Early humans out of Africa had only base-initial numerals

Marc Allassonnière-Tang, One-Soon Her, Yung-Ping Liang, Eugene Chan, Hung-Hsin Hsu, Anthony Chi-Pin Hsu

▶ **To cite this version:**

Marc Allassonnière-Tang, One-Soon Her, Yung-Ping Liang, Eugene Chan, Hung-Hsin Hsu, et al.. Early humans out of Africa had only base-initial numerals. Humanities and Social Sciences Communications, 2024, 11 (1), pp.254. 10.1057/s41599-023-02506-z . hal-04599404

HAL Id: hal-04599404
https://u-paris.hal.science/hal-04599404v1

Submitted on 3 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-02506-z OPEN

Early humans out of Africa had only base-initial numerals

One-Soon Her[1,2], Yung-Ping Liang[2], Eugene Chan[3], Hung-Hsin Hsu[2,4], Anthony Chi-Pin Hsu[1] & Marc Allassonnière-Tang[5]✉

The vast majority of languages have numerals involving multiplication. Cross-linguistically, a numeral that involves a multiplier and a numeral base can be base-final, e.g., *three hundred* [three × hundred] in English, or base-initial, e.g., *ikie ita* [hundred × three] in Ibibio (Niger-Congo). A worldwide survey of 4099 languages reveals that 39% of the languages are base-initial, 48% are base-final, 4% use both orders, and 8% are without numeral bases. As the first step towards explaining this diversity and worldwide distribution, we offer convergent evidence to support the hypothesis that the languages of early humans in Africa had base-initial numerals. From a linguistic point of view, linearization is necessary for the verbal expression of multiplicative numerals. Between the two linear orders of multiplication, we demonstrate that the base-initial order has an initial advantage in communicative efficiency. We also offer typological evidence from the dominant head-initial word order in present-day numeral systems and nominal phrases in African languages. Finally, results from a phylogenetic analysis based on a global tree of human languages show that the base-initial order is more stable diachronically and more likely to be at the root of the reconstructed tree of languages in Africa between 100 and 150 thousand years ago. The dominant base-final order in non-African languages of modernity is thus likely to be a development after the Out-of-Africa exodus between 60 and 80 thousand years ago.

[1] Department of Foreign Languages and Literature, Tunghai University, Taichung, Taiwan. [2] Graduate Institute of Linguistics, National Chengchi University, Taipei, Taiwan. [3] Independent researcher, Hong Kong, China. [4] Institute for Language and Communication, Université catholique de Louvain, Louvain-la-Neuve, Belgium. [5] Lab Eco-Anthropology UMR 7206, National Museum of Natural History, Paris, France. ✉email: marc.allassonniere-tang@mnhn.fr

Worldwide distribution of base orders

The arithmetic functions of addition and multiplication as cognitive concepts do not necessarily involve linearization. However, complex numerals formed by addition and/or multiplication do require linearization when expressed verbally. The most common pattern of complex numerals is $(n \times \text{base}) + m$, where $m < \text{base}$ and $n \leq \text{base}$ (Comrie, 2013), e.g., *three hundred and two*, which expresses the arithmetic relation $(3 \times 10^2) + 2$. The commutative property of multiplication dictates that languages may adopt either the base-final order ($n \times$ base), e.g., *three hundred* in English, or the base-initial order, e.g., *ikie ita* [hundred × three] in Ibibio (Niger-Congo).

Curiously, while word order variations have been a focal point of typological studies, the base order parameter has long been overlooked and regarded as a trivial feature in linguistic typology, e.g., Comrie states, "The order of elements is irrelevant, as are the particular conventions used in individual languages to indicate multiplication and addition" (Comrie, 2013). Consequently, there were no large-scale surveys focusing on this typological feature. It is widely recognized that numeral systems are even more endangered than languages (Comrie, 2007). The documentation of linguistically and geographically precise information on numeral systems is thus urgently needed, as many indigenous numeral systems have lost their uniqueness and converged to the familiar dominant fully place-valued base-final decimal system (Freitas and Shell-Gellasch, 2012). Therefore, we conducted a manual survey of language grammar and relevant literature, along with an automatic survey of digitized grammatical descriptions. The manual survey benefited from data from existing surveys on numeral systems (Allassonnière-Tang and Her, 2020), WALS (Comrie, 2013), WACL (Her et al., 2022), and especially Eugene Chan's site Numeral Systems of the World's Languages. The automatic survey was mostly based on automatic searches in the DReaM Corpus (Virk et al., 2020).

For each language in Glottolog, the research team had to determine first whether there is sufficient data on the numeral system it employs. For each of the 4099 languages identified with sufficient data, we carefully examined the data available and determined whether its numeral system employs multiplicative numeral bases at all; if so, whether the word order is consistent; if so, whether the order is base-initial or base-final. A language thus belongs to one of these four categories: initial, final, split, and not_used. In the majority of cases, especially languages associated with prime cultures, e.g., Germanic and Sinitic languages, or languages associated with writing systems, the numeral systems employed have a rather rich inventory of numeral bases with a clearly identifiable and consistent base order. However, there are also languages with no multiplicative bases or a rather restricted set of bases. In such cases, the minimal criterion is that at least one multiplicative numeral base is employed with at least two examples of different multipliers. Mandarin Chinese and English are good examples of base-final languages, e.g., *liang bai* and *two hundred* are both [two × hundred] and *san bai* and *three hundred*, [three × hundred]. A good example of base-initial languages, besides Ibibio mentioned earlier, is Kilivila (Austronesian), e.g., *lakatu-yu* [hundred × two] and *lakatu-tolu* [hundred × three] (Senft, 1986). Rabha (Sino-Tibetan) is a good example of a language with both orders, as Rabha has an older base-initial system and a more recent base-final system (Joseph, 2007). Another example of such a split system is Rongga (Austronesian), where multiplicative numerals 20–90 are base-final, and those above 100 are base-initial. Pirahã (Muran) (Everett and Madora, 2012) and Andegerebinha (Pama-Nyungan) are two examples of languages without multiplicative numerals at all. Figure 1a for the first time shows the global distribution of 4099 languages in terms of base orders, based on an extension of the database in World Atlas of Classifier Languages (WACL) (Her et al., 2022).

Given the Out of Africa hypothesis (Liu et al., 2006; Haber et al., 2019; Scheinfeldt et al., 2010; Nielsen et al., 2017; Mellars, 2006; Gell-Mann and Ruhlen, 2011) and the hypothesis that all human languages derived from a single earlier language (Gell-Mann and Ruhlen, 2011; Campbell and Poser, 2008; Atkinson, 2011), a speculative set of hypotheses has been proposed to explain the distribution pattern in Fig. 1a. First, assuming that concepts of number and arithmetic may require cultural mediation to develop (Núñez, 2017), we hypothesize that given our current knowledge of early humans out of Africa between 50 and 80 thousand years ago (ka), they already had additive and multiplicative numerals, both of which are base-initial. The aim of the paper is to demonstrate that the converging evidence currently available is consistent with this base-initial hypothesis.

In order to have at least a road map, however tentative, towards explaining the worldwide distribution of base orders in Fig. 1a, we venture to further conjecture that the base-initial language groups out of Africa spread to various parts of the world via vertical inheritance; later, however, perhaps as late as the last 6000 years, certain language groups outside of Africa started to switch to

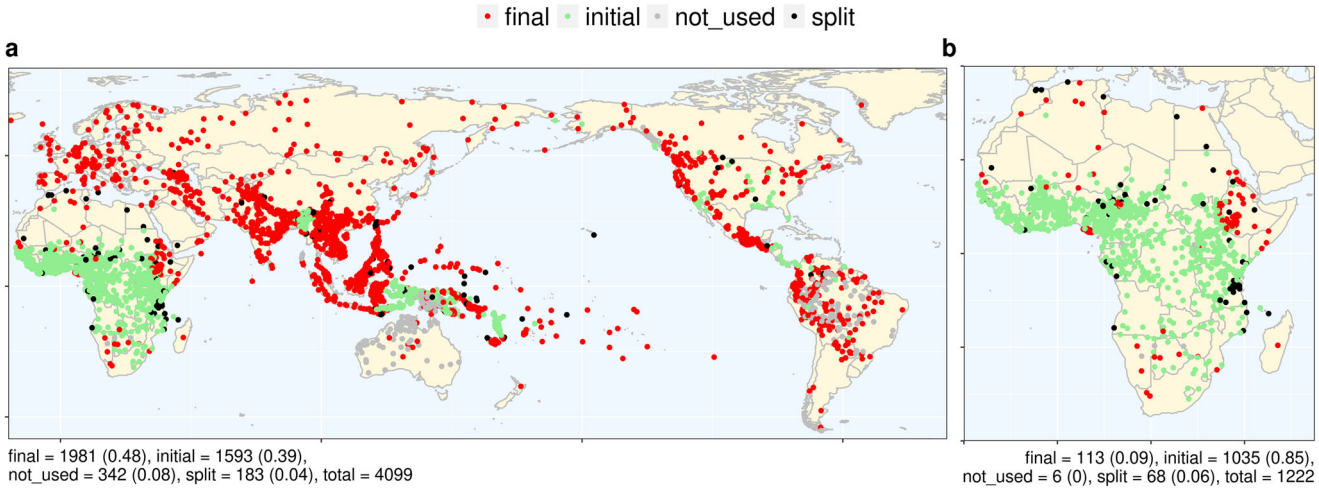

Fig. 1 Worldwide distribution of languages ($n = 4099$) with different base orders. Red dots represent base-final systems; green dots are base-initial numeral systems; gray dots are systems without numeral bases; black dots indicate systems with both orders. **a** shows a worldwide view while **b** shows Africa.

base-final, possibly due to the invention of numerical notations and/or writing systems. This hypothesis is motivated by the regularity numbered I5 Chrisomalis put forth in his seminal work on numerical notations, that all such systems are universally base-final (Chrisomalis, 2010). These base-final languages then spread via horizontal diffusion as well as vertical inheritance. Subsequently, all present-day base-split systems outside of Africa are thus in a transitional initial-to-final stage, but the situation in African languages may be mixed in both directions. Languages with no multiplicative numeral bases have lost them due to disuse. We fully acknowledge that these highly speculative and contentious claims require enormous further research, which is beyond the scope of the paper.

This study is thus the first serious step in solving the jigsaw puzzle in Fig. 1a and addresses the research question of what evidence is available to support the hypothesis that early humans Out of Africa had base-initial numerals, not base-final numerals. To that aim, we will offer evidence from several fronts. First, the obvious reason is that African languages today overwhelmingly (84.3%) employ the base-initial order (Fig. 1b). Second, we demonstrate that the base-initial order offers significant advantages over the base-final order in terms of communicative efficiency when multiplication first emerged in numeral systems. Third, typologically, the dominant head-initial order in nominal phrases in African languages today also lends support. And, finally, we offer support from a phylogenetic analysis based on a global tree of human languages. The paper is organized accordingly.

The dominant base-initial order in African languages

Most base-initial African languages belong to a single language family, Niger-Congo, the largest language family in the world rooted in Africa (68.5%, 840/1225) (Blench, 2006). These facts suggest the likelihood that Proto-Niger-Congo (PNC) has base-initial numerals. The reconstructed base-initial forms of 6, 7, and 9 as '5 + 1', '5 + 2', and '5 + 4', respectively (Pozdniakov, 2018), where 5, reconstructed as 'hand', is a base-initial additive numeral base, and 20 as 'person', thus implying 'hands (two) feet (two)'. In the Khoisan family, another ancient language family indigenous to Africa, six out of 13 are base-initial, and seven are base-final, likely due to recent contact with colonial languages. Note that the 9% base-final African languages are largely Afro-Asiatic (67.8% base-final, 78/115), a language family originated either in the Levant area or in east/northeast Africa with close ties to Western Asia and the early back-to-Africa migration (Hodgson et al., 2014). We thus hypothesize that Proto-Afro-Asiatic is base-final and the base-initial feature is due to contact with Niger-Congo.

With studies of mitochondrial DNA (mtDNA) and the autosomes, it is found that the Khoe-San people possess the most distinct lineage divergence among all human populations (Slebusch and Jakobsson, 2018). The Khoe-San split is estimated to have occurred between 200 and 300 thousand years ago (Slebusch and Jakobsson, 2018). Besides a Eurasian back-migration through Egypt between 15 and 10 thousand years ago, there was another Eurasian back-migration across the Mandab Strait reaching the Ethiopian Highlands, which is estimated to have occurred around 3000 years before the present era (Slebusch and Jakobsson, 2018). The group that migrated back from Eurasia assimilated with both Eurasian and East African genetic lineages in the Ethiopian Highlands, proceeding southward to reach southern Africa 2000 years before the present era (Slebusch and Jakobsson, 2018). The Bantu expansion commenced in western Africa around 5000–3000 years ago and reached southern Africa around 1500 years before the present era (Slebusch and Jakobsson, 2018). The back-migration from Eurasia through the Ethiopian Highlands left genetic imprints primarily in populations in Eritrea, Ethiopia, Somalia, eastern Africa, and contemporary southern Africa (Slebusch and Jakobsson, 2018). The Bantu expansion, in contrast, left genetic marks across western, central, and southern Africa (Slebusch and Jakobsson, 2018).

If those who migrated out of Africa around 50 and 80 thousand years ago had initially employed a multiplicative base in their language(s) and this base evolved into a final form outside Africa, the distribution of initial and final base languages could be explained by the earlier Eurasian back-migration and the subsequent Bantu expansion. Initially, the back-migration introduced base-final languages. Subsequently, the Bantu expansion introduced base-initial languages, reshaping the linguistic landscape from western to southern Africa, thereby intersecting with the base-final languages. This historical sequence in Africa's prehistoric context likely underlies the contemporary distribution of base-initial and base-final languages. The disjunctive distribution of base-final languages in Eritrea, the Ethiopian Highlands, and southern Africa originates from the earlier Eurasian back-migration, while the continuous distribution of base-initial languages traces back to the later Bantu expansion. These migrations contribute predominantly to the present state, with minimal influence from horizontal diffusion.

The initial advantage

When a multiplicative complex numeral first emerged in the course of language evolution, a choice between ($n \times$ base) and (base $\times n$) had to be made. Even in the most extreme generativist view that the fundamental design of language is for thought (Baker, 2001; Berwick and Chomsky, 2015), the externalization of language via sensory-motor mechanisms nonetheless affords humans by far the most powerful means of communication among all known species. Processing and communicative factors must thus play crucial roles in the grammar of the externalized language (Gibson et al., 2013; Kemp et al., 2017; Xu et al., 2020). We argue that the initial choice of the (base $\times n$) order offers such an advantage.

Note first that the law of commutativity likewise applies to the concept of addition, and it is well-established that the existence of multiplication in a numeral system implies the existence of addition (Greenberg, 1978), indicating that cognitively addition is a foundation of multiplication. The order of elements in additive numerals is thus instrumental to the order of elements in multiplicative numerals.

Between the two addends in an additive numeral, languages generally favor the [larger + smaller] order (Greenberg, 1978; Liu and Xu, 2019), also known as the Packing Strategy (Hurford, 2007) and the Ordering Principle (Chrisomalis, 2010). The cognitive advantage this order has over the reverse order is this: given three natural numbers x, y, and z, where $x > y$ and $x + y = z$, x presents the closest approximation of the final product z; giving x first is thus more conducive for the listener to grasp the ultimate number the speaker aims to convey. Considering also unexpected interruptions in speech, the [larger + smaller] order favoring the constituent with more information first receives a communication-based explanation (Gibson et al., 2013; Liu and Xu, 2019). The [larger + smaller] order of the two underlined constituents in [[a dozen] and [one] roses] is preferable over [[one] and [a dozen] roses], and likewise for [[one hundred] and [one] guests].

Thus, when humans with a language that had only simple numerals, e.g., 1, 2, … ,10, first developed the concept of an additive numeral base, e.g., 10, the natural choice of externalizing additive numerals is base-initial, hence 10 + 1, 10 + 2, … ,10 + 9. Note that, besides the advantage of efficient delivery and

reception of the target number, the base-initial order also allows minimal disturbance to the existing system, as the newly added additive numerals, i.e., $10+1$, $10+2, \ldots, 10+9$, always begin with the base. Thus, the listener, when hearing a simple numeral, can be sure that it is the target number. In contrast, base-final additive numerals, i.e., $1+10$, $2+10, \ldots, 9+10$, which always start with a simple numeral, would create uncertainty for the target number, which can be the simple numeral or the simple numeral plus a base yet to appear. In our dataset, among the 182 languages that have additive numerals but no multiplicative numerals, 153 (84%) employ the base-initial order, only 5 (3%) use the base-final order, 4 (2%) have both orders, and 20 (11%) have data which are hard to determine. Such an overwhelming bias favoring the base-initial additive numerals thus receives a functional explanation. Note also that among the 182 languages, only 6 are in Africa, 3 (50%) employ the base-initial order, only 1 (17%) use the base-final order, and 2 (33%) have both orders.

Remarkably, to our knowledge, previous works have not considered the preferable order within a multiplicative numeral composed with a multiplier n and a numeral base. Logically, the base-initial order ($\text{base} \times n$) is consistent with the [larger + smaller] principle and enjoys the same advantage of efficient delivery and reception of the target number. This principle can be formulated more accurately in terms of the numerical distance to the target number, as in (1).

(1) The shortest distance principle of number naming

Between the two numerals in an additive or multiplicative numeral, the one with the shortest numerical distance to the target number is preferred to appear first.

For example, given the additive number *twenty-one* as the target, there are two possible orders between the numbers 20 and 1. In a large-initial order, [[20] [1]], the distance between the actual number and the first number mentioned is equal to $21-20=1$. In a large-final order, [[1] + [20]], the distance between the actual number and the first number mentioned is equal to $21-1=20$. By comparison, the large initial order is preferred since it has the shortest distance to the target number. Likewise, for multiplicative numbers, in the number 20 of a decimal system, two orders are possible between the multiplier 2 and the base 10. In a base-initial order [[10] × [2]], the distance between the actual number and the first number mentioned is equal to $20-10=10$. In a base-final order, [[2] × [10]], the distance between the actual number and the first number mentioned is $20-2=18$. The base-initial order is preferred since it has the shortest distance from the actual number. This amounts to an efficient strategy to consistently offer the best approximates of the target number until the target is reached.

We thus argue that when early humans with a language that had only simple numerals, e.g., $1, 2, \ldots, 10$, and additive numerals, e.g., $11, 12, \ldots, 19$, first developed the concept of a multiplicative numeral base, e.g., 10, the natural order in externalizing multiplicative numerals is base-initial, e.g., $10\times1, 10\times2, \ldots, 10\times9$. An additional significant advantage of such a numeral system is that all complex numerals, whether additive, e.g., $10+3$, or multiplicative, e.g., 10×3, or a combination of both, e.g., $10\times3+3$, begin with the base. Thus, the listener, when hearing a simple numeral, can still be sure that it must be the target number. This allows the simple numerals, which are undoubtedly the most frequently used numerals in any numeric society (Kemp et al., 2017), to remain straightforward and unambiguous. Again, in contrast, base-final multiplicative numerals, e.g., $1\times10, 2\times10, \ldots, 9\times10$, on the other hand, would create considerable ambiguity for the listener when they hear the first numeral. For example, given a limited base-final decimal numeral system ranging from 1 to 99, hearing the numeral 3 gives three possibilities: 3, 30, and 33.

However, we acknowledge that the preference for the base-initial order depicted above would benefit tremendously from psycholinguistic experimental evidence. Here, we find the study by Cooperrider et al. (2017) as an inspiration. Such a study would certainly require significant time and effort and deserves a separate paper. We therefore leave it as a future study, for which we are actively seeking collaboration.

We further note that having both additive numerals and multiplicative numerals in a base-initial order, e.g., $10+2$ and 10×2, may create ambiguity between addition and multiplication if neither operation is overtly marked, as both would appear as 10 immediately followed by 2. Thus, the account proposed above predicts that such a numeral system should at least mark either addition or multiplicative to avoid ambiguity. Greenberg observed that overt expression of the operation of addition, called 'link for addition', is rather frequent but rarely for multiplication (Greenberg, 1978). Among the 153 languages with only base-initial additive numerals but without multiplicative numerals, 121 (79%) overtly mark addition already. Within the 3 African languages among the 153, 2 mark addition. These facts suggest that base-initial languages can easily remedy this possible drawback.

Arguments based on the cognitive and communicative advantages of the base-initial order lend support to the hypothesis that early humans employed only base-initial multiplicative numerals, and this feature has largely remained stable in Africa, while elsewhere, many languages flipped the order, possibly due to the invention of numerical notations and subsequently writing systems, a speculative hypothesis currently under intensive investigation by the research team.

This transition from the base-initial to base-final order may stem from human working memory coupled with visual symbols aiding decipherment. For instance, consider symbol 卄 for 10; decoding 卄卄卄卄卄卄 involves recalling its numeral and counting. Two approaches emerge: recalling 卄 then counting, or counting symbols then recalling 卄. For the base-initial order, the decipherment involves the recollection of the symbol 卄 for 10 and the counting sequence "10; 1, 2, 3, 4, 5, 6,…,60", where the base "10" is separated by "1" to "5" in the counting sequence. For the base-final order, the counting sequence directly places the numeral count "6" before base "10" as "1, 2, 3, 4, 5, 6; 10…60". The base-final order requires less working memory load, distinguishing it from the base-initial order.

The head-initial order

Word order typology offers additional support, as the order between a head and its modifier within a nominal phrase tends to be consistent. Base as the head in a multiplicative numeral and noun as the head in a nominal phrase are expected to be synchronized. Thus, if the noun-head order is initial at the root, it strengthens the probability of also having base-initial order at the root. Taking the order of nouns and adjectives as an example, in a nominal phrase formed by an adjectival element (Adj) and a noun (N), e.g., *strong women*, N is seen as the head of the phrase, while Adj serves as a modifier. It has long been recognized that between a numeral (Num), including numerical quantifiers, and the quantified noun, e.g., *three women* and *many children*, the relation is also modifier-head, like the Adj–N relation, as seen in the parallel among these three phrases: *strong women*, *three women*, and *three strong women*, the only difference is that numerals quantify nouns and adjectives qualify them (Stampe, 1976). The modifier-head relation likewise exists within a multiplicative numeral, where the multiplier n quantifies the numeral base. The word *dozen* in English in fact can be used as a numeral base as well as a noun, e.g., *three dozen roses* and *three dozens of roses*, respectively. Thus, given the dominant head-initial order within

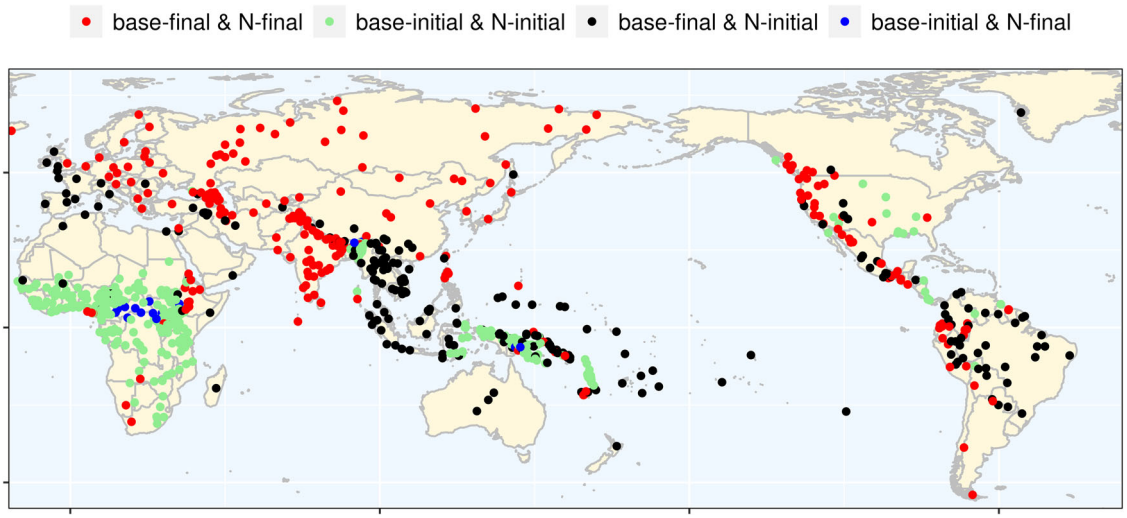

Fig. 2 Worldwide distribution of languages when comparing the alignment of numeral bases and adjectives in the noun phrase (Dryer, 2013). Red dots indicate languages with base-final and N-final orders. The green dots indicate languages with base-initial and N-initial orders. The black dots indicate languages with base-final and N-initial orders. The blue dots indicate languages with base-initial and N-final orders..

multiplicative numerals in African languages, a comparable prevalent head-initial order of N–Adj is also found (see Fig. 2).

Multiple analyses show that the order of numeral bases is significantly harmonized with the order of adjectives and nouns, not only within languages spoken in Africa but also at the global level. We first use a Chi-square test and a logistic regression as representatives of non-parametric and parametric tests to assess if the distribution of base order and adjective order is significantly different from a random distribution and if this difference has a large effect size. Second, we use generalized linear mixed models (Bates et al., 2015) to assess if there is a significant interaction between base order and adjective order as fixed effects. In this process, we control for language family and geographic area by setting these variables as random effects. Finally, we use conditional inference trees (Hothorn et al., 2019) to extract the hierarchical interaction between the variables. Conditional inference trees are an algorithm of decision trees that use a recursive binary split of the dependent variables. At each step, the algorithm uses a permutation test to evaluate the association between predictors and the response variable. The predictor with the strongest association is used to perform a binary split on the response variable. This procedure is repeated until the data cannot be split further. During this process, we did not conduct cross-validation since conditional inference trees use permutations, which can be considered to fulfill cross-validation. We compare the predictive performance of the generalized linear mixed models and decision trees with the majority baseline, which is what a model would get by randomly guessing that all tokens of the data set belong to the largest category, i.e., base-final (0.48).

At the global level, the Chi-square test ($X^2 = 217.98$, df = 3, p-value < 0.001) indicates a medium-large effect size (Cramer's $V = 0.47$) for the harmonization between the order of adjectives and numeral bases. Logistic regression predicting base order with adjective order shows that a language is 18 times more likely to be base-initial if it is N-initial. Generalized linear mixed models controlling for family and macroarea as a random effect also show that the model predicting base order with adjective order is significantly better than a null model (AIC 551 vs. 641, p-value < 0.001). The model considers that the N-initial order is positively associated with the base-initial order (Est. = 3.3514, Std. error = 0.4293, $p < 0.001$, $R^2 = 0.6$). Furthermore, conditional inference trees predicting base order based on adjective order also show that the N-initial order can predict the base-initial order with an accuracy significantly above the majority baseline (0.61 vs. 0.50). Decision trees additionally trained with family and area information along with adjective-order still show that the adjective-order is relevant for predicting base-order (accuracy 0.79 vs. majority baseline 0.50). The prediction is thus borne out that African languages today overwhelmingly follow the N-initial order, producing a harmonization between numeral bases and nouns. This suggests that the nominal phrase in the languages of early humans was consistently head-initial. We shall now employ phylogenetic methods to further test the base-initial hypothesis.

Phylogenetic analysis

For the phylogenetic analysis, we use the world tree sample generated by Bouckaert et al. (2022). On the one hand, we consider the world tree sample pruned by keeping all the languages that are available in our data set. On the other hand, we extract a subset of the world tree sample by filtering languages located in Africa based on Glottolog coordinates (Hammarström et al., 2022). This reduced tree sample is used to conduct phylogenetic analyses on African languages. Finally, we conduct a second filter on both tree samples (world and Africa) to keep only languages for which we have data for both base order and adjective order (including only final and initial orders). A simplified view of the world tree with the languages included in the analysis is shown in Fig. 3. Additional data and code are available in the supplementary materials.

For reconstructing the order of base at the root of both tree samples (world tree and tree with languages in Africa), two methods are used. First, we conduct ancestral character estimation (Paradis and Schliep, 2019), in which we consider an equal-rates model with discrete characters. The method is expected to infer the probability of each base-order at the root of the tree samples. Second, we use reverse jump Markov Chain Monte Carlo (Gowri-Shankar and Rattray, 2007), which gives not only the probability of each base-order at the root of a tree sample but also the transition rates between each base-order. This procedure uses a Continuous Time Markov Chain process that considers possibilities of reversed change between base orders, i.e., the model not only scores the probability that a language switches from base-initial to base-final, but it also scores the probability that a language switches from base-initial to base-final but re-switches to base-initial later on. The parameters are set with

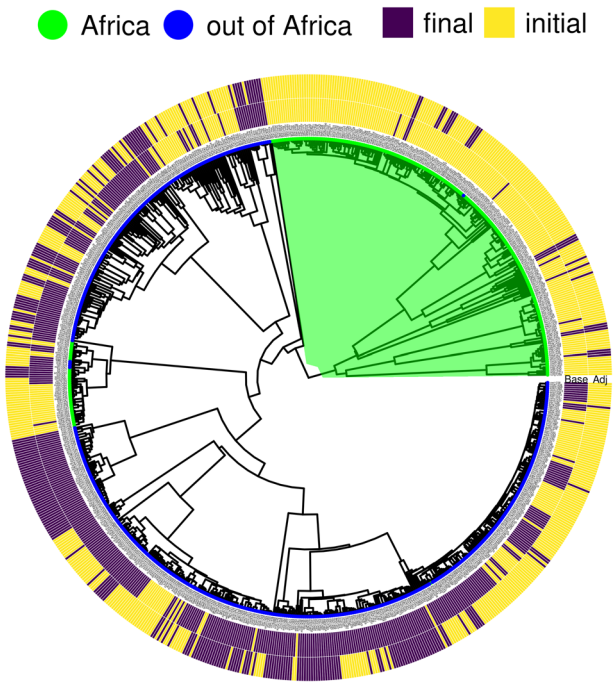

Fig. 3 The world tree (Bouckaert et al., 2022) pruned to keep 770 languages for which both the information of base order and adjective order is available. The colors of the visualization refer to the base (inner circle) and noun–adjective order (outer circle). The languages highlighted in green indicate the languages located in Africa that are considered for the analyses focusing on Africa.

1,000,000 generations and the first half being discarded as a burn-in. The sample frequency is set to 1000 iterations, which results in an output of 500 iterations. The stepping stone sampler is set with 100 tones and 1000 iterations per stone to estimate the marginal likelihood.

The output of both methods is compared to assess the robustness of the results. Then, to assess the correlated evolution between base order and adjective order, we combined the status of both variables for each language. For example, if a language is base-final and N-final, it is assigned the vector of [11]. If it is base-initial and N-final, it is assigned [01], among others. We then use reverse jump MCMC to assess the transition rates between the four possible states, i.e., [11], [10], [01], [00]. Two models are compared: an independent model that assumes no interaction between base order and adjective order and a dependent model that infers a correlated evolution between base order and adjective order. We use Bayes factors (Burnham and Anderson, 2002) to evaluate which model better explains the variance in the data. If the independent model is preferred, we can then infer the transition rates between the four states to visualize what are the correlated evolutionary trends between base order and adjective order. Additional data and code are available in the supplementary materials.

Both results suggest that the phylogenetic signal is not strong partially due to the uncertainty of the world tree sample, but that the presence of initial base order at the root of both the world tree sample and the Africa tree sample is more likely. Across the trees of the sample, while ancestral character estimation cannot successfully distinguish the most likely status at the root (between base-final, base-initial, mixed, none), the mean probability of having base-initial at the root is consistently the highest (but not by much) when using reverse jump MCMC at the worldwide level ($p = 0.251$) and in Africa ($p = 0.251$), respectively. This probability is significantly higher than the probability of having base-final numerals at the root when using a *t*-test with Bonferroni correction. Second, we use reverse jump MCMC to infer the likelihood of correlated evolution between base order and adjective order in the world tree sample and the Africa tree sample. The results at the worldwide level show that the dependent model is more likely than the independent model with very strong evidence (Log marginal likelihood of the dependent model = −522.3302, log marginal likelihood of the independent model = −548.5822, Bayes factor = 52.50392). Strong evidence supporting the dependent model is also found when considering languages in Africa (Log marginal likelihood of the dependent model = −127.5091, log marginal likelihood of the independent model = −134.2548, Bayes factor = 13.49136). The reverse jump MCMC also shows that the combination of base-initial and N-initial orders is more likely at the root of the world tree (mean probability = 0.261) and in African languages (mean probability = 0.253). In terms of transition rates, the reverse jump MCMC shows similar results on the world tree sample and the tree sample considering African languages: the order of numeral bases and adjective order tends to be harmonized as either final or initial. However, in the Africa tree sample, the base-final and N-final combination is less stable than the base-initial and N-initial combination. The order of the adjective and the noun is likely to switch from N-final to N-initial. At this stage, the order can go back to be harmonized as final or initial. Nevertheless, if the order is harmonized as base-initial and N-initial, it is less likely to change.

Conclusion

We hypothesized that the base-initial order in additive and multiplicative numerals is more likely to be at the ancestral state of languages worldwide. This hypothesis is supported by evidence from two perspectives. First, the shortest distance principle suggests that larger numbers are more likely to be said first in additive and multiplicative numerals, which supports the presence of base-initial systems at the emergence of multiplicative numerals. Second, synchronic and diachronic evidence from worldwide languages suggests that the order of the head in a nominal phrase is highly harmonized with the order of the numeral base, supporting the presence of head-initial and base-initial orders at the root of worldwide languages and thus languages in Africa. We expect that this research will contribute to the discussion of the origin and evolution of human languages.

Data availability

The data, the R code, and the full output of the models used in this study are available at the following repository: https://osf.io/wqhk7/?view_only=9dce74309890444ebccd515964d5c2a3 .

Received: 31 August 2023; Accepted: 11 December 2023;
Published online: 12 February 2024

References

Allassonnière-Tang M, Her OS (2020) Numeral base, numeral classifier, and noun: word order harmonization. Lang Linguist 21(4):511–556
Atkinson QD (2011) Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science 332:346–349
Baker M. The atoms of language (Basic Books, 2001)
Bates D, Maechler M, Bolker BB, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48. https://doi.org/10.18637/jss.v067.i01
Berwick RC, Chomsky N (2015) Why only us: language and evolution (MIT Press)
Blench R (2006) Archaeology, language, and the African past. AltaMira Press
Bouckaert R et al (2022) Global language diversification is linked to socio-ecology and threat status. Preprint at https://doi.org/10.31235/osf.io/f8tr6

Burnham KP, Anderson DR (2002) Model selection and multimodel inference. Springer-Verlag
Campbell L, Poser WJ (2008) Language classification: history and method. Cambridge University Press
Chrisomalis S (2010) Numerical notation. A comparative history. Cambridge University Press
Comrie B (2007) Endangered numerals. OGMIOS newsletter of the foundation for endangered languages: issue 34 https://www.ogmios.org/ogmios/Ogmios_034.pdf. Accessed 31 Dec 2007
Comrie B (2013) Numeral bases. In: Dryer MS, Haspelmath M (eds.) The world atlas of language structures. Max Planck Institute for Evolutionary Anthropology
Cooperrider K, Marghetis T, Núñez R(2017) Where does the ordered line come from? Evidence from a culture of Papua New Guinea Psychol Sci 28(5):599–608. https://doi.org/10.1177/0956797617691548
Dryer MS (2006) Order of adjective and noun. In the World Atlas of language structures online framework. J Comput Graph Stat 15(3):651–674. https://doi.org/10.1198/106186006X133933
Dryer MS (2013) Order of adjective and noun. In: Dryer MS, Haspelmath M (eds.) The World Atlas of language structures. Max Planck Institute for Evolutionary Anthropology
Everett C, Madora K (2012) Quantity recognition among speakers of an anumeric language. Cogn Sci 36(1):130–141
Freitas PJ, Shell-Gellasch A (2012) When a number system loses uniqueness: the case of the Maya. Convergence. The MAA Mathematical Sciences Digital Library
Gell-Mann M, Ruhlen M (2011) The origin and evolution of word order. Proc Natl Acad Sci USA 108(42):17290–17295
Gibson E et al. (2013) A noisy-channel account of crosslinguistic word-order variation. Psychol Sci 24:1079–1088
Gowri-Shankar V, Rattray M (2007) A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. Mol Biol Evol 24(6):1286–1299
Greenberg J (1978) Generalizations about numeral systems. In: Denning K, Kemmer S (eds.) On language: selected writings of Joseph H. Greenberg. Stanford University Press, 1990, pp. 271–309 [First published in Universals of the Human Language 3, 249–295 (Stanford University Press, 1978)]
Haber M et al. (2019) A rare deep-rooting D0 African Y-chromosomal haplogroup and its implications for the expansion of modern humans out of Africa. Genetics 212(4):1421–1428. https://doi.org/10.1534/genetics.119.302368
Hammarström H, Forkel R, Haspelmath M, Bank S (2022) Glottolog 4.7. Max Planck Institute for Evolutionary Anthropology
Her OS, Hammarström H, Allassonnière-Tang M (2022) Defining numeral classifiers and identifying classifier languages of the world. Languist Vanguard 8.1:151–164
Hodgson JA, Mulligan CJ, Al-Meeri A, Raaum RL (2014) Early back-to-Africa migration into the Horn of Africa. PLoS Genet 10(6):e1004393. https://doi.org/10.1371/journal.pgen.1004393
Hothorn T, Hornik K, Zeileis A(2019) Unbiased Recursive Partitioning: A Conditional Inference Framework. J Comput Graph Stat 15(3):651–674
Hurford JR (2007) A performed practice explains a linguistic universal: counting gives the packing strategy. Lingua 117:773–783
Joseph (2007) UV Rabha. Brill, Leiden
Kemp C, Xu Y, Regier T (2017) Semantic typology and efficient communication. Annu Rev Linguist 4:109–128
Liu E, Xu Y (2019) Rapid information gain explains cross-linguistic tendencies in numeral ordering. In: Goel AK, Seifert CM, Freksa C (eds.) Proceedings of the 41st annual conference of the Cognitive Science Society. Cognitive Science Society, pp. 2166–2173
Liu H, Prugnolle F, Manica A, Balloux F (2006) A geographically explicit genetic model of worldwide human-settlement history. Am J Hum Genet 79(2):230–237
Mellars P (2006) Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. Proc Natl Acad Sci USA 103:9381–9386
Nielsen R et al. (2017) Tracing the peopling of the world through genomics. Nature 541:302–310
Núñez RE (2017) Is there really an evolved capacity for number? Trends Cogn Sci 21(6):409–424. https://doi.org/10.1016/j.tics.2017.03.005
Paradis E, Schliep K (2019) Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinform 35:526–528
Pozdniakov K (2018) The numeral system of Proto-Niger-Congo: a step-by-step reconstruction. Language Science Press
Scheinfeldt LB, Soi S, Tishkoff SA (2010) Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. Proc Natl Acad Sci USA 107:8931–8938
Senft G (1986) Kilivila: the language of the Trobriand Islanders. de Gruyter Mouton
Slebusch CM, Jakobsson M (2018) Tales of human migration, admixture, and selection in Africa. Annu Rev Genom Hum Get 19:405–428
Stampe D (1976) Cardinal number systems. In: Mufwene SS, Walker CA, Steever SB (eds.) The proceedings of CLS 12. Chicago Linguistic Society, pp. 594–609
Virk SM, Hammarström H, Forsberg M, Wichmann S (2020) The DReaM corpus: a multilingual annotated corpus of grammars for the world's languages. In: Calzolari N et al (eds.) Proceedings of the 12th language resources and evaluation conference. European Language Resources Association, pp. 871–877
Xu Y, Liu E, Regier T (2020) Numeral systems across languages support efficient communication: from approximate numerosity to recursion. Open Mind 4:57–70. https://doi.org/10.1162/opmi_a_00034

Acknowledgements

O-SH offers his heartfelt thanks to the graduate students and researchers for their help in building the database of base orders, including part-time RAs: Hsieh, Chen-tien; Lai, Wan-Chun; Chen, Meng-Ying; Wang, Wei; Chen, Ching-Perng; Chen, Yun-Ju; Liao, Jia-Yu; Li, Bing-Tsiong; Chia, Cheng-Pin; Allassonnière-Tang, Marc; Lin, Kun-Han; Huang, Yu-Min; Chen, Chia-Chi; Yeh, Chu-Hsien; Yang, Wen-Chi; Huang, Tsung-Chia; Chen, Shen-An; Jheng, Jhih Siou; Liang, Yu-Ting; Gao, Zhong-Liang; Cao, Zi-Yun; Hsu, Hung-Hsin; Liang, Yung-Ping; Lo, I-Chieh; Chen, Yi-Ju; Chen, Wei-You; Cheng, Yu-Ching; full-time RAs: Chen, Ying-Chun; Lin, Yen-Tse; Ho, Pei-Hsuan; and post-docs: Hsieh, Fu-Tsai; Tsai, Hui-Chin; Hsiao, Pei-Yi; Hsu, Chi-Pin. O-SH gratefully acknowledges the following grants awarded to him as PI by Taiwan's National Science and Technology Council (NSTC): 101-2410-H-004-184-MY3, 102-2811-H-004-023, 103-2811-H-004-003, 103-2633-H-004-001, 103-2410-H-004-136-MY3, 104-2811-H-004-004, 104-2633-H-004-001, 104-2410-H-004-164-MY3, 106-2410-H-029-077-MY3, 107-2811-H-004-517, 108-2811-H-004-521, 108-2410-H-029-062-MY3, 109-2811-H-004-522, 111-2410-H-029-009-MY3.

Author contributions

O-SH and MA-T conceived and designed the study. All authors contributed to establishing the data, with EC, Y-PL, and H-HH having leading roles. MA-T conducted the quantitative analyses. O-SH, Y-PL, H-HH, ACH, and MA-T were involved in the discussion and interpretation of the results. O-SH, Y-PL, ACH, and MA-T contributed to the writing of the paper.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Marc Allassonnière-Tang.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024