the likely ancestors of later Altai Neanderthals. The ingressed sequence included a FOXP2 segment but, significantly, the uniquely human FOXP2 mutation, plausibly implicated in speech, was not part of the introgression of human DNA into Neanderthal or Denisovan genomes. The regulatory change must therefore have occurred after the Levantine interbreeding between humans and Neanderthals happened but we left Africa.

Acknowledgements

I would like to thank two anonymous reviewers, Rudie Botha, Norbert Corver, Sam Epstein, Martin Everaert, Marcel den Dikken, Shigeru Miyagawa, and Charles Yang for their valuable comments on an earlier draft of the manuscript, and particularly Bob Berwick, Noam Chomsky, and David Lightfoot for stimulating discussion and encouragement. Needless to say, not everyone—and perhaps even no one—may agree with all or any of the views expressed in the present article.

References


