UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

How communicatively optimal are exact numeral systems? Once more on lexicon size and morphosyntactic complexity

Permalink

https://escholarship.org/uc/item/4hz9j5f8

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Cathcart, Chundra
Rubehn, Arne
Bocklage, Katja
et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

How communicatively optimal are exact numeral systems? Once more on lexicon size and morphosyntactic complexity

Chundra Cathcart[1,*], Arne Rubehn[2], Katja Bocklage[2], Luca Ciucci[2], Kellen Parker van Dam[2], Alžběta Kučerová[2], Jekaterina Mažara[2], Carlo Y. Meloni[2], David Snee[2] & Johann-Mattis List[2]

[1]Institute for the Interdisciplinary Study of Language Evolution, University of Zurich, Zurich, Switzerland

[2]Chair for Multilingual Computational Linguistics, University of Passau, Passau, Germany

[*]chundra.cathcart@uzh.ch

Abstract

Recent research argues that exact recursive numeral systems optimize communicative efficiency by balancing a tradeoff between the size of the numeral lexicon and the average morphosyntactic complexity (roughly length in morphemes) of numeral terms. We argue that previous studies have not characterized the data in a fashion that accounts for the degree of complexity languages display. Using data from 52 genetically diverse languages and an annotation scheme distinguishing between predictable and unpredictable allomorphy (formal variation), we show that many of the world's languages are decisively less efficient than one would expect. We discuss the implications of our findings for the study of numeral systems and linguistic evolution more generally.

Keywords: Numeral systems; morphological complexity; communicative efficiency

Introduction

In the world's languages, numeral systems — the inventories of terms used to denote different quantities — show considerable diversity along a number of dimensions. Languages differ in terms of whether their numeral systems are exact, i.e., containing a unique term referring to every possible cardinality, or approximate, with one-to-many mappings between numeral terms and numerosities (Núñez, 2017; O'Shaughnessy et al., 2022; Pica et al., 2004). Different languages build their numeral systems around different arithmetic bases; while decimal systems are the most frequent cross-linguistically, vigesimal (base twenty) systems are also common and a variety of other bases are found as well (Comrie, 2022). Numerals also differ in terms of their morphosyntactic properties, exhibiting different orders of elements (i.e., modifiers with respect to bases) within numeral terms (Allassonnière-Tang & Her, 2020). Much effort has been made to account for this variation. A popular line of research argues that numeral systems are subject to trade-offs between domains of complexity ultimately stemming from a pressure to facilitate efficient communication, similar to tradeoffs in other domains (e.g., Kemp & Regier, 2012).

Here, we focus on another dimension of variation that has attracted some attention in recent years, namely that of formal variation within numeral systems (Cathcart, 2026; Koile & Blasi, 2025; Rubehn et al., 2025). Specifically, we build on existing work which argues that languages optimize a tradeoff between the size of the lexicon used to construct numerals, comparing attested numeral systems to hypothetical artificially generated systems exhibiting a diverse range of configurations. We argue that previous studies have operationalized complexity in problematic ways and do not take into account a representative sample of languages exhibiting the full range of complexity found cross-linguistically. Improving upon these issues, we demonstrate that not all languages of the world exhibit optimality in the sense described by previous work. Our results serve as a reminder that putative universals rooted in communicative efficiency are not always as straightforward as they seem at first blush. We conclude with an appeal to take into account the full range of historical contingencies that shape linguistic evolution in at times unexpected ways.

Background

We take as a point of departure recent work (Denić & Szymanik, 2024) which argues that exact numeral systems must manage a trade-off between (1) a pressure to keep the numeral lexicon as small as possible and (2) a need to keep numeral terms from 1 to 99 on average as morphosyntactically simple (i.e., involving the smallest combination of numeral terms and arithmetic operators) as possible. As an example of this dual pressure, languages like Japanese can generate numerals 1–99 with a small inventory of elements (the modifiers 1–9 and the base 10), but numbers above 20 are generated by concatenating together three elements (e.g., *ni jū go* '25'), whereas a language with dedicated forms for each decade would have a larger inventory but would only need to combine two elements to express similar quantities (e.g., Swedish *tjugo fem* '25'). Drawing upon earlier work on communicative efficiency in numeral systems (Xu et al., 2020), the authors use a computational algorithm to generate a large number of hypothetical unattested numeral systems ranging from highly non-optimal (e.g., having a large numeric lexicon and high average morphosyntactic complexity) to optimal, and show that attested numeral systems lie along a Pareto frontier estimated from these synthetic data, meaning that they exhibit a near-optimal solution to the trade-off between the two variables.

Subsequent work has modified aspects of this approach. Yang and Regier (2025) note that the evolutionary algorithm used by Denić and Szymanik (2024) to generate hypothetical numeral systems, based on Hurford's (1975) grammar of numerals, lacks some properties of natural languages, including the use of suppletive forms (e.g., Ukrainian *sorok* '40', originally 'bundle of [40] pelts', replacing earlier *čotyr-desjat* '4×10'; Fałowski 2011). Altering the algorithm to incorpo-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

rate this and other properties, the authors find that the original results hold. Prasertsom et al. (2025) call into question the metrics used in the original work, proposing alternative measures of complexity argued to better capture differences between attested and unattested systems (for the purpose of this paper we keep with the original study in investigating the trade-off between lexicon size and average morphosyntactic complexity, given the straightforwardness of operationalizing these properties of numeral systems under the approach we use).

In this paper, we address and improve upon a major limitation of Denić and Szymanik's (2024) study involving the treatment of the data. As mentioned previously, one dimension of complexity which systems conceivably work to minimize is the size of the numeric lexicon. Curiously, the authors stress that they are concerned only (in their words) with the number of lexicalized numeral concepts found in a language, and not with allomorphic variation in form-meaning relationships (e.g., *ten* vs. *-teen* vs. *-ty*). In our view, this decision is not well motivated: at the very least, it strikes us as misguided to operationalize complexity whilst excluding information regarding unpredictable formal variation that learners of numeral systems must confront. After all, a central concern of most grammatical frameworks regardless of their theoretical orientation involves accounting for the exceptions that a language user must learn in order to produce and comprehend utterances.

Aside from these concerns, a more serious problem can be identified. As we will demonstrate, the premise that analyses can focus on lexicalized number meanings while ignoring form-meaning mappings is fundamentally flawed: any attempt to isolate lexicalization from the presence of irregular morphology is untenable, as the two phenomena are intertwined. It is not always possible to determine how Denić and Szymanik's (2024) coding scheme relates to the formal properties of the numerals under study, since they do not provide explicit information regarding how they arrived at their analyses. Nonetheless, a few examples will serve to illustrate the inconsistency that results from attempting to ignore irregular morphology.

In many languages, the early teens are irregular exceptions within an otherwise transparent and predictable system. In the authors' coded data, English *eleven* is treated as a non-decomposable term lexicalizing the quantity 11; however, German *elf* is represented as decomposed 10 + 1. This analysis entails that *elf* contains allomorphs (*e-* and *-lf*?) of *ein(s)* '1' and *zehn* '10', even though the word contains no element formally resembling or cognate to the latter term, descending from Proto-Germanic **aina-lifa-*, cf. Gothic *ain-lif* '11', lit. 'one left (over)' (Kroonen, 2013). Persian *bist* '20' is analyzed as multimorphemic 2×10, despite not being straightforwardly decomposable into the numbers *du* '2' and *dah* '10'. Whether or not these are coding errors is beside the point: under an analysis that concerns itself with the lexicalization of number meanings but is not tasked with explaining the distribution of formal elements found in a numeral system (i.e., which allomorphs appear in which contexts), any analysis is valid for any number, as a concept can have any number of formal realizations. Concretely, there is nothing preventing us from analyzing *sorok* '40', arguably a simplex form, as a complex form (e.g., *sor-ok* '4×10', with *sor-* and *-ok* treated as allomorphs of '4' and '10', respectively).[1] At worst, this analytical framework has the potential to collapse, treating all languages as identical in their numeral morphosyntax. At best, it leads to the inconsistencies we have highlighted above. This coding scheme likely captures some interesting variation stemming from the cross-linguistic use of different bases (where consistently coded) and arithmetic operators (e.g., subtraction), but remains far from characterizing the range of formal complexity that the world's languages display.

We adopt a more principled approach, reconcilable with most morphological theories, which distinguishes between phonologically predictable and unpredictable (i.e., lexically specified) allomorphy resulting from sound change or suppletion (that is, the introduction of unrelated lexical material; Kim 2016). The former phenomenon consists of alternations such as that between Spanish *diec*[i]*siete* '17' and *diec*[j]*ocho* '18', where a segment surfaces as [j] before vowels and [i] elsewhere. An extreme example of phonologically unpredictable allomorphy is found in Dhivehi (dhiv1236, Maldives; Fritz 2002; Gnanadesikan 2017), where '7' has *s-* and *h*-initial allomorphs seen in *satā-ra* '17' and *hatā-vīs* '27' (numerals are base-final). There is no straightforward phonological context that predicts when *s-* and *h*-initial variants surface; rather, speakers must remember to use the correct variant for the correct lexical item, just as Ukrainian speakers remember to use the suppletive form *sorok* '40' instead of a hypothetical form composed from the numerals 4 and 10. These unpredictable elements add to the amount of lexical material that is needed to characterize the entire system, and hence reflect the burden the system places on language users. We do not assume a holistic representation for irregulars, but allow for the possibility that variant forms are stored as sub-word units that can be abstracted from multiple irregular forms (e.g., English *thir-*, *fif-*, *-teen*, etc.).

We work with phonemically transcribed, segmented data with manually annotated morphological glosses, making it straightforward to operationalize the morphological complexity of numerals in a language's system as well as the size of the system's inventory of lexical items. Hand-curated approaches of this sort often suffer from the problem of non-unique analyses (Round, 2017; Wu et al., 2019); e.g., an annotator may choose an analysis *twelve* or *twe-lve*. We remedy this issue in several ways, including via inter-annotator reliability values and by showing that our results are robust to different approaches (permissive vs. strict) to morphological segmentability.

[1] Denić and Szymanik (2024) identify English, Persian, Russian, and other languages as unclear cases which are excluded from some analyses. This raises a further concern that certain results are biased in favor of simpler systems within their sample.

Language	Concept	Source Form	Surface Form	Underlying Form	Morpheme Gloss
French	TEN	dix	dis	/dis/	TEN
French	SIXTEEN	seize	sɛz	/sɛz/	SIXTEEN
French	SEVENTEEN	dix-sept	disɛt	/dis-sɛt/	TEN SEVEN
French	EIGHTEEN	dix-huit	dizɥit	/dis-ɥit/	TEN EIGHT
Balochi	TEN	də	də	/də/	TEN
Balochi	SIXTEEN	šāzdə[g]	ʃãzdə	/ʃãz-də/	SIX_2 TEN
Balochi	TWENTY	bist	bist	/bist/	TWENTY
Balochi	TWENTY SIX	bist w šəšš	bistʊʃəʃː	/bist-ʊ-ʃəʃː/	TWENTY AND SIX_1
Dhivehi	TWELVE	bāra	baːra	/baː-ra/	TWO_2 TEN_2
Dhivehi	NINETEEN	onavihi	onavihi	/ona-vihi/	MINUS_ONE TWENTY_1
Dhivehi	TWENTY	vihi	vihi	/vihi/	TWENTY_1
Dhivehi	TWENTY TWO	bāvīs	baːviːs	/baː-viːs/	TWO_2 TWENTY_2

Table 1: Selected examples of annotation scheme from French (stan1290; Doherty 2016), Balochi (west2368; Barker and Mengal 1969) and Dhivehi (dhiv1236; Gnanadesikan 2017), slightly modified for increased clarity. Phonologically predictable allomorphs (e.g., French [di] before /s/ vs. [diz] before a voiced segment vs. [dis] elsewhere) are coded as underlyingly the same, whereas allomorphs with no clear phonological conditioning environment (e.g., Balochi [ʃəʃː] vs. [ʃãz] '6', Dhivehi [vihi] vs. [viːs] '20') are coded as separate underlying morphemes. We exclude cognate codings, as cognacy is not relevant to the research question investigated here, which depends only on the synchronic configurations exhibited by languages.

Materials and methods

Manually segmented and glossed data

We employ a revised and enlarged collection of manually segmented and glossed numerals, building upon the "Compositional Structures in Numeral Systems" (CoSiNuS, Rubehn et al. 2025) database and extending the sample to include the forms for numbers 1 to 99 in 52 genetically and geographically diverse languages. Word forms are transcribed in a unified transcription system with words segmented into morphemes, which are assigned to language-internal cognate sets and annotated with morpheme glosses (Hill & List, 2017), as in Table 1. The morpheme annotation scheme distinguishes between different unpredictable morphemic variants pertaining to the same meaning; e.g., English *five* (FIVE_1) and *ten* (TEN_1) vs. *fif-teen* (FIVE_2 TEN_2) and *fif-ty* (FIVE_2 TEN_3). The annotation furthermore distinguishes between phonologically unpredictable and predictable allomorphy by checking the literature for phonological rules that could generate variant surface forms from the same underlying representation, explicitly documenting potentially controversial decisions. All analyses were checked and critiqued by multiple coders. We calculated the inter-annotator agreement scores for both a 10% sample of all data points and for the complete lists of numerals for 10 randomly sampled languages. Agreement between annotators was 87% overall (N = 529) and 94.5% for the complete datasets of 10 languages (N = 1003), indicating consistency of the segmentation scheme. Data are stored in Cross-Linguistic Data Formats (Forkel et al., 2018).

Some cases in which annotator-level idiosyncrasies may impact results warrant discussion. For example, in situations where we are forced to infer the operation of productive synchronic processes (e.g., assimilation, voicing, vowel raising, reduction, etc.), our analyses may underestimate the number of unpredictable allomorphs in a given system. In general, our analyses tend to lump rather than split; we treat German *-zig* and *-ßig* (in *zwan*[ts]*ig* '20', *drei*[s]*ig* '30', *vier*[ts]*ig* '40', etc.) as underlyingly the same, since the latter occurs only after a vowel, despite the lack of secure evidence for similar alternations elsewhere in the lexicon. However, as we discuss later, this does not harm the overall thrust of (and in fact further strengthens) the argument we make. Another issue is the following: while we have taken pains to be consistent in only treating recurrent phonological elements as segmentable morphemes, it could be argued that we over-segment in some cases, inflating morphosyntactic complexity. As an example, we segment the early teens in Dhivehi as *egā-ra* '11', *bā-ra* '12', *tē-ra* '13', etc. on the basis of recurrent *-ra* (TEN_2); this creates a unique morpheme *egā-* found only in this context (*bā-* and *tē-* appear elsewhere). It could be argued that *egāra* should not be segmented. For robustness, we recode the data such that segmented forms where at least one morpheme is found only once are treated as holistic, non-decomposed irregular forms. We present results of analyses carried out using this recoded data with its NARROW construal of segmentable allomorphy in addition to the original BROAD representation.

Lexicon size and morphosyntactic complexity

We closely follow the studies cited above in calculating lexicon size and morphosyntactic complexity for the broad and narrow representations of our data. We calculate lexicon size by counting up the number of unique unpredictable allomorphs in a numeral system. Morphosyntactic complexity is a weighted average for numerals 1 to 99 defined as proportional to $\sum_{i=1}^{99} |w_i| \cdot i^{-2}$, where $|w_i|$, the length of the term denoting cardinality i, is defined as the number of morphemes and arithmetic operators (covert and overt) that it contains (i.e., German *ein-und-zwan-∅-zig* '1 + 2 × 10' and English *twen-∅-ty-∅-one* '2 × 10 + 1' have the same length), and i^{-2} is

proportional to the approximate frequency of cardinality i in usage. We exclude forms not pertaining to numeral meaning or arithmetic operations (e.g., morphemes involved in subtraction) from the calculation of lexicon size and complexity, as they generally reflect language-specific artifacts that have nothing to do with numeral complexity (e.g., citation forms for numerals in Georgian contain a case suffix which should not count toward the size of the numeral lexicon or the number of morphemes in a numeral term). In cases where languages contain multiple terms for the same cardinality, our annotation schema marks a default form, which we use in calculating these quantitites.

Synthetic numeral systems

We generate hypothetical non-optimal and optimal numeral systems using Yang and Regier's (2025) implementation of the evolutionary algorithm originally proposed by Denić and Szymanik (2024). The algorithm generates hypothetical numeral systems using Hurford's (1975) grammar of numeral phrases, randomly sampling candidate bases, digits, and irregular suppletive forms to denote numerals. We modify this implementation to allow up to 98 bases, digits, and suppletive forms in a given system. We evolve 100 hypothetical systems for 300 generations, randomly permuting the numeral grammar and selecting more optimal (i.e., reducing complexity for one variable without increasing complexity for the other) outcomes of this permutation, the end result of which consists of 100 hypothetical numeral systems lying along the Pareto frontier representing the optimal tradeoff between lexicon size and complexity. For ease of visualization, we compare attested systems only to the optimal systems populating this frontier.

Note that the representation of numeral systems are in some ways not directly comparable to the attested data that we analyze; while hypothetical systems contain suppletive forms replacing a numeral expression with a single non-compositional form (e.g., German *elf* '11', Kashmiri *kəh* '11', Persian *bist* '20'), this process does not simulate unpredictable allomorphy at the sub-word level (e.g., *thir-teen*). Nonetheless, the optimal languages generated serve as a valid baseline against which to compare attested systems under our analysis, as they indicate, given a numeral lexicon size, how a highly economical language would distribute numeral forms across expressions.

Mixture modeling

We employ Bayesian mixture modeling to infer whether the data in our sample can be partitioned into separate components exhibiting different types of behavior. Our rationale is that some languages in our sample may cluster into different groups characterized by more vs. less optimal relationships between lexicon size and average morphosyntactic complexity. We assume a mixture of regression models with the following generative story: for each language $\ell = 1..\mathcal{L}$, a component label is generated $z_\ell \sim \text{Categorical}(\boldsymbol{\pi})$, where $\pi_k = K^{-1} : k = 1..K$ (K denotes the number of mixture components; we fixed this parameter to mitigate identifiability and convergence issues); the average morphosyntactic complexity value for a given language is generated on the basis of the sampled component label and the language's lexicon size, $\texttt{morphosyntactic\_complexity}_\ell | z_\ell = k \sim \text{LogNormal}(\alpha_k + \beta_k \cdot \texttt{lexicon\_size}_\ell, \sigma_k)$, with $\text{Normal}(0, 1)$ priors over $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and a $\text{HalfNormal}(0, 1)$ prior over $\boldsymbol{\sigma}$. During inference we marginalize out the discrete parameter z, and reconstruct it from the fitted posterior samples of continuous parameters. We use RStan (Carpenter et al., 2017) to fit mixture models to the values for these variables found in the attested languages in our sample for $K = 1..4$ over four chains of 2000 iterations of the No U-Turn Sampler (NUTS), discarding the first half of samples and assessing convergence via standard diagnostics (e.g., Gelman & Rubin, 1992). We carry out model comparison via leave-one-out expected log pointwise density (ELPD) values (Vehtari et al., 2017) and model stacking (Yao et al., 2017), which averages predictive distributions of different models to generate weights representing their relative predictive power. In general, models with $K > 2$ did not converge, so we compare a two-component mixture model to a null one-component model (the equivalent of a standard regression model). The two-component mixture model is a better fit to the data, with a higher expected log pointwise density (broad: $\text{ELPD}_{K=2} = 122.2, \text{ELPD}_{K=1} = 107.8, \Delta_{\text{ELPD}} = -14.4, SE_\Delta = 6.6$; narrow: $\text{ELPD}_{K=2} = 132.8, \text{ELPD}_{K=1} = 104.2, \Delta_{\text{ELPD}} = -28.6, SE_\Delta = 6.3$) and stacking weight (broad: 0.870 vs. 0.130; narrow: 1.0 vs. 0.0). We classify languages according to the posterior probability of their component membership.

Results

Figure 1 plots lexicon sizes against average morphosyntactic complexity values for attested languages in our sample, under broad and narrow coding schemes. The Pareto frontier representing optimal tradeoffs between lexicon size and complexity estimated by the evolutionary algorithm is shown by the black curve along the lower left side of the plot. As is typically the case in studies of this sort, most languages lie slightly above the Pareto frontier rather than directly on the curve itself. This configuration represents a near-optimal solution to the problem of minimizing both lexicon size and complexity. At the same time, a number of points lie relatively far from the curve. To our knowledge, despite a large body of literature comparing near-optimal systems to optimal and non-optimal ones, there are no carefully developed criteria for determining how far away from the frontier an attested language must lie in order to no longer be considered near optimal. Furthermore the magnitude of these deviations is small and difficult to interpret in meaningful terms.

We turn to the results of our mixture model in order to assess whether there are statistically detectable differences in the behavior of different languages that can be interpreted in a straightforward fashion. The points representing joint values of the variables of interest in attested languages are colored

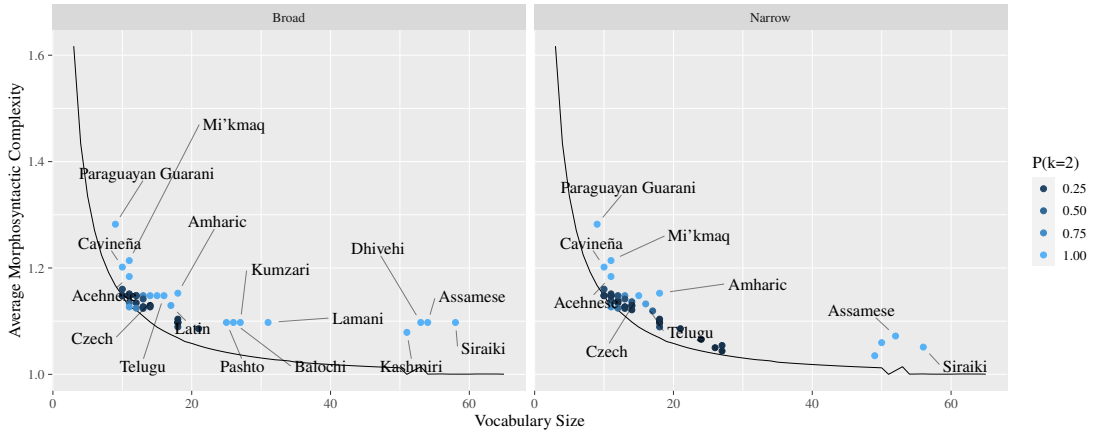

Figure 1: Lexicon sizes plotted against morphosyntactic complexity values for languages in our sample, under broad and narrow data formats, with selected languages labeled. Colors represent posterior probabilities of component membership (lighter values indicate higher $P(k = 2)$). The solid line represents the Pareto frontier estimated by the evolutionary algorithm.

according to their posterior probability of mixture component membership. Visually, there is a clear separation between data points assigned to component $k = 1$ (dark points, tightly straddling the frontier) and $k = 2$ (light points, further from the frontier). The parameters of the different mixture components are even more telling. The mean slope of the lognormal-link regression function associated with component $k = 1$ is negative (broad: $\hat{\beta}_1 = -0.005$, 95% percentile interval $= [-0.006, -0.004]$; narrow: $\hat{\beta}_1 = -0.005$, 95% PI $= [-0.005, -0.005]$), while the posterior distribution of the slope for component $k = 2$ is effectively (or at least closer to) zero (broad: $\hat{\beta}_2 = -0.001$, 95% PI $= [-0.002, -0.0007]$; narrow: $\hat{\beta}_2 = -0.002$, 95% PI $= [-0.003, -0.001]$). What this indicates is that for languages associated with component $k = 2$, average morphosyntactic complexity is not expected to decrease (or increase) as lexicon size increases.

This is borne out visually when we inspect the behavior of this component: when we look at languages with a lexicon size of around 30 or greater, they seem to follow a straight line with a slope of zero. In this group of languages, no matter how large their inventory, they do not decrease in complexity. These are languages in our sample (primarily Indo-European and Semitic) containing irregular forms, either in the teens (e.g., English *eleven*, Spanish *once*, Persian *yaz-dah*) or beyond, all of which increase the lexicon size of the system without adding to the number of usable bases and thus reducing morphosyntactic complexity. Outlying Indo-Aryan languages (Lamani, Kashmiri, Dhivehi, Siraiki) tolerate a high degree of irregularity as well. Languages like Paraguayan Guaraní exhibit higher complexity than expected given its lexicon size because of its hybrid use of different bases. The languages in $k = 1$ (dark points in a straight line showing a clearer negative relationship between the two variables of interest) consist of languages from other families with highly transparent systems and different bases (quinary, decimal, vigesimal, etc.), where the tradeoff between the two variables is far more linear.

Discussion

Our results show that while a number of languages present a near-optimal solution to the pressure to minimize both lexicon size and average morphosyntactic complexity, many languages do not, lying relatively far from the Pareto frontier and forming a distinct, statistically detectable group to the exclusion of languages with more efficient systems. Denić and Szymanik (2024) argue that while all languages' numeral systems, exact and approximate, must optimize a tradeoff between the number of distinct numerosities they can express and the informativity of the system (Xu et al., 2020), exact systems must manage an orthogonal tradeoff between lexicon size and morphosyntactic complexity. We have shown that while many exact numeral systems manage this tradeoff efficiently, a considerable number do not; this result holds despite the potential underestimation of irregularity on our part (e.g., German *-zig* vs. *-βig*) and under two data coding schemes. The key dimension underlying this difference is regularity: when taking into account the number of unpredictable elements in languages' numeral term inventories — and we have argued above that it is not possible to operationalize lexicon size without taking allomorphy into account — it turns out that certain languages tolerate a varying degree of irregularity in numeral systems with no apparent benefit (in terms of the variables discussed above). At the extreme end of this tolerance for irregularity are Indo-Aryan numeral systems; though long recognized for their unusual degree of irregularity (Bright, 1972; Comrie, 2011; Hammarström, 2010; Hurford, 1987; Overmann, 2025), these languages do not figure prominently in quantitative studies of complexity in numeral systems. The Indo-Aryan languages make up only a small, phylogenetically restricted fraction of the 7000-odd languages spoken today, yet they are spoken by a large proportion of the world's population. Although the history of these systems is relatively well documented and the historical developments affecting them are reasonably well

understood (Andrijanić, 2024; Berger, 1992), the reasons for the development of such a high degree of complexity remain unclear, as well as the mechanisms involved in its persistence for the last 1000 years or so. It is possible that a mechanistically grounded reason for this irregularity can be found, given the irregularity is argued to have several communicative benefits, such as earlier cues to form identity that allow listeners to process forms more quickly (Blevins et al., 2017). It remains to be seen whether the preservation of irregularity in these systems serves to enhance the discriminability of individual numeral forms.

Given that the irregularity exhibited by Indo-Aryan numeral systems breaks many generalizations regarding communicative efficiency, what are some properties that all numeral systems have in common? It is widely observed that morphological irregularity is more likely to emerge and persist in highly frequent words than in less frequent ones (Sims-Williams, 2021). This generalization also holds for numeral systems like that of English, where irregular suppletive forms like *twelve* have lower cardinality (and therefore higher frequency; Dehaene and Mehler 1992) than regular, decomposable forms like *ninety-two* (Brysbaert, 2005). Indo-Aryan numeral systems are different from those of English, Persian, and Neo-Aramaic in that there is no clear binary opposition between regularity and irregularity: irregular-looking forms can be found beyond the teens, all the way up to 99, and numerals seem to occupy a cline spanning from somewhat irregular to highly irregular. Nevertheless, it has been shown that this generalization holds for Indo-Aryan numeral systems as well: despite their higher overall irregularity in comparison to other systems, higher-cardinality (lower-frequency) numerals are more predictable (according to a variety of computational models) and re-use more recurrent elements than lower-cardinality ones (Cathcart, 2017, 2026). Impressionistically, this makes sense when inspecting forms in selected Indo-Aryan languages: for example, forms in the Dhivehi twenties display ample allomorphic variation (e.g., *vihi* 'twenty', *ekā-vīs* 'twenty one', *sab-bīs* 'twenty six') in comparison to the seventies (e.g., *haẘtari* 'seventy', *ekā-haẘtari* 'seventy one', *sa-haẘtari* 'seventy six'). This intuition has robust statistical support.

This issue is related to some of the pressures involved in minimizing morphosyntactic complexity discussed by Denić and Szymanik (2024). The average morphosyntactic complexity measure proposed by the authors and used in this paper, though an aggregate measure, reflects the tendency to keep more frequent elements at lower cardinalities shorter (in terms of the number of elements combined) than less frequent elements at higher cardinalities. The authors link this pressure to a number of general principles in language use, invoking Zipf's law of abbreviation (Zipf, 1949) — the tendency of more frequent elements to be shorter — as well as the idea that more frequently used elements are accessed whole as opposed to stored in a decomposed representation. We generally concur with these ideas: it seems to be the case that less frequent numerals of higher cardinality are phonologically longer, more transparent, and more decomposable than more frequent ones. This can be operationalized in a variety of ways, including those that do not necessitate the manual segmentation of words into morphemes (Baayen et al., 2018; Beniamine & Guzmán Naranjo, 2021; Guzmán Naranjo, 2024), as we have done here, and future research can aid in understanding the evolutionary mechanisms that maintain this state of affairs. However, as we have shown here, the drive to keep more frequent numerals phonologically shorter and more transparent does not necessarily correspond to differences in the number of segmentable morphological units in a given word. Our analyses find quasi-recurrent elements (and evidence of multimorphemicity) even in the highly irregular teens. Furthermore, we show that the pressure to minimize the average morphosyntactic complexity of utterances does not interact with the pressure to minimize complexity of the numeral lexicon in a meaningful way across languages: there exist a number of languages with a large inventory of terms that do not distribute them in a way that minimizes length in morphemes. Ultimately, we are likely observing not only the effects of cognitive pressure but of social factors as well (cultural diffusion, trade networks, contact resulting in hybrid systems). Our understanding of pressures toward efficiency should accommodate these local historical contingencies.

Conclusion

This study questions the notion that all numeral systems provide a near-optimal solution to the twin pressures of lexicon size and morphosyntactic complexity minimization. Our research departs from previous work by taking into account the full range of morphological complexity these systems display, which we argue cannot be ignored. Our results show that irregularity complicates the picture, particularly the high degree of irregularity displayed by Indo-Aryan languages.

While our results challenge the lexicon size/complexity trade-off from a synchronic perspective — there exist contemporary languages that fly in the face of this generalization — they do not rule out the possibility that these pressures exert themselves in the context of diachronic linguistic evolution. Many Indo-Aryan languages have been highly irregular for at least a millennium, but what is to say that they won't evolve to balance these pressures over the next thousand years? Further work is needed to elucidate the dynamics of these pressures in language change, and to better understand the origins and life cycle of deviations from efficiency.

Supplementary Materials

The data underlying our analyses are curated on GitHub under https://github.com/numeralbank/cosinus (v2.0) and archived on Zenodo under https://doi.org/10.5281/zenodo.18684984. The code used in our analyses is curated on GitHub under https://github.com/chundrac/numeral-complexity-tradeoff and archived on Zenodo under https://doi.org/10.5281/zenodo.18696081.

Acknowledgments

We are grateful to David Yang for guidance on modifying his code. C.C. was supported by the NCCR Evolving Language (SNSF Agreement No. 51NF40_180888), and gratefully acknowledges Swiss National Science Foundation grant No. 207573. A.R., K.B., L.C., A.K., C.Y.M., D.S., and J.-M.L. were supported by the ERC Consolidator Grant ProduSemy (PI J.-M.L., Grant No. 101044282, see https://doi.org/10.3030/101044282). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.

Author contributions: C.C.: Conceptualization, Methodology, Software, Visualization, Data curation, Writing (Original Draft); A.R., J.-M.L.: Project administration, Data curation, Writing (Review & Editing); K.B., L.C., K.P.D., A.K., C.Y.M., D.S.: Data curation; J.M.: Validation.

References

Allassonnière-Tang, M., & Her, O.-S. (2020). Numeral base, numeral classifier, and noun: Word order harmonization. *Language and Linguistics*, *21*(4), 513–558.

Andrijanić, I. (2024). Hindi cardinal numerals in a historical and comparative perspective. In I. Andrijanić & M. Browarczyk (Eds.), *Between language and literature: Hindī in classroom and beyond* (pp. 151–170). FF Press.

Baayen, R. H., Chuang, Y.-Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The mental lexicon*, *13*(2), 230–268.

Barker, M. A.-R., & Mengal, A. K. (1969). *A course in Baluchi*. Institute of Islamic Studies, McGill University.

Beniamine, S., & Guzmán Naranjo, M. (2021). Multiple alignments of inflectional paradigms. *Proceedings of the Society for Computation in Linguistics 2021*, 216–227.

Berger, H. (1992). Modern Indo-Aryan. In J. Gvozdanović (Ed.), *Indo-European numerals* (pp. 243–287). Mouton de Gruyter.

Blevins, J. P., Milin, P., & Ramscar, M. (2017). The Zipfian paradigm cell filling problem. In *Perspectives on morphological organization* (pp. 139–158). Brill.

Bright, W. (1972). Hindi numerals. In M. E. Smith (Ed.), *Studies in linguistics in honor of George L. Trager* (pp. 222–230). Mouton.

Brysbaert, M. (2005). Number recognition in different formats. In J. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 23–42). Psychology Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76(1)*(1), 1–32.

Cathcart, C. (2017). Decomposability and Frequency in the Hindi/Urdu Number System. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, London*, 1733–1738.

Cathcart, C. (2026). Complexity counts: Global and local perspectives on Indo-Aryan numeral systems. *Journal of Cultural Cognitive Science*. https://doi.org/10.1007/s41809-026-00197-x

Comrie, B. (2011). Typology of numeral systems. https://lingweb.eva.mpg.de/channumerals/TypNum_Latest_21ho.pdf

Comrie, B. (2022). The arithmetic of natural language: Toward a typology of numeral systems. *Macrolinguistics*, *10*(1), 1–35.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.

Denić, M., & Szymanik, J. (2024). Recursive numeral systems optimize the trade-off between lexicon size and average morphosyntactic complexity. *Cognitive Science*, *48*(3), e13424.

Doherty, L. (2016). Wikidict-fr: Wikipedia Bilingual Reference Data (French). https://github.com/open-dict-data/wikidict-fr

Fałowski, A. (2011). The East-Slavonic sorok '40' revisited. *Studia Etymologica Cracoviensia*, *16*(1), 7–15.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., & Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, *5*(180205), 1–10. https://doi.org/https://doi.org/10.1038/sdata.2018.205

Fritz, S. (2002). *The Dhivehi language*. Ergon.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Gnanadesikan, A. E. (2017). *Dhivehi: The language of the Maldives* (A. E. David, Ed.). De Gruyter Mouton.

Guzmán Naranjo, M. (2024). An analogical approach to the typology of inflectional complexity. *Journal of Language Modelling*, *12*.

Hammarström, H. (2010). Rarities in numeral systems. In J. Wohlgemuth & M. Cysouw (Eds.), *Rethinking universals: How rarities affect linguistic theory* (pp. 11–60). De Gruyter Mouton.

Hill, N. W., & List, J.-M. (2017). Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting*, *3*(1), 47–76. https://doi.org/https://doi.org/10.1515/yplm-2017-0003

Hurford, J. R. (1975). *The linguistic theory of numerals*. Cambridge University Press.

Hurford, J. R. (1987). *Language and number: The emergence of a cognitive system*. Blackwell.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kim, R. I. (2016). From sound change to suppletion: Case studies from Indo-European languages. *Slovo a slovesnost*, *77*(4), 354–370.

Koile, E., & Blasi, D. (2025). Decimal systems around the world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *380*(1937).

Kroonen, G. (2013). *Etymological dictionary of Proto-Germanic*. Brill.

Núñez, R. E. (2017). Is there really an evolved capacity for number? *Trends in cognitive sciences*, *21*(6), 409–424.

O'Shaughnessy, D. M., Gibson, E., & Piantadosi, S. T. (2022). The cultural origins of symbolic number. *Psychological review*, *129*(6), 1442.

Overmann, K. A. (2025). Numbers in India. In *Cultural number systems: A sourcebook* (pp. 163–169). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-83383-0_25

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an amazonian indigene group. *Science*, *306*(5695), 499–503.

Prasertsom, P., Silvi, A., Culbertson, J., Johansson, M., Dubhashi, D., & Smith, K. (2025). Recursive numeral systems are highly regular and easy to process. *arXiv preprint arXiv:2510.27049*.

Round, E. R. (2017). Matthew K. Gordon: Phonological typology. *Folia Linguistica*, *51*(3), 745–755. https://doi.org/doi:10.1515/flin-2017-0027

Rubehn, A., Rzymski, C., Ciucci, L., Bocklage, K., Kučerová, A., Snee, D., Stephen, A., Van Dam, K. P., & List, J.-M. (2025). Annotating and inferring compositional structures in numeral systems across languages. *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, 29–42.

Sims-Williams, H. (2021). Token frequency as a determinant of morphological change [Publisher: Cambridge University Press]. *Journal of Linguistics*, 1–37. https://doi.org/10.1017/S0022226721000438

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*(5), 1413–1432.

Wu, S., Cotterell, R., & O'Donnell, T. J. (2019). Morphological irregularity correlates with frequency. *arXiv preprint arXiv:1906.11483*. http://arxiv.org/abs/1906.11483

Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, *4*, 57–70.

Yang, D., & Regier, T. (2025). Re-examining the tradeoff between lexicon size and average morphosyntactic complexity in recursive numeral systems. *Proceedings of the 47th Annual Conference of the Cognitive Science*, 863–869.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2017). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*. https://doi.org/10.1214/17-BA1091

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.