Ilja A. Seržant | George A. Moroz

Universal attractors in language evolution provide evidence for the kinds of efficiency pressures involved

Suggested citation referring to the original publication:
Humanities & Social Sciences Communications 9 (2022) 1, Art. 58
DOI https://doi.org/10.1057/s41599-022-01072-0
ISSN 2662-9992

Journal article | Version of record

Secondary publication archived on the Publication Server of the University of Potsdam:
Zweitveröffentlichungen der Universität Potsdam : Philosophische Reihe 180
ISSN: 1866-8380
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-583976
DOI: https://doi.org/10.25932/publishup-58397

Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-022-01072-0 OPEN

Universal attractors in language evolution provide evidence for the kinds of efficiency pressures involved

Ilja A. Seržant[1]✉ & George Moroz[2]

Efficiency is central to understanding the communicative and cognitive underpinnings of language. However, efficiency management is a complex mechanism in which different efficiency effects—such as articulatory, processing and planning ease, mental accessibility, and informativity, online and offline efficiency effects—conspire to yield the coding of linguistic signs. While we do not yet exactly understand the interactional mechanism of these different effects, we argue that universal attractors are an important component of any dynamic theory of efficiency that would be aimed at predicting efficiency effects across languages. Attractors are defined as universal states around which language evolution revolves. Methodologically, we approach efficiency from a cross-linguistic perspective on the basis of a world-wide sample of 383 languages from 53 families, balancing all six macro-areas (Eurasia, North and South America, Australia, Africa, and Oceania). We explore the grammatical domain of verbal person–number subject indexes. We claim that there is an attractor state in this domain to which languages tend to develop and tend not to leave if they happen to comply with the attractor in their earlier stages of evolution. The attractor is characterized by different lengths for each person and number combination, structured along Zipf's predictions. Moreover, the attractor strongly prefers non-compositional, cumulative coding of person and number. On the basis of these and other properties of the attractor, we conclude that there are two domains in which efficiency pressures are most powerful: strive towards less processing and articulatory effort. The latter, however, is overridden by constant information flow. Strive towards lower lexicon complexity and memory costs are weaker efficiency pressures for this grammatical category due to its order of frequency.

[1] University of Potsdam, Potsdam, Germany. [2] National Research University Higher School of Economics, Moscow, Russian Federation. ✉email: serzant@uni-potsdam.de

Introduction

Language provides a means for communication. It is crucial that communication be not only successful but also efficient, i.e., with minimal effort for both parts and obeying high transmission accuracy (Gibson et al., 2019).

We distinguish between two linguistic levels at which the effects of efficiency obtain: *online*, contextual effects produced by individual speakers and *offline* effects that are found in the mental grammar and lexicon of speakers (see Jaeger and Buz (2018)). Online effects are found, e.g., in the pronunciation of words in a spontaneous speech: if predictable in the particular context, words may be articulated with less care and be reduced (inter alia, Aylett and Turk, 2004; Aylett and Turk, 2006; Pluymaekers et al., 2005). Online effects pertain to particular communication events and individual speakers. By contrast, offline effects emerge over time via conventionalization of the more efficient and, therefore, more frequently selected variant in the online efficiency management (Gibson et al., 2019; Kirby, 2001; Pierrehumbert, 2001; Diessel, 2007; Seyfarth, 2014; Currie et al., 2018; Seržant, 2021b). Crucially, offline effects pertain to the population level of commonly shared linguistic culture. They are thus subject not only to the individual-level effects but are also constrained by the complex sociological and interactional effects emerging on the population level.

Moreover, conventionalized, offline strings are not static but constantly changing over time (Hopper, 1987; Bybee and Hopper, 2001; Seržant, 2021a). Change may be driven by semantic change or various external and sociolinguistic factors (Seržant, 2021b). As a consequence, the distribution and frequency of lexical and grammatical items is not at all stable. Thus, the question arises whether efficiency pressures themselves may essentially change over time, and, accordingly, whether the outcomes of these processes may be expected to largely parallel each other within and across languages.

Offline efficiency effects have most prominently been observed in the lexicon. The Zipfian effect that the length of a word tends to be a function of its inverse frequency (Zipf, 1935; Bentz and Ferrer-i-Cancho, 2016) or informativity (Piantadosi et al., 2011) is the result of various historical processes from which the more efficient word lengths have been conventionalized. The association with the original form is often lost here, as in English *pants* from *pantaloons* or *pub* from *public house* ("opacification" in Kanwal et al., 2017). This is especially true of grammatical items, which tend to be entirely dissociated from their origin (e.g., the indefinite article *a* and its source *one*).

In addition to the distinction between online and offline efficiency effects, efficiency pressures operate on different stages of production. While the information-theoretic approach to efficiency primarily relies on the articulatory efficiency (boiling down to the length of the message), it does not take into account the processing efficiency or the planning efficiency, which may require signs that are less efficient from the articulatory perspective. For example, when minimizing the articulatory effort online, the speaker has to assess at the same time whether or not the particular reduced form will achieve its communicative goal before it actually goes into articulation. This also requires that larger chunks must first be pre-planned before a cue goes into production (Bornkessel-Schlesewsky and Schlesewsky, 2014: p. 107; Jaeger and Tily, 2010: p. 325). This requires processing costs. Potential ambiguities are also costly for the hearer who can correctly interpret an efficient but ambiguous cue only once enough context has been uttered (Bornkessel-Schlesewsky and Schlesewsky, 2014: p. 107; Jaeger and Tily, 2010: p. 324). Thus, ambiguities created by articulatory efficient signs may require more processing effort because speech is generated and decoded incrementally. Languages respond to these processing efforts by developing systems of context-independent cues to resolve potential rather than actual ambiguity (cf. Malchukov, 2008; Seržant, 2019). This unavoidably leads to mismatches between the length of a cue and its predictability in certain contexts (Seyfarth, 2014; Sóskuthy and Hay, 2017).

To sum up, efficient cues result online from an interaction of various trade-offs between the processing, planning and articulatory efficiency pressures (see, however, Levshina, 2021). Offline-efficient cues, in turn, emerge on the population level via selection and conventionalization of one of the efficient variants emerged online. Here, social factors play an important role as well.

There is no integrative theory combining these different efficiency effects and their conventionalization mechanisms that would be able to predict cross-linguistic data. Here, we suggest that an essential component of such a theory is universal attractors. Attractors are a notion borrowed from dynamic models of cognition, in which they are defined as states that related states prefer to develop into but not develop away from (Norton, 1995: p. 56). We extend this notion by using it for diachronic linguistic processes. Attractors are universal properties of conventionalized cues within a particular domain. The motivation behind attractor states is that languages tend to organize meanings and functions space in certain ways. A corollary is that languages tend to develop semantically and functionally similar items that, in effect, have similar distributional frequencies and are therefore subject to similar efficiency pressures across languages.

In this paper, we provide evidence for the attractor in one particular grammatical domain: subject indexing on the verb as found, for example, in Latin: *vide-ō* (see-1SG) meaning "I see", *vidē-s* (see-2SG) "you see", *vide-t* (see-3SG) "(s)he sees", *vidē-mus* (see-1PL) "we see", *vidē-tis* (see-2PL) "you see", *vide-nt* (see-3PL) "they see". We show that language evolution revolves around this attractor. The attractor is characterized by at least two universal properties: (1) preferred absolute lengths of the indexes and (2) preference for the cumulative coding (i.e., non-compositional, atomic coding). The attractor is internally structured and caused by efficiency pressures, which are thus universal.

Data

In order to establish the attractor in this domain we manually compiled a database We restricted our study to intransitive verbs only. We analyzed the six subject indexes (endings/prefixes/clitics) that encode the person and number (and in some languages masculine gender, as well) of the subject participant on the verb. We excluded the dual. The six person–number indexes found in the morphologically unmarked (typically present) tense were entered into the database: first person singular (1SG), second person singular (2SG), 3SG, 1PL, 2PL, 3PL. In total, these data have been manually collected from 383 languages from 53 families, covering all six macro-areas of the world: Eurasia, North and South America, Australia, Africa, and Oceania (Fig. 1, Moroz, 2017, the entire list is presented in the Appendix 1 in the online supplement; the entire dataset is published in Seržant, 2021c).

Methods

15 families contribute each 10-50 languages to the database in order to exclude language-specific effects and in order to control for family effects. Other families are represented with only few languages (sometimes only one, e.g., with isolates). Two extremely large and diverse families are split into subfamilies: Nuclear Trans New Guinea (Sogeram, Awyu-Dumut, Oceanic, and

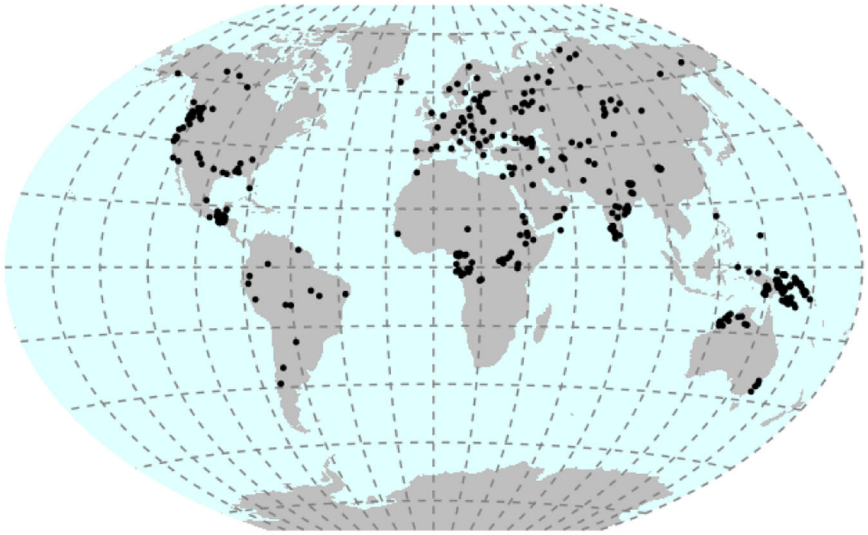

Fig. 1 Languages in the database. Dots represent languages in our database.

(other) Nucear Trans New Guinea) and Afroasiatic (Semitic and (other) Afroasiatic). Likewise, Atlantic-Congo family is represented only by its Bantu subfamily. Furthermore, in order to explore the dynamics we have entered the person–number indexes of the respective proto-languages (Proto-Indo-European, Proto-Athabaskan, Proto-Semitic, Proto-Salishan, Proto-Muskogean, Proto-Bantu, Proto-Dravidian, etc.; 15 in total) found in the authoritative literature.[1] Since there is a great deal of controversy on the reconstruction of the Proto-Tibeto-Burman indexes, we adopted only the reconstructions for two subfamilies Gyalrongic and Kiranti, over which there is no controversy in the literature. The remaining 38 families were excluded from the diachronic analysis because no commonly accepted reconstructions for these families have been found. All computations have been carried out in the R environment (R Core Team, 2015).

Attractor lengths were modeled with Poisson mixed effects model with person and number as fixed effects. The results from a model that neglects the information on person and number significantly differ from the observations (Fisher exact test). When measuring length we only relied on the number of segments (proxied as the number of letters except for French and English). Long segments have been assigned 1.5.

Evolution towards the attractor was tested by comparing the proto and the modern forms in order to see whether verbal person–number indexes tend to move towards (or remain within) the attractor or away from it. In order to do so, we established for each form whether or not the difference between its modern length and the attractor length became smaller than the length difference between the attractor and the proto-form. Whenever the difference remains the same and the length of the proto-form is very close to the attractor we counted it as a movement towards attractor. After we thus obtained the direction of change for each modern form we applied a logistic mixed effects model predicting the direction of change with person and number as fixed effects and clade as a random effect.

Preference for cumulative coding was established by testing the diachronic preference for and against compositionality. The data points were divided into four categories for each person: (i) *no compositionality*—compositionality is found neither in the proto-form nor in the modern form; (ii) *compositionality disappears*—compositionality is present in the proto-form and disappeares in the modern form; (iii) *compositionality remains*—compositionality is present in both the proto-form and in the modern form; (iv) *compositionality appears*—compositionality is absent from the proto-form but appears in the modern form. Subsequently, we applied a logistic mixed effects model to obtain the probabilities for the three persons to disprefer compositionality.

The *properties of the attractor* thus obtained are interpreted with regard to efficiency effects at different stages of production (articulatory, processing, memory retrieval, etc.).

Results

Indexes lengths for each person–number combination do not vary much across languages. The dispersion around the average lengths across languages is quite small. This is illustrated in Fig. 2. We evaluated the Poisson regression model with person and number as fixed effects and clade as a random effect in order to obtain an exact formula for the observed relation between length of the index, person, and number. 1SG form was selected as a baseline for the regression. The lme4 (Bates et al., 2015) formula used for this model is as follows:

index length ~ person * number + (1|clade)

The overall predictions of our model are presented in Fig. 2, with the estimated values and a 95% confidence interval (model printouts are presented in the supplementary materials). Both variables person and number are statistically significant. Since all variables are statistically significant and differ from zero, we can conclude that our attractor model is supported by our data. This allows us to compute the lengths of the attractors. The absolute average lengths computed by the model are presented in Fig. 2.

While the lengths predicted by the model for all families represent the static evidence for the attractor, we have also tested whether languages tend to develop towards this state if they happen to deviate from it in their proto-languages or whether the lengths are preserved in the modern languages if the proto-language already adhered to the attractor. It has been repeatedly argued that linguistic universals are not language states but rather the accumulation of the diachronic processes and the mechanisms of change that lead to these states (Bybee, 1988; Bybee, 2006; Bybee, 2008; Creissels, 2008; Cysouw, 2010; Dunn et al., 2011; Givón, 1979; Greenberg, 1966; Greenberg, 1978; Haspelmath, 1999; Maslova, 2000; Maslova, 2004; Cristofaro, 2012; Cristofaro, 2014; Bickel et al., 2014).

If the attractor lengths exist as suggested by the model on the basis of the synchronic data above, then the attractor should also become visible in the transitional probability of languages to adhere to the attractor lengths over the course of time. In order to

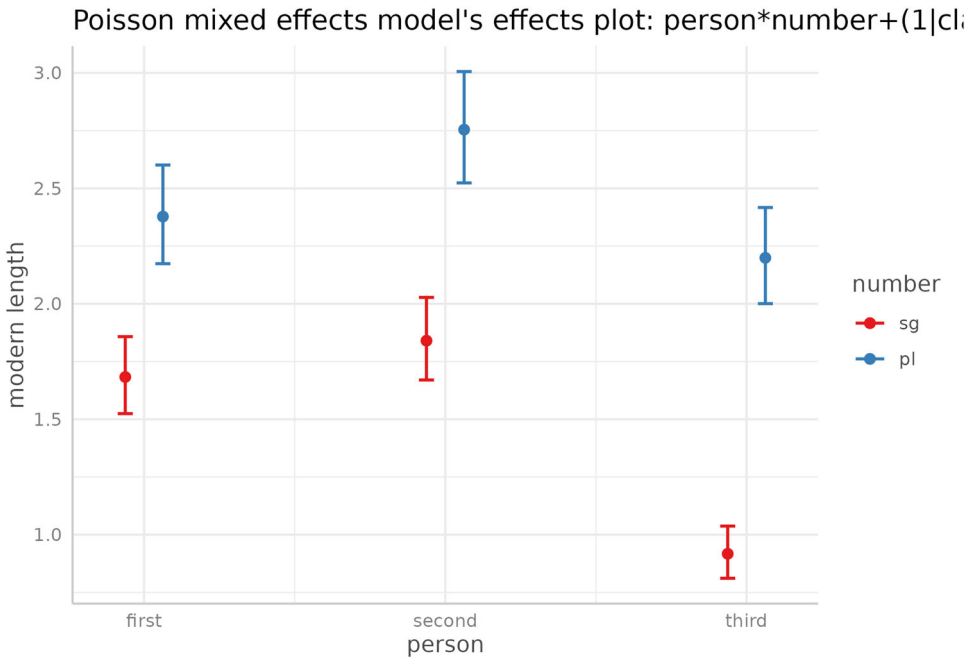

Fig. 2 Predictions of the Poisson mixed effects model for the number of segments based on person and number (clade is used as a random effect).

test whether there is indeed a diachronic pressure towards the attractor lengths, we have compared two idealized diachronic stages: Stage 0 and Stage 1. Stage 0 consists of the lengths of each of the six person–number indexes in the proto-language reconstructed by the historical-comparative method in the authoritative literature for 15 (sub)families (see fn. 1 for the references). Stage 1 is the lengths of each of the six person–number indexes across all modern languages of the respective (sub)family (10–50 languages per family). The lengths at Stage 0 is in principle subject to accidental, language-specific pressures, since there is only one proto-language per family. By contrast, the lengths at Stage 1 may be taken as indicative of universal pressures, since we take 10 to 50 modern languages per family, thus leveling out possible language-specific effects.

We find that the modern forms, on average, develop towards the attractor over the course of time. We also do not observe any significant source determination. Modern languages either "fix" the original proto-lengths via (i) shortening or (ii) enlarging, or they retain the lengths if these adhered to the attractor lengths already in the proto-language. For example, Uralic had singular proto-forms that were too short: 1SG *-m*, 2SG *-n*, 3SG *-ø* (Janhunen, 1982: p. 35). Accordingly, some modern Uralic languages enlarged them to two segments in the 1SG and 2SG and to one segment in the 3SG (e.g., Saami, Erzya, Komi-Permyak). Observe that this enlargement is differential: in contrast to the singular forms, the first and second plural forms (both three segments in Proto-Uralic) have not been enlarged in modern Uralic languages on average. The enlargement only takes place if the proto-forms considerably deviate from the attractor state.

By contrast, families with proto-forms considerably longer than the attractor shorten their lengths. For example, second singular in Proto-Indo-European was three segments (**-e-si*). It was accordingly shortened to 1.57 segments on average in the modern Indo-European languages. The same applies to first and second person in Proto-Mayan: with 2.5 (a segment plus a long segment) it was somewhat too long and was accordingly shortened to around two segments on average in the modern languages. At the same time, the respective plural proto-form was somewhat too short with two segments and was enlarged in a number of modern Mayan languages (yielding the modern average of 2.64 segments). Finally, indexes adhering to the attractor remain largely unchanged as to their lengths. For example, the length of 1SG in modern Sogeram, Athabaskan, or Semitic languages does not deviate considerably from its proto-forms. We thus observe that indexes are not randomly affected by reduction or enlargement (via, for example, analogical extensions).

In order to model the tendencies between Stage 0 and Stage 1, we computed for each language whether or not its indexes have changed toward the attractor estimated in the previous model, as a binary variable: *moving towards* or remaining in the attractor vs. *not moving towards the attractor*. Subsequently, we applied a logistic mixed effects model to predict the probability of movement towards (and remaining within) the attractor by person and number. The 1SG form was again selected as a baseline for the regression. The lme4 (Bates et al., 2015) formula used for this model is as follows:

movement towards attractor or being in the attractor range ~ person * number + (1|clade)

The overall predictions of our model are presented in Fig. 3, with the estimated values and a 95% confidence interval (model printouts can be found in the Supplement).

The model reveals that in all person–number combinations there is a high probability to obey the attractor. There is no statistically significant difference among persons. We conclude that the model supports our hypothesis that indexes are obeying the attractor lengths in their diachronic developments. Note that the probability of obeying the attractor length of the given person is extremely high in the singular forms (around 90–100%) and less so in the plural forms (around 65–90%). The distinction between singular and plural forms is also statistically significant.

To summarize, despite continuous processes of various phonetic and morphological changes and restructurings (Seržant, 2021a), there is a stable blueprint in the coding of person–number indexes. Regardless of the lengths in the respective proto-language, modern languages on average stick to the attractor lengths by the right combination of diachronic processes leading to reduction, enlargement, or retention (see Moroz, 2021 for an exception). Importantly, while many studies since Zipf (1935)

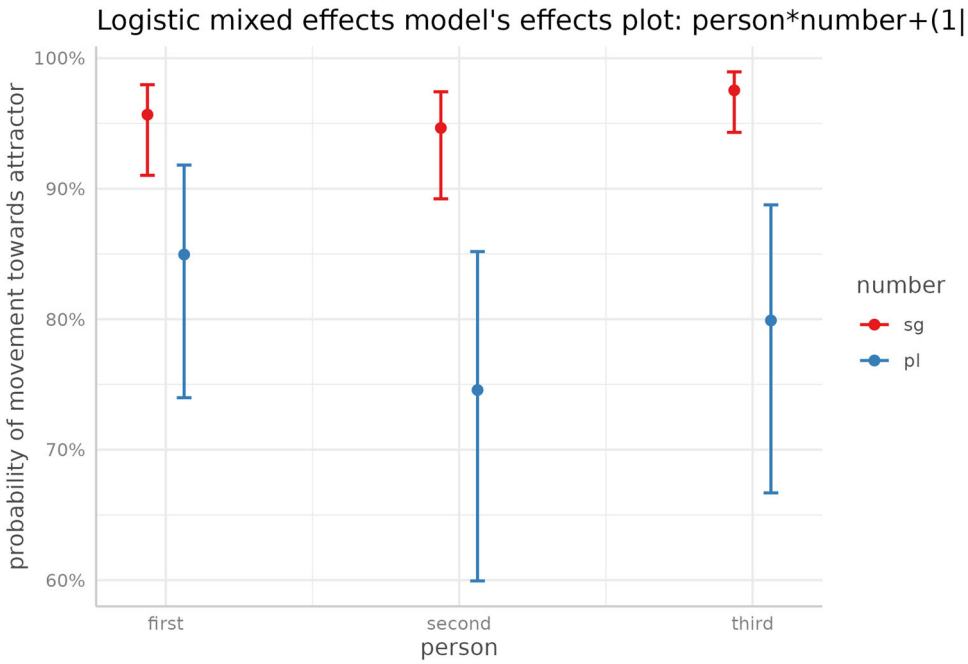

Fig. 3 Logistic mixed effects model's predictions for the number of segments based on person and number (clade is used as a random effect).

assume that frequency effects on coding length only manifest themselves via reduction (Diessel, 2007; Jaeger and Tily, 2010; Bybee, 2001; Bybee, 2003; Cohen Priva and Jaeger, 2018), the length optimization discussed here is a more complex process that may result not only from reduction but from retention or enlargement as well. For example, the Polish 1Pl *-my* (from Proto-Slavic **-mū*) is the result of the lengthening of the final vowel, which was originally hyper-short *-mŭ* (with the reduced vowel *ŭ*) in Proto-Slavic and thus much shorter than the attractor. The lengthened variant most probably emerged by analogy to the independent 1PL pronoun *my* (<*mū*) 'we' already in Early Slavic. Importantly, no other person-number combination underwent this kind of lengthening.

The second universal property of the attractor is the preference for compositionality. Compositionality is found when the person (1st vs. 2nd vs. 3rd) and the number (singular vs. plural) are transparently and separately coded. For example, the indexes in Russian show no compositionality (i.e., are cumulative), cf. 1SG *-u* vs. 1PL *-m* or 2SG *-š'* vs. 2PL *-te*. By contrast, Maalula, a Western Aramaic language does show compositionality: 2SG *či-* vs. 2PL *či- … -un* or 3SG *yi-* vs. 3PL *yi-…un*. In this language, second person is marked by *či-*, third person by *yi-* and number is marked by zero in the singular and by *-un* in the plural. These forms are thus compositional.

We coded changes in compositionality into four values: no compositionality (neither the proto-language nor the modern language has compositionality), compositionality disappears (compositionality of the proto-language decreased in the modern language), compositionality remains (both the proto-language and the modern language have some compositionality and its degree remains unchanged), compositionality appears (the modern language develops some compositionality). Results are presented on Fig. 4.

Both green bars stand for the preference of compositionality while both blue bars indicate dispreference for compositionality. Overwhelmingly, compositionality tends to be avoided. We also applied logistic mixed effects model to predict compositionality of the modern form depending on the person and the compositionality of the proto-form. For this, we merged the blue values into the value "dipreferred" and the green values into the value "preferred." The lme4 (Bates et al., 2015) formula used for this model is as follows:

compositionality of modern language ~ person * compositionality of proto-language + (1|clade)

The overall predictions of our model are presented in Fig. 5, with an estimated values and a 95% confidence interval (see supplement).

It follows from Figs. 4 and 5 that compositionality is dispreferred in the long run. The model predicts an extremely high probability of non-compositional coding (over 95%) for each person.

Discussion

Although the coding of indexes in particular languages is subject to various independent and language-specific processes including various types of reduction, reanalyses, analogical extensions, etc. (Seržant, 2021a), there are universal pressures that channel their development over time. More specifically, we provided synchronic and dynamic evidence for a universal attractor in the domain of indexing. The attractor is characterized by the absolute lengths for each person–number combination (Fig. 2) and cumulative (non-compositional) coding. Finally, subject indexes are almost never optional in the languages of the world as has been shown earlier (Karlsson, 1986; Siewierska, 1999). From these characteristics of the attractor the following conclusions about the universal principles constraining the interaction between underlying efficiency pressures can be drawn.

First, despite an extremely high corpus frequency, indexes nevertheless are not all equal in their lengths. The absolute lengths are structured: (i) the third person tends to be the shortest, and (ii) the plural indexes are longer than their respective singular indexes (Greenberg, 1966: pp. 33–38). These asymmetries correlate with the asymmetries in the corpus frequencies of these forms as predicted by Zipf's Law of Abbreviation: the more frequent form is shorter than the less frequent one. Consider the corpus frequencies from the oral subcorpus of the Russian National Corpus (216,112 words) as a proxy (Table 1). In comparison to other persons, third person is the most frequent person in both number sets, with 69% in the singular and 62% in

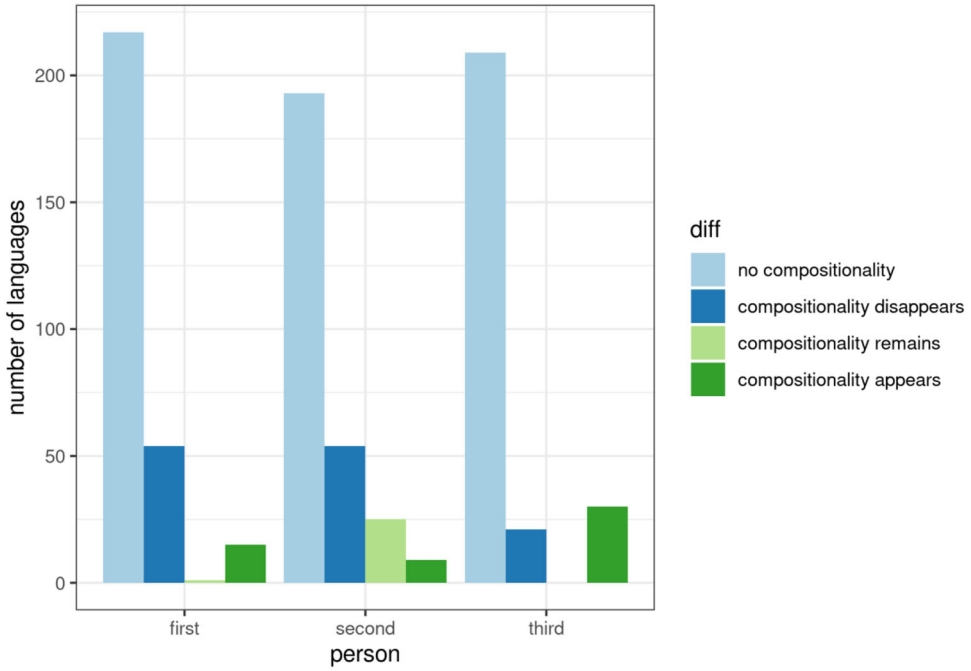

Fig. 4 Number of languages that increased/decreased number of compositional persons.

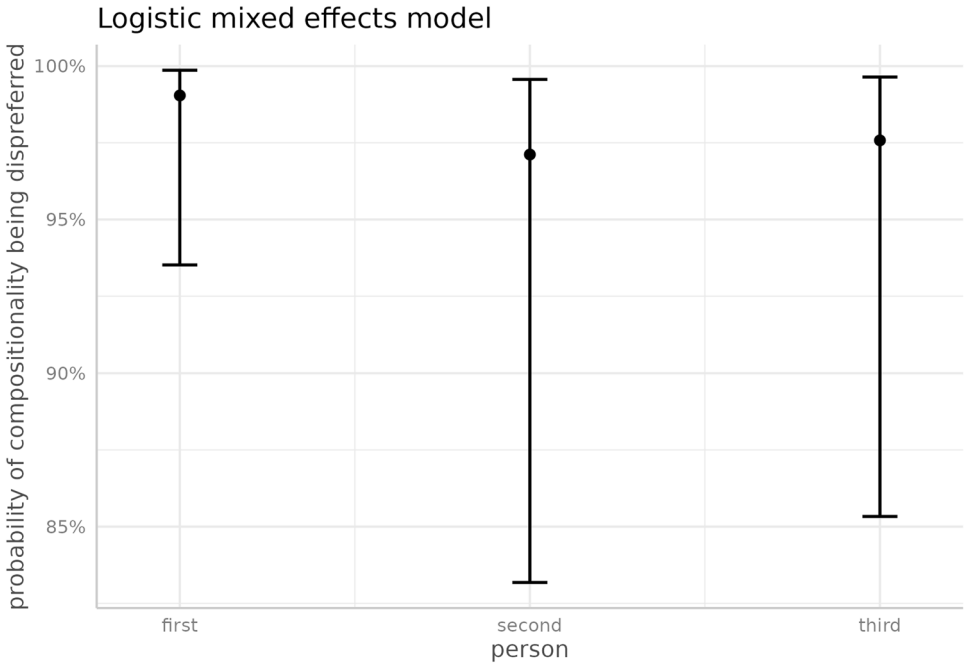

Fig. 5 Probability of compositionality of the modern form depending on the person and compositionality of the proto-language.

the plural. Likewise, the singular forms are much more frequent than the plural ones, with 69% singulars vs. 31% plurals of all forms. Both frequency asymmetries (3rd vs. 1st or 2nd and singular vs. plural) are statistically significant ($p = 0.002$, χ^2). Similar frequency asymmetries have been obtained for other languages, such as spoken Spanish (Bybee, 1985: p. 71), Finnish (on the basis of *olla* "to be" in Karlsson, 1986: 24), and some other languages (Greenberg, 1966: p. 37).

These figures show that articulatory efficiency plays an important role here: the more expected the sign is the shorter it is. Nevertheless, zero is not preferred. The most frequent third-person form is more frequently coded with a segment than with zero as one would expect if only the articulatory efficiency were at play. We did not observe any dynamic bias towards zero (only the weaker, reverse statement is true: zeros, if at all, are more probable in the third singular than elsewhere, Siewierska, 2010; Bickel et al., 2015). In fact, some subfamilies even entirely replace the third-person zero inherited from their proto-languages. For example, Proto-Uralic had zero-coded third-person singular index (Janhunen, 1982: p. 35) while a number of modern Uralic languages, including the entire Finnic subfamily, developed a non-zero coding here.

While zero would be the most efficient in terms of articulation, non-zero coding of the third-person singular must be motivated

Table 1 Person–number frequencies in the oral subcorpus of the RNC.

	Singular	Plural
1	26% (2.276)	15% (601)
2	5% (471)	23% (926)
3	**69%** (6.021)	**62%** (2.493)
Total	**69%** (8768)	**31%** (4020)

Bold indicates the most frequent combinations.

by processing and planning efficiency overriding articulation ease. Sending the hearer a non-zero phonetic cue facilitates the processing effort on the part of the hearer and thus increases the chances of a successful transmission of information. A non-zero form is also more planning-efficient for the speaker because it provides a straightforward link from meaning to coding, while zero is inherently ambiguous by being linked to various meanings and domains. Non-zero coding also alleviates the planning process because it makes the assessment of whether or not the context provides enough information unnecessary.

Secondly, it also is the planning efficiency that must be responsible for the fact that verbal indexes are almost never optional in the languages of the world (Siewierska, 1999; Haig, 2018). This obligatoriness yields redundant uses in those contexts that provide enough information for the identification of the subject referent, as in *ven-ī, vid-ī, vic-ī* "came-1SG", "saw-1SG", "conquered-1SG" (the last two occurrences of -1SG are increasingly redundant because they can be guessed from the previous context anyway). Planning efficiency overrides articulatory efficiency here as well.

Thirdly, the most articulatory efficient paradigm that would also warrant unambiguous information transmission would not require the plural to have longer forms than the singular. Thus, theoretically a morphological system of coding all six distinctions (1SG, 2SG … 3PL) with one segment—e.g., 1SG *-a*, 2SG *-t*, 3SG *-i* (or zero), 1PL *-k*, 2PL *-o*, 3PL *-r*—would perfectly fulfill the requirement of accurate information transmission under the lowest articulatory effort. Thus, the effect of articulatory efficiency alone does not explain why cross-linguistically the plural forms require more segments than the singular forms if they all may be sufficiently disambiguated by just one segment. Multiple segments, however, allow the speakers to gain more production time and the hearer more comprehension time with the less expected meanings (plural in this case). The longer forms of the plural fulfill here the function of according the message with constant information flow (Aylett and Turk, 2004; Levy and Jaeger, 2007; Pluymaekers et al., 2005; Uniform Information Density hypothesis in Coupé et al., 2019). In turn, the selection of particular phonetic segments serves the distinguishability function.

Fourthly, while it is known that high-frequency items as opposed to low-frequency items do not require transparent, compositional coding (Kirby, 2001: p. 108; Christiansen and Chater, 2008: p. 499), our cross-linguistic diachronic evidence suggests that items as frequent as person–number indexes in fact prefer cumulative coding (number and person being coded by one atomic sign): those families that were not compositional in the proto-language (e.g., Indo-European) did not develop compositionality in any of the modern languages, and some of those families that did have compositionality in the proto-language (e.g., Awyu-Dumut) removed it in the modern languages at least to some extent. This "opacification" is also observed in independent words, such as *pub* from *public house* (Kanwal et al., 2017). Cumulative coding requires higher complexity of the lexicon and comes at higher memory and learnability costs because it requires six signs (1SG, 2SG… 3PL) while compositional coding would require only four signs (three signs for the three persons and one plural sign applicable to all of them). While both options are equally informative, it is only the first one that is cross-linguistically preferred. This fact allows uncovering the specific efficiency processes involved. Languages structure their lexica optimally such that the trade-off between the processing costs and the lexicon complexity is resolved within the Pareto frontier either in favor of higher processing costs (more compositional) or in favor of higher lexicon complexity and memory costs (more cumulative coding) (Kemp and Regier, 2012; Kemp et al., 2018; Xu et al., 2020). Yet, languages prefer the specific choice (corner) within the Pareto frontier in high-frequency domains such as the indexing domain: processing efficiency outweighs lexicon complexity and, thus, memory (and learnability) costs with linguistic items of this order of frequency. The reason for this is that higher processing costs are not efficient with high-frequency items that are easily learnable and retrievable from the memory anyway (Kirby, 2001: p. 109). This ties in with Kemp et al. (2018: p. 114) who claim that the preference for the cumulative coding within the Pareto frontier is found when the lexical domain is important for the culture, if "important for the culture" means that the items of this lexical domains are frequent in this culture (similarly in Xu et al., 2020 for number signs). We conclude from this that processing ease outweighs lexicon simplicity and, thus, memory (and learability) costs with linguistic items of this order of frequency.

To sum up, first, we have established that there is a universal attractor state for indexing around which the evolution revolves. Second, the properties of the attractor uncover two domains in which efficiency pressures are most powerful: strive towards less processing and articulatory effort while strive towards lower lexicon complexity and lower memory costs are weaker efficiency pressures for this grammatical category due to its order of frequency. Having said this, our evidence is cross-linguistic comparative evidence. Ideally, our conclusions should be supported by experimental evidence.

Data availability

All data analyzed are included in the manuscript and supplementary information file.

Received: 13 September 2021; Accepted: 24 January 2022;
Published online: 17 February 2022

Note

1 Proto-Indo-European (Meier-Brügger, 2010: pp. 173–184), Proto-Turkic (Róna-Tas, 1998: p. 75; Old Turkic in Abduraxmanov, 1997: p. 68; Erdal, 2004: p. 232; Tuguševa, 1997: p. 59), Proto-Mayan (Bricker, 1977: p. 2; Schele, 1982: p. 9), Proto-Uralic (Honti, 2010: p. 21; Janhunen, 1982: p. 35; Kulonen, 2001; Laanest, 1982 [1975]: pp. 229–30), Proto-Dravidian (Andronov, 2009: pp. 224–231), Proto-Semitic (Hasselbach, 2004: p. 32; Huehnergard, 2000; Lipiński, 2001: p. 378), Proto-Oceanic (Blust, 1972; François, 2016: p. 32; Ross, 1988: p. 366, 2002 : 60; Starosta et al., 1981), Proto-Bantu (Meeussen, 1967: pp. 97–99; Schadeberg, 2003 [2014]: p.151), Proto-Sogeram (Daniels, 2015: p. 155), Proto-Awyu-Dumut (Wester, 2014: pp. 78–85), Proto-Athabaskan (Hoijer, 1971: pp. 127–132; Leer, 2006: p. 429), Proto-Muskogean (Booker, 1980: p. 33), Proto-Worroran (McGregor and Rumsey, 2009: p. 68), and Proto-Salishan (Newman, 1979: p. 213, 1980: p. 156), Proto-Kiranti and Proto-rGyalrongic (DeLancey, 2010: p. 15, 2011: p. 2, 2014; Jacques, 2012, 2016; LaPolla, 2003: p. 30).

References

Abduraxmanov GA (1997) Karaxanidsko-ujgurskij jazyk. In: Tenišev ÈR, Poceluevskij EA, Kormušin IV, Kibrik AA (eds.) Jazyki Mira. Tjurkskie jazyki. "Kyrgyzstan", Bishkek, pp. 64–74

Andronov MS (2009) A comparative grammar of the dravidian languages. Beiträge zur Kenntnis südasiatischer Sprachen und Literaturen 7. Otto Harrassowitz, Wiesbaden

Aylett M, Turk A (2004) The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Lang Speech 47:31–56
Aylett M, Turk A (2006) Language redundancy predicts syllable duration and the spectral characteristics of vocalic syllable nuclei. J Acoust Soc Am 119:3048–3058
Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48
Bentz CH, Ferrer-i-Cancho R (2016) Zipf's law of abbreviation as a language universal. In: Bentz CH, Jäger G, Yanovich Y (eds.) In: Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics. University of Tubingen, online publication system. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/6855814
Bickel B, Witzlack-Makarevich A, Zakharko T (2014) Typological evidence against universal effects of referential scales on case alignment. In: Bornkessel-Schlesewsky I, Malchukov A, Richards M (eds.) Scales and hierarchies: a cross-disciplinary perspective on referential hierarchies. De Gruyter, Mouton, Berlin, pp. 7–44
Blust RA (1972) Proto-Oceanic addenda with cognates in non-Oceanic Austronesian languages: a preliminary list. WPLUH 411:1–43
Booker KM (1980) Comparative muskogean: aspects of proto-muskogean verb morphology. University of Kansas dissertation, Lawrence, KS
Bornkessel-Schlesewsky I, Schlesewsky M (2014) Competition in argument interpretation: evidence from the neurobiology of language. In: MacWhnney B, Malchukov A, Moravcsik E (eds.) Competing motivations in grammar and usage. Oxford University Press, pp. 107–126
Bricker VR (1977) Pronominal inflection in the Mayan languages. Occasional Paper 1. Middle American Research Institute, New Orleans
Bybee JL (1988) The diachronic dimension, chapter 13. In: Hawkins JA (ed.) Explaining language universals. OUP, pp. 350–379
Bybee JL (2001) Phonology and language use. Cambridge University Press, Cambridge
Bybee JL (2003) Mechanisms of change in grammaticization: the role of frequency. In: Joseph BD, Janda RD (eds.) The Handbook of Historical Linguistics. Blackwell, Oxford, pp. 602–623
Bybee JL (2006) From usage to grammar: the mind's response to repetition. Language 82(4):711–733
Bybee JL (2008) Formal universals as emergent phenomena: the origins of structure preservation. In: Good J (ed.) Language universals and language change. Oxford University Press, pp. 108–121
Bybee J, Hopper P (2001) Introduction to frequency and the emergence of linguistic structure. In: Bybee J, Hopper P (eds.) Frequency and the emergence of linguistic structure [Typological studies in language 45]. John Benjamins, pp. 1–27
Bybee J (1985) Morphology: A Study of the Relations between Meaning and Form. Amsterdam/Philadelphia: John Benjamins
Christiansen MH, Chater N (2008) Language as shaped by the brain. Behav Brain Sci 31:489–558
Cohen Priva U, Jaeger TF (2018) The interdependence of frequency, predictability, and informativity in the segmental domain. Linguist Vanguard 4(2):1–13
Coupé CH, Oh YM, Dediu D, Pellegrino F (2019) Different languages, similar encoding efficiency: comparable information rates across the human communication niche. Sci Adv 2594
Creissels D (2008) Direct and indirect explanations of typological regularities: the case of alignment variations. Folia Linguistica 42(1):1–38
Cristofaro S (2012) Cognitive explanations, distributional evidence, and diachrony. Stud Lang 36(3):645–670
Cristofaro S (2014) Competing motivation models and diachrony: what evidence for what motivations? In: MacWhnney B, Malchukov A, Moravcsik E (eds.) Competing motivations in grammar and usage. Oxford University Press, Oxford, pp. 282–298
Currie KH, Hume E, Jaeger TF, Wedela A (2018) The role of predictability in shaping phonological patterns. Linguist Vanguard 4(s2):1–15
Cysouw M (2010) On the probability distribution of typological frequencies. In: Ebert CH, Jäger G, Michaelis J (eds.) Math Lang. Springer, Heidelberg, pp. 29–35
Daniels, D (2015) A Reconstruction of Proto-Sogeram. Phonology, Lexicon, Morphosyntax. A dissertation in partial satisfaction of the requirements for the degree Doctor of Philosophy in Linguistics. Santa Barbara: University of California
DeLancey S (2010) Towards a history of verb agreement in Tibeto-Burman. Himalayan Linguist 9(1):1–38
DeLancey S (2011) Agreement prefixes in Tibeto-Burman. Himalayan Linguist 10(1):1–29
DeLancey S (2014) Second person verb forms in Tibeto-Burman. Linguist Tibeto-Burman Area 37(1):3–33
Diessel H (2007) Frequency effects in language acquisition, language use, and diachronic change. New Idea Psychol 25:108–127
Dunn M, Greenhill SJ, Levinson SC, Gray RD (2011) Evolved structure of languages shows lineage-specific trends in word-order universals. Nature 473:79–82
Erdal M (2004) A grammar of old-turkic. Handbook of oriental studies. Handbuch der Orientalistik. Section eight. Central Asia. Vol 3. Brill, Leiden/Boston
François A (2016) The historical morphology of personal pronouns in northern Vanuatu. In: Pozdniakov K (ed.) Comparatisme et reconstruction: tendances actuelles. Faits de Langues. Peter Lang, Bern, pp. 25–60
Gibson E, Futrell R, Piantadosi ST, Dautriche I, Mahowald K, Bergen L, Levy R (2019) How efficiency shapes human language. Trend Cogn Sci 23(5):389–407
Givón T (1979) On understanding grammar. Academic Press, New York, NY
Greenberg JH (1966) Language universals, with special reference to feature hierarchies. Mouton, The Hague
Greenberg JH (1978) Diachrony, synchrony and language universals. In: Greenberg JH, Ferguson CA, Moravcsik EA (eds.) Universals of human language, Vol. 1: method and theory. Stanford University Press, Stanford, pp. 61–92
Haig G (2018) The grammaticalization of object pronouns: why differential object indexing is an attractor state. Linguistics 56(4):781–818
Haspelmath M (1999) Optimality and diachronic adaptation. Zeitschrift für Sprachwissenschaft 18(2):180–205
Hasselbach R (2004) Final vowels of pronominal suffixes and independent personal pronouns in semitic. J Semit Stud 49(1):1–20
Hoijer H (1971) Athapaskan morphology. In: Sawyer J (ed.) Studies in American Indian Languages. University of California Publications in Linguistics 65. University of California Press, Berkley, pp. 113–147
Honti L (2010) Personae ingratissimae? A 2. személyek jelölése az uráliban. Nyelvtudományi Közlemények 107:7–57
Hopper P (1987) Emergent Grammar. Berkley Linguistic. Society 13:139–157
Huehnergard, J. 2000. Comparative Semitic Linguistics. Unpublished. Cambridge, Mass
Jacques G (2012) Agreement morphology: the case of Rgyalrong and Kiranti. Lang Linguist 13(1):83–116
Jacques G (2016) Le sino-tibétain: polysynthétique ou isolant? Faits de langues 47(1):61–74
Jaeger TF, Tily H (2010) On language 'utility': processing complexity and communicative efficiency. Cogn Sci 2:323–335
Jaeger TF, Buz E (2018) Signal reduction and linguistic encoding. In: Fernández EM, Smith Cairns H (eds.) The Handbook of Psycholinguistics. John Wiley & Sons
Janhunen J (1982) On the structure of Proto-Uralic. Finnisch-ugrische Forschungen 44:23–42
Kanwal J, Smith K, Culbertson J, Kirby S (2017) Zipf's Law of Abbreviation and the Principle of Least Effort: language users optimise a miniature lexicon for efficient communication. Cognition 165:45–52
Karlsson F (1986) Frequency considerations in morphology. STUF –Lang Typol Univ 39(1):19–28
Kemp C, Regier T (2012) Kinship categories across languages reflect general communicative principles. Science 336:1049–1054
Kemp C, Xu Y, Regier T (2018) Semantic typology and efficient communication. Ann Rev Linguist 4:109–128
Kirby S (2001) Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. IEEE Trans Evol Comput 5:102–110
Kulonen UM (2001) Zum n-Element der zweiten Personen besonders im Obugrischen. Finnisch-Ugrische Forschungen 56:151–174
Laanest A (1982) Einführung in die ostseefinnischen Sprachen. Autorisierte Übertragung aus dem Estnischen von Hans-Hermann Bartens. Buske, Hamburg
LaPolla R (2003) Overview of Sino-Tibetan morphosyntax. In: Thurgood G, Matisoff JA, Bradley D (eds.) Linguistics of the Sino-Tibetan area: The state of the art. Pacific Linguistics Series C, 87. Department of Linguistics, Australian National University, Canberra, pp. 22–42
Leer J (2006) Na-Dene languages. In: Asher RE, Simpson JMY (eds.) The encyclopedia of language and linguistics. Pergamon, Oxford, pp. 428–430
Levshina N (2021) Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. Front Psychol 12:648200
Levy R, Jaeger FT (2007) Speakers optimize information density through syntactic reduction. Adv Neural Inform Process Syst 19:849–856
Lipiński E (2001) Semitic Languages: outline of a comparative grammar, 2ed. Peeters, Leuven
Malchukov AL (2008) Animacy and asymmetries in differential case marking. Lingua 118:203–221
Maslova E (2000) A dynamic approach to the verification of distributional universals. Linguist Typol 4(3):307–333
Maslova E (2004) Dinamika tipologičeskix raspredelenij i stabil'nost' jazykovyx tipov [Dynamics of typological distributions and stability of language types]. Voprosy jazykoznanija 5:3–16

McGregor WB, Rumsey A (2009) Worrorran revisited: the case for genetic relations among languages of the Northern Kimberley region of Western Australia. The Australian National University, Canberra
Meeussen, AE (1967) Bantu grammatical reconstructions. Africana Linguistica 3:79–121
Meier-Brügger M (2010) Indogermanische Sprachwissenschaft. 9., durchgesehene und ergänzte Auflage. Unter Mitarbeit von Matthias Fritz und Manfred Mayrhofer. De Gruyter, Berlin
Moroz G (2021) Length of East Caucasian subject indexes: a quantative research. In: Majsak TA, Sumbatova NR, Testelec YG (eds.) Durqasi xazna. Sbornik statej k 60-letiyu R. O. Mutalova. Buki Vedi, Moscow, pp. 258–282
Moroz G (2017) lingtypology: easy mapping for Linguistic Typology. https://CRAN.R-project.org/package=lingtypology
Newman S (1979) A History of the Salish Possessive and Subject Forms. Int J Am Linguist 45(3):207–223
Newman S (1980) Functional changes in the Salish pronominal system. Int J Am Linguist 46(3):155–167
Norton A (1995) Dynamics: an introduction. In: Port RF, Van Gelder T (eds.) Mind as Motion: explorations in the dynamics of cognition. MIT Press, pp. 44–68
Piantadosi ST, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. Proc Natl Acad Sci USA 108(9):3526–3529
Pierrehumbert J (2001) Exemplar dynamics: word frequency, lenition and contrast. In: Bybee J, Hopper P (eds.) Frequency effects and the emergence of lexical structure: studies in language. John Benjamins, 137–157
Pluymaekers M, Ernestus M, Baayen RH (2005) Articulatory planning is continuous and sensitive to informational redundancy. Phonetica 62:146–159
R Core Team (2015) R: A language and environment for statistical computing. Austria, Vienna. https://www.R-project.org/
Róna-Tas A (1998) The Reconstruction of Proto-Turkic and the Genetic Question. In: Johanson L, Csató ÉÁ (eds.) The Turkic Languages. CUP, Cambridge, pp. 67–80
Ross M (1988) Proto-Oceanic and the Austronesian languages of western Melanesia. Pacific Linguistics: Series C, 98. Australian National University dissertation. Research School of Pacific and Asian Studies, Canberra
Ross M (2002) Proto Oceanic. In: Lynch J, Ross M, Crowley T (eds.) The oceanic languages. Routledge, London/New York, pp. 54–91
Schadeberg T (2003) Historical linguistics. In: Nurse, D & G Philippson (eds.), The Bantu Languages. London/New York: Routledge, pp. 143–163
Schele L (1982) Maya Glyphs. The Verbs. University of Texas Press, Austin
Seržant IA (2019) Weak universal forces: the discriminatory function of case in differential object marking systems. In: Schmidtke-Bode K, Levshina N, Michaelis SM, Seržant I (eds.) Explanation in typology: diachronic sources, functional motivations and the nature of the evidence [Conceptual Foundations of Language Science 3]. Language Science Press, Berlin, pp. 149–178
Seržant IA (2021b) The dynamics of Slavic morphosyntax is primarily determined by the geographic location and contact configuration. Scando-Slavica 67(1):65–90
Seržant IA (2021a) Cyclic changes in verbal person-number indexes are unlikely. Folia Linguistica Historica 42(1):49–86
Seržant IA (2021c) Dataset for the paper "Universal attractors in language evolution provide evidence for the kinds of efficiency pressures involved" [Data set]. Version 4. Zenodo. https://doi.org/10.5281/zenodo.6028260
Seyfarth S (2014) Word informativity influences acoustic duration: effects of contextual predictability on lexical representation. Cognition 133(1):140–155
Siewierska A (1999) From anaphoric pronoun to grammatical agreement marker: why objects don't make it. Folia Linguistica 33(1/2):225–251
Siewierska A (2010) Person asymmetries in zero expression and grammatical functions. In: F Floricic (ed.), Essais de typologie et de linguistique générale. Mélanges offerts à Denis Creissels. Paris: Presses de L'Ecole Normale Supérieure, pp. 471–485
Sóskuthy M, Hay J (2017) Changing word usage predicts changing word durations in New Zealand English. Cognition 166:298–313
Starosta S, Pawley AK, Reid LA (1981) The evolution of focus in Austronesian. Paper presented to the Third International Conference on Austronesian Linguistics, Bali. Abridged version published. In: Halirn A, Carrington L, Wurm SA (eds.) Papers from the Third International Conference on Austronesian Linguistics. Vol. 2. Tracking the travellers. Dept. of Linguistics, Australian National University, Canberra, pp. 145–170
Tuguševa LJ (1997) Drevneujgurskij jazyk. In: Tenišev, ÈR, JeA Poceluevskij, IV Kormušin, AA Kibrik (eds.), Jazyki mira. Tjurkskie jazyki. Biškek: Izdatel'skij dom "Kyrgyzstan", pp. 54–63
Wester R (2014) A linguistic history of Awyu-Dumut: Morphological Study and Reconstruction of a Papuan Language Family. Doctoral dissertation, Vrije Universiteit Amsterdam
Xu Y, Liu E, Regier T (2020) Numeral systems across languages support efficient communication: from approximate numerosity to recursion. Open Mind 4:57–70
Zipf G (1935) The psychobiology of language. Routledge, London

Acknowledgements

The first author has received funding by the Heisenberg grant SE 2838/1-1 "Exploring linguistic diversity" of the German Research Foundation (Deutsche Forschungsgemeinschaft). The second author greatly acknowledges the support he has received within the Basic Research Program of the National Research University Higher School of Economics.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Ethical approval

Not applicable.

Informed consent

Not applicable.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-022-01072-0.

Correspondence and requests for materials should be addressed to Ilja A. Seržant.

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.