Consonant lengthening marks the beginning of words across a diverse sample of languages

Frederic Blum, Ludger Paschen, Robert Forkel, Susanne Fuchs, Frank Seifart

nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01988-4

Consonant lengthening marks the beginning of words across a diverse sample of languages

Received: 20 December 2023

Accepted: 14 August 2024

Published online: 24 September 2024

Check for updates

Frederic Blum [1,2] ✉, Ludger Paschen [3], Robert Forkel [1], Susanne Fuchs [3] & Frank Seifart [4,5]

Speech consists of a continuous stream of acoustic signals, yet humans can segment words and other constituents from each other with astonishing precision. The acoustic properties that support this process are not well understood and remain understudied for the vast majority of the world's languages, in particular regarding their potential variation. Here we report cross-linguistic evidence for the lengthening of word-initial consonants across a typologically diverse sample of 51 languages. Using Bayesian multilevel regression, we find that on average, word-initial consonants are about 13 ms longer than word-medial consonants. The cross-linguistic distribution of the effect indicates that despite individual differences in the phonology of the sampled languages, the lengthening of word-initial consonants is a widespread strategy to mark the onset of words in the continuous acoustic signal of human speech. These findings may be crucial for a better understanding of the incremental processing of speech and speech segmentation.

Speech is a continuous stream of acoustic signals that transmit linguistic meaning with the purpose of spoken communication. The intricate process of comprehending speech demands the sequential segmentation of the acoustic signal into discrete units such as words and phrases, which are the basic building blocks of language[1–4]. This segmentation is supported by a complex interaction of factors that operate on the levels of sound structure, lexicon and grammar, both for the speaker and for the listener. Several of these factors have been identified in previous research, but few have been studied across a wide range of languages. Most previous studies on speech production and processing focus on 'Western, educated, industrial, rich and democratic (WEIRD)' people and their languages, which undermines the potential to make species-wide generalizations about human language and cognition[5,6]. For the factors that affect speech production, this emerges as a particularly severe limitation in light of the huge variability of grammars and sound systems of the world's ~7,000 languages[7–9].

Word onsets play a special role in speech segmentation and word recognition. In the lexicon, word-initial segments are known to be more informative than later segments for distinguishing the intended word from other words[10], and listeners exploit this for continuously updating hypotheses regarding word identity and boundaries as the phonetic signal progresses[11]. At the level of phonology, word-initial positions generally exhibit more 'fortition' (stronger articulation) and fewer 'lenition' (weaker articulation) processes than word-internal or word-final positions and are thus assigned a prominent status in phonological theories[12–15]. Complex consonant clusters that are restricted to word onsets through phonotactic constraints may serve as additional cues for word segmentation[16]. However, there is considerable cross-linguistic variation in this respect, and many languages lack consonant clusters altogether. This implies that clusters cannot be a universal method to segment speech into word units. Other, more general strategies may be more relevant instead.

Acoustic features such as modulations of segment duration and changes in fundamental frequency play a major role in structuring speech into different units. Among these features, the lengthening of vowels at the ends of prosodic phrases, clauses or utterances is

[1]Department of Linguistic and Cultural Evolution, Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [2]Chair for Multilingual Computational Linguistics, University of Passau, Passau, Germany. [3]Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany. [4]Structure et Dynamique des Langues, CNRS, INALCO, IRD, Villejuif, France. [5]Institut für Deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin, Germany. ✉e-mail: frederic_blum@eva.mpg.de

https://doi.org/10.1038/s41562-024-01988-4

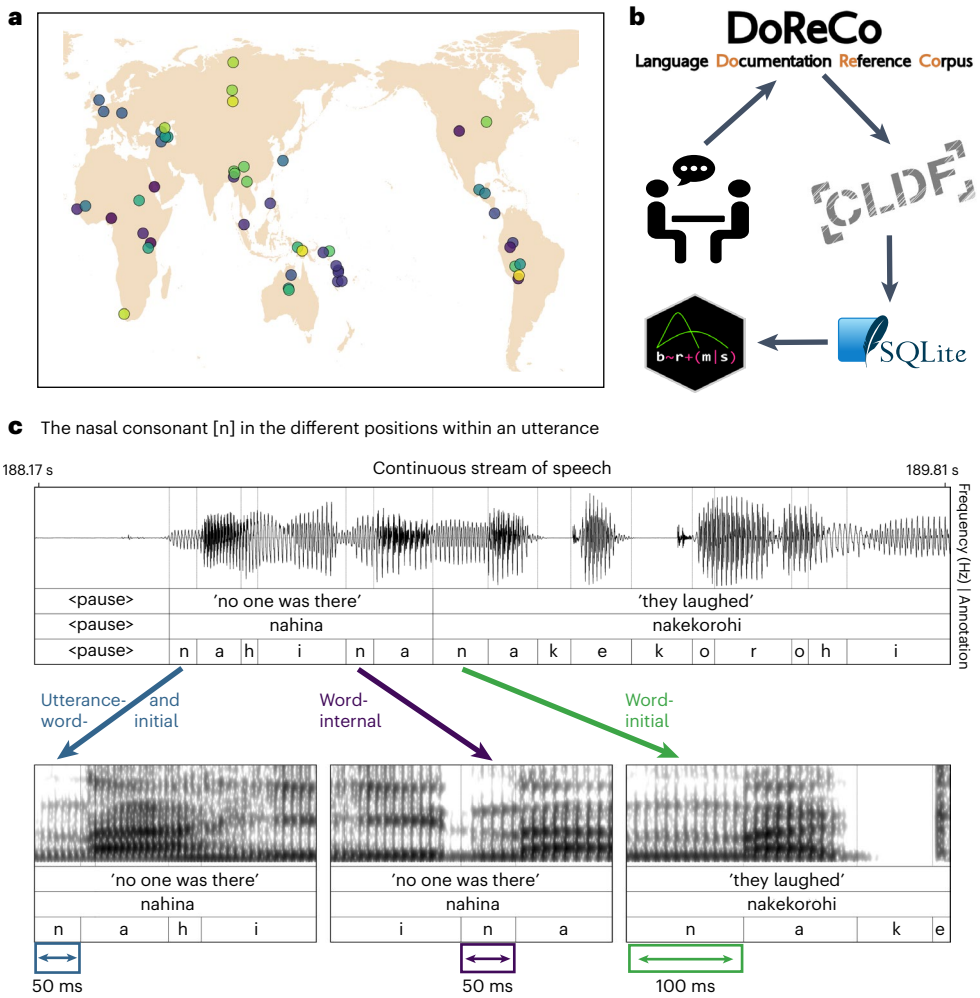

Fig. 1 | Workflow and language sample. a, The geographic distribution of the 51 languages in our sample. The colors indicate the 30 different language families in the sample. **b**, The workflow from fieldwork-based language documentation to the data sample analysed in the present study. **c**, An example (doreco_trin1278_T06, from second 188.17 to 189.81.) of word-initial lengthening in Mojeño Trinitario, an Arawakan language spoken in the Amazonian region of Bolivia[96].

attested across a wide variety of languages[17,18] and is often assumed to be universal[19]. At the word level, the acoustic properties of word-initial phones have been argued to be particularly relevant for the prosodic organization of some languages, including English, Korean and French[19–21]. The realization of these word-initial phones may depend on language-specific properties, such as prosodic systems and consonant inventories, but also on between-speaker variation[22,23]. However, so far most of the evidence for these features comes from a handful of languages, most of them Indo-European.

Two closely related features of word-initial phones that have been reported for individual languages are initial lengthening and strengthening. While initial strengthening implies a stronger articulation[19,23,24], initial lengthening refers to the duration of consonants. This is illustrated in Fig. 1 from the Amazonian language Mojeño Trinitario, which is also included in our sample. The example illustrates the same consonant /n/ in three different positions: utterance-initial (50 ms), word-internal (50 ms) and word-initial (100 ms). In artificial language learning experiments, it has been shown that speakers of Hungarian, Italian and English can use word-initial consonant lengthening as a cue to locate word boundaries[21]. Similarly, word-initial strengthening has been found to facilitate disambiguation between similar lexical items[25]. However, very little is known about the extent and degree of word-initial lengthening across languages. For words in utterance-initial position, it is not clear whether they display any additional temporal changes. In previous studies, utterance-initial consonants have been found to sometimes be lengthened or shortened, but with an overall small change in duration[26,27]. Indeed, from a functional perspective, it makes sense that no additional cue to word segmentation is necessary at the beginning of utterances, especially after a pause[26,28]. To our knowledge, the cross-linguistic evidence for initial lengthening processes remain scarce, and neither word- nor utterance-initial lengthening has been investigated in a worldwide sample of languages.

Our main research question is whether we can find cross-linguistic evidence for word-initial lengthening or shortening effects in observed speech across a wide range of languages. We also investigate whether we can find such an effect at utterance-initial positions. Following this, we analyse the cross-linguistic distribution of any emergent effects. To be able to make valid generalizations across languages, we also control for between-speaker variability and analyse the lengthening and shortening effects across segments with different places and manners of articulation.

Results

Evidence for word-initial lengthening across languages

We used a comprehensive corpus consisting of spontaneous speech from 51 languages, shown in Fig. 1a, recorded from 393 speakers (195 female, 198 male) of an age range between 16 and 100 years[29]. Of these 51 languages, 49 are spoken by non-WEIRD populations[5,6]. The languages in our sample display a wide range of sound inventories and prosodic systems and cover a wide spectrum of grammars. The main units of our

https://doi.org/10.1038/s41562-024-01988-4

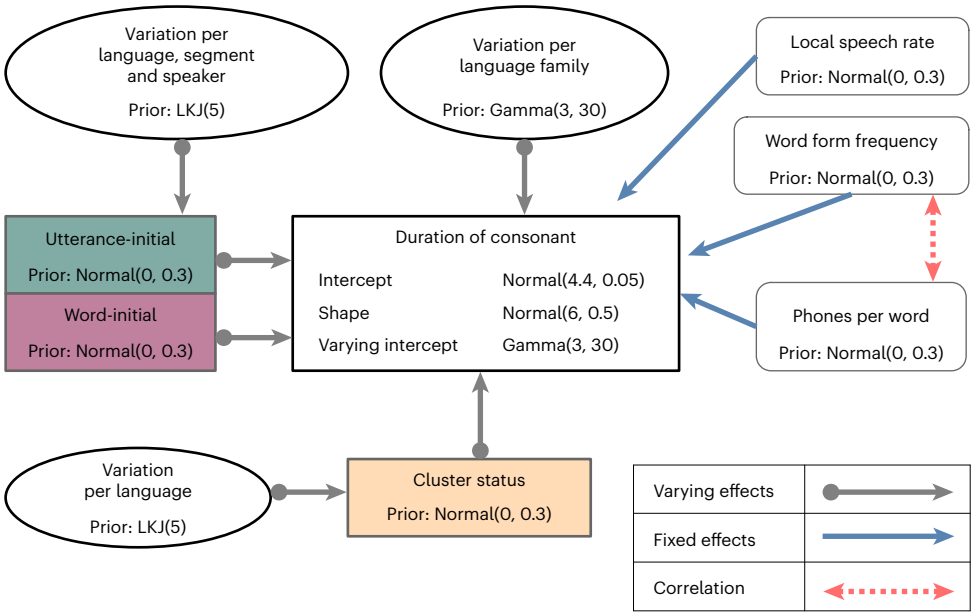

Fig. 2 | Model architecture. Fixed and varying effects of all parameters in the model including their prior distributions. The prior for the varying slopes is given as Lewandowski-Kurowicka-Joe (LKJ) distribution. The colored boxes indicate the various slopes that were added to the model, varying per language.

analysis are phones (discrete segments of speech); words, as defined by experts on each language; and utterances, which we define as inter-pausal units—that is, chunks of speech that are not interrupted by a silent pause. The entire corpus consists of over two million phones, all of which have been time-aligned semi-automatically[30]. Of these, we used 874,627 phones for this study (see Methods for information on data filtering). For 49 of 51 languages, our analysis included more than 10,000 data points.

We used Bayesian linear regression to estimate the effect of word-initial and utterance-initial positions on the duration of consonants, compared with word-internal positions. We modelled the effect of both positions with a population-level estimate that is allowed to vary between all languages in the sample. For a more conservative analysis, we allowed for variation of the effects between speakers of the same language. This ensures that any inference drawn from the model can be generalized over different speakers. Similarly, we allowed the model to vary between segments of different places and manners of articulation since lengthening effects influence each kind of segment differently[20]. We also controlled for consonant clusters and distinguished between three levels: the consonant is (1) at the beginning of a cluster, (2) in a cluster but not at the beginning or (3) not in a cluster. All levels are modelled as varying between each language. As fixed parameters, we controlled for word length (the number of phones in a word), word form frequency (of forms in the DoReCo corpus of each language) and local speech rate. The full model including prior distributions and likelihood function is given as Fig. 2. The likelihood function defines the response variable using a gamma distribution, which transforms the response variable (duration in milliseconds) to a log scale. Converting to a log scale is a common transformation for duration measures in linguistics to compare orders of magnitude instead of comparing absolute differences in milliseconds[31]. The posterior distributions of parameter values in Bayesian regression studies are defined via their highest posterior density interval (HPDI), which describes the area of the distribution in which most of the sampled posterior values are represented[32–34]. In Bayesian statistics, the type S error rate for the posterior intervals is much lower than in comparable frequentist methods[35]. Another measure to exclude spurious effects and to produce reliable results is to include a region of practical equivalence to 0 (ROPE)[36]. The ROPE is values near 0 (−0.01 to 0.01 on the log scale) that we consider not to be meaningful. In the complete absence of an effect, the posterior distribution would be fully within the ROPE[37]. We interpret 89% HPDIs not overlapping the ROPE as evidence in favour of an effect. If the 89% HPDI overlaps the ROPE, we take the evidence as inconclusive.

The fitted model shows evidence for the word-initial lengthening of consonants in utterance-medial position for 43 of the 51 sampled languages. No language shows evidence in favour of word-initial shortening. For the languages for which we have evidence, the 89% HPDI does not intersect with zero or the values defined in the ROPE. The mean of the HPDI for the 43 languages ranges mostly between 0.1 and 0.3 on the log scale, which translates to an average effect between 8 ms and 18 ms for a segment 84 ms long (the mean duration of phones in the data). The cross-linguistic distribution provides us with high confidence in the reliability of our results. They strongly imply that the observation of lengthening of word-initial consonants in comparison with their word-internal counterparts can be generalized across languages. We show the posterior distributions for the word-initial parameter in all languages in Fig. 3.

Regarding utterance-initial positions, no language in our sample shows evidence in favour of lengthening. However, 15 languages show evidence for utterance-initial shortening. In these languages, the duration of consonants tends to be shorter in utterance-initial than in utterance-medial or final position. For the other 36 languages, the results are inconclusive. The HPDI of this distribution displays a weak tendency towards the shortening of utterance-initial consonants for some languages, but for others, the HPDI indicates a weak tendency towards their lengthening. None of those are interpretable, and no uniform cross-linguistic pattern emerges across the sample. We present the individual posterior distributions in Fig. 4.

Posterior distribution of control variables

The distribution of parameter values across the whole dataset is presented in Fig. 5. All values are on the log scale. Since the model was parameterized as treatment coding, the 'non-initial' level is modelled as the intercept, and both 'utterance-initial' and 'word-initial' compare directly to the 'non-initial' baseline. For the average consonant of 84.35 ms in our data, a lengthening on the log scale of 0.14 (the mean of the word-initial parameter) results in a lengthening of ~13 ms.

https://doi.org/10.1038/s41562-024-01988-4

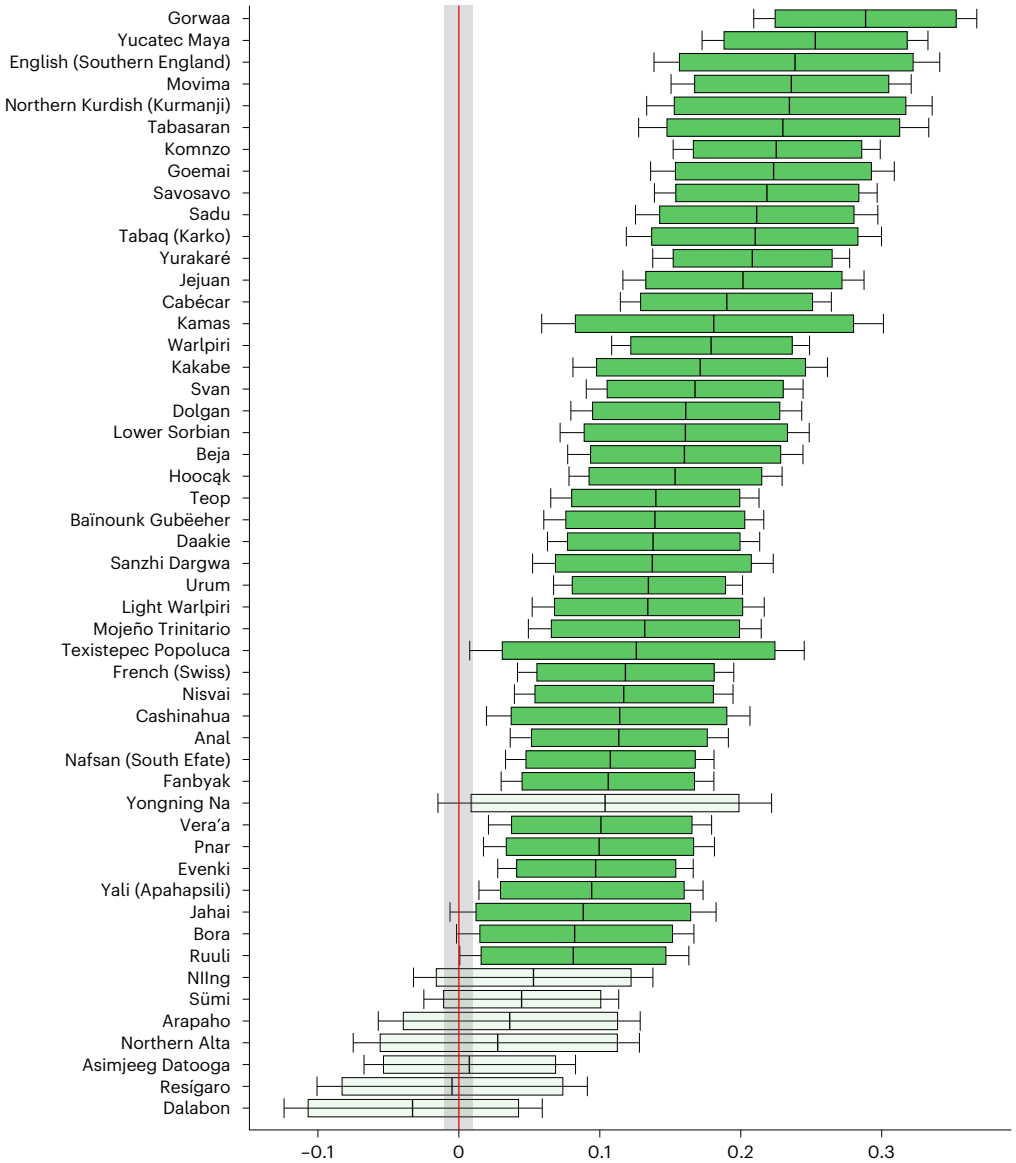

Fig. 3 | Main results for word-initial lengthening. The value on the x-axis indicates the lengthening effect of the word-initial position on the log-scale. Mean (vertical line), 89% HPDI (box) and 95% HPDI (error bars) (n = 6,000 Markov chain Monte Carlo (MCMC) samples) of the posterior distribution for word-initial lengthening across 51 languages. Faded colouring indicates that the 89% posterior interval intersects with the ROPE (grey shading).

Word-form frequency has a small negative effect on duration with a mean of −0.02 (95% HPDI from −0.02 to −0.02) on the log scale. Similarly, word length in phones, measured as phones per word, has a small negative effect on duration with a mean of −0.03 (95% HPDI from −0.03 to −0.03) on the log scale. This is exactly as predicted: segments in longer words are shortened (polysyllabic shortening), and more frequent words are uttered faster. There is a strong correlation (ρ = 0.61) between both parameters[31], in that many phones per word correlates with a lower word-form frequency. Incidentally, this confirms the cross-linguistic validity of Zipf's law of abbreviation that more frequently used words are shorter[38–41]. Given the strong correlation between both parameters, the effects in the model should not be interpreted separately but should always be considered together statistically. Local speech rate has the expected large effect on duration in the model (−0.19, 95% HPDI from −0.20 to −0.19). As duration per sound is a central part of calculating speech rate, it is not surprising that this predictor is the strongest of all three. It is important to remember that all three predictors are modelled to be uniform across the whole dataset—that is, they are modelled not to vary between individual languages. The effects for cluster-internal consonants show more variation. Consonants outside of a cluster are shorter (−0.03, 95% HPDI from −0.05 to −0.00) than consonants at the beginning of a cluster. Consonants within a cluster are even shorter (−0.07, 95% HPDI from −0.09 to −0.04). The results per language are presented in Supplementary Information section B. Figure 5 further shows that the utterance-initial and word-initial parameters have a large standard deviation at the population level. This indicates that these predictors do not behave uniformly across languages, as we have already seen for the language-specific distributions.

Posterior evaluation of the model

We ran posterior predictive simulations to confirm that on average, we expect word-initial consonants to be longer than consonants in other positions. A common way to evaluate a Bayesian linear regression model is to run posterior predictions with simulated data[33,34]. We present such posterior predictions in Fig. 6, where we can observe a higher average duration for word-initial consonants than for the other positions. On average, the word-initial consonants in the simulated dataset

https://doi.org/10.1038/s41562-024-01988-4

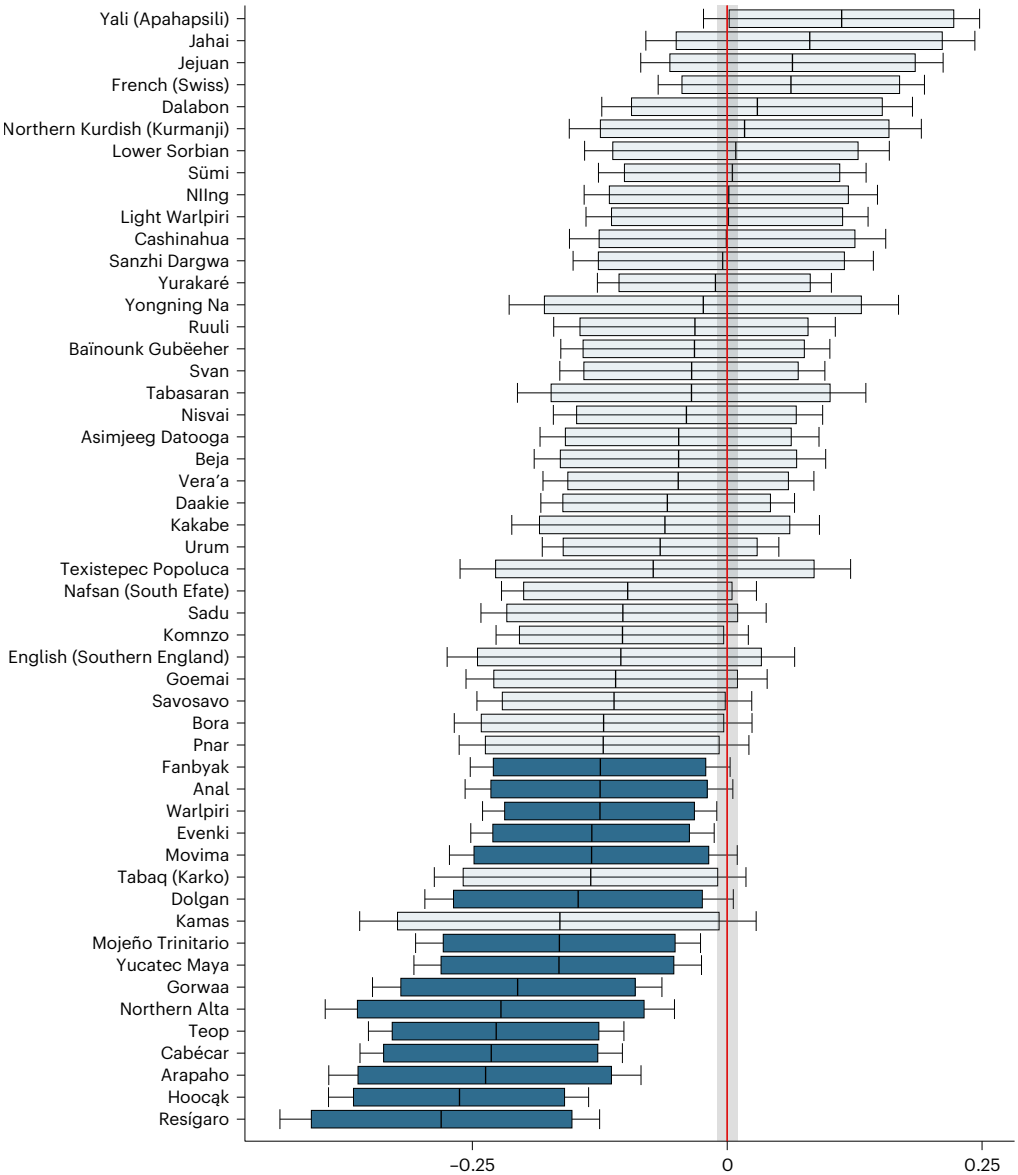

Fig. 4 | Main results for effects in utterance-initial position. The value on the x-axis indicates the shortening or lengthening effect of the utterance-initial position on the log-scale. Mean (vertical line), 89% HPDI (box) and 95% HPDI (error bars) (n = 6,000 MCMC samples) of the posterior distribution for effects in utterance-initial position across 51 languages. Faded colouring indicates that the 89% posterior interval intersects with the ROPE (grey shading).

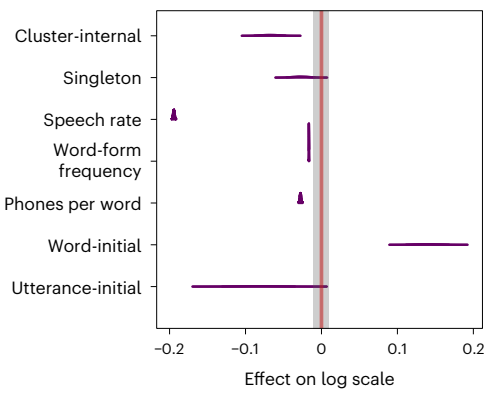

Fig. 5 | Results for all parameters at the population level. The boxes show the 89% HPDIs (n = 6,000 MCMC samples) of the posterior distribution for all parameters at the population level. The point estimates represent variation of less than 0.02 in the estimate of the posterior distribution.

are expected to be around ~13 ms longer (~106 ms) than consonants in other positions (~93 ms). Full posterior predictive checks according to the Bayesian Analysis Reporting Guidelines[42] are presented in Supplementary Information section B.

To control for possible non-independence of data points, we carefully analysed the genealogical and spatial relations in our dataset. Our sample includes data from 30 different language families. While eight language families are represented by multiple languages (for example, seven Austronesian, four Indo-European and four Sino-Tibetan languages), there are 22 language families with only one language in our sample. In the model, we added a varying intercept per language family, which shows a very small variance between language families (0.04 on the log scale). This shows that the model cannot identify systematic patterns across language families and attributes most of the durations to variation between languages, segments or speakers. Further approximations of potential correlations between language families are provided by controlling for spatial autocorrelation, since most of the languages in our sample that are related to each other genealogically

https://doi.org/10.1038/s41562-024-01988-4

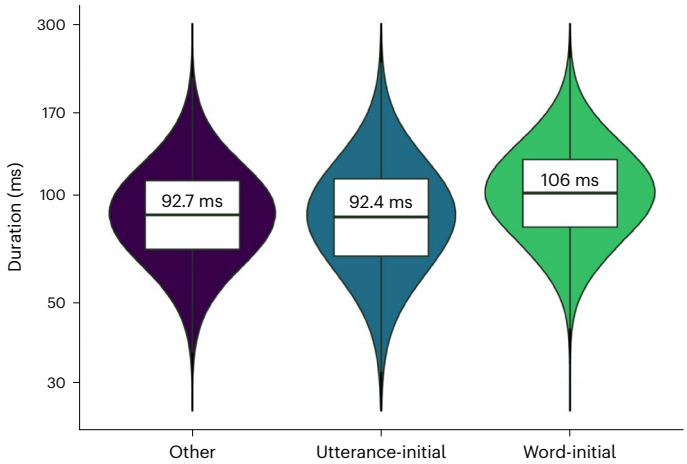

Fig. 6 | Expected duration across model predictions. Posterior predictions for expected draws (n = 6,000 draws from the posterior distribution) given the fitted model and simulated data. The horizontal bar represents the mean with the value printed above it, the box represents the 25th and 75th percentiles, and the violin represents the whole estimated distribution for each parameter. Note that the y axis is log scaled.

(especially Austronesian, Indo-European and Sino-Tibetan languages) are also geographically close to each other.

We also verified that the model is not biased through spatial autocorrelation. This type of bias is frequent in linguistic typology and can arise through the borrowing of structural features between languages[43,44]. The amount of spatial autocorrelation in data used for regression models can be measured through the Moran coefficient[45–49]. We based the computation of the Moran coefficient on the geodesic distance between the language coordinates as provided by Glottolog[50], following suggestions in the literature[51]. We computed this coefficient using the geostan package[46]. In all cases, the coefficient was close to 0, indicating very little or no spatial bias in our data. The full report for each macro area is presented in Supplementary Information section B.

Discussion

The current study reports acoustic evidence that speakers from vastly different cultural, geographic and linguistic backgrounds produce longer word-initial consonants. While languages differed in the magnitude of lengthening, evidence could be observed across a large part of the sample: 43 languages provided evidence in favour of word-initial lengthening, and none provided evidence for word-initial shortening. The effect in those languages was observed while controlling for the known between-speaker variability in prosodic boundary marking[23] and the intrinsic differences of lengthening effects of different segments. Since the current study is based on a comprehensive dataset consisting of languages from predominantly non-WEIRD communities from all parts of the world, the distribution of the effect indicates a universal tendency in spoken languages.

Our findings are consistent with models that argue for the dual importance of word-initial lengthening for segmenting speech. First, word-initial lengthening might directly indicate word boundaries. Second, lengthening would facilitate word recognition through the prominent pronunciation of word-initial segments, which are the most informative ones for word identification[10,21]. One potential reason why speakers' word-initial lengthening is so widespread is that it can promote these two processing requirements for the listener simultaneously[11]. There may be additional articulatory reasons for slowing down in the vicinity of boundaries, but how exactly language comprehension and production interact in this respect remains unclear[28,52]. While the influence of initial lengthening on speech processing has been shown in experimental studies for speakers of some languages[21], the cross-linguistic evidence for the role of initial lengthening in speech processing would ultimately have to be confirmed in perception studies. Word-initial lengthening could then emerge as an additional key factor for the segmentation of speech in the multi-faceted process of speech recognition[53].

Regarding speech production, our results partially support and partially contradict predictions made by current models of articulatory phonology, such as the π-gesture model. This model predicts that articulatory gestures are slowed down at prosodic boundaries, manifested in acoustic data as lengthening effects[52,54]. Our findings are, in general, consistent with this view. For larger prosodic boundaries in contrast to smaller ones, the π-gesture model would predict longer durations. If we assume that a word boundary after a pause corresponds to a major prosodic boundary compared with a word boundary with no preceding pause, longer durations should be found in the former than in the latter. However, we did not find a lengthening effect for consonants utterance-initially compared with word-initial positions. For 15 of 51 languages, we even found evidence for shortening of utterance-initial consonants. These findings go against the π-gesture model predictions. The findings do, however, mirror reports on the disappearing effect of final lengthening at strong prosodic boundaries with long pauses[18]. This suggests that speakers systematically modulate the segmental duration of initial consonants at the word level but do not always mark boundaries of higher prosodic levels at the beginning of an utterance. The absence of additional lengthening in utterance-initial position suggests that consonant lengthening is more closely linked to the segmentation and identification of word units than to prosodically structuring speech into larger units such as prosodic phrases. Since utterances are operationalized as chunks of speech surrounded by silent pauses in our study, we interpret the lack of an effect as being related to the lack of functional ambiguity: the first segment following a pause will necessarily also be the first segment of a word, without the need for further segmentation.

Our findings align with several strands of linguistic research about the phonological role of initial segments. At the level of the syllable, onsets have long been recognized as privileged positions. They show several characteristics that other positions do not show, such as resistance to phonological change[15,16,55]. From a diachronic perspective, word-initial consonants tend to be more resistant to phonemic change than consonants in other positions. For example, initial consonant retention is far more typical than initial consonant loss, with some notable exceptions found in Indo-European and across Australian languages[56–58]. Initial consonant deletion as a productive synchronic process is even less common (but see ref. 59 for a counterexample). Regarding explanations for such asymmetries, our results lend support to models of evolutionary phonology that view initial strengthening as a cause for the historical development and preservation of 'strong' and distinctive word-initial sounds in the phonology and lexicon[60]. There have also been attempts to relate the role of phonological properties to the functional load of syllable onsets compared with syllable codas, and the word-initial position compared with the word-final position[61–63]. One such study investigated the lexical inventories of 12 mostly Indo-European languages and found that syllable onsets have a considerably higher functional load, giving them an extraordinary status[61]. Conversely, word-final positions have been shown to have a reduced degree of structural complexity[63]. These long-term evolutionary processes are consistent with the special role of word-initial segments during the online incremental processing of words.

While the data showed a clear cross-linguistic trend for lengthening at the beginning of words, 8 of 51 languages showed a certain degree of resistance to durational modulations at word-initial position, as evidenced by the intersection of the 89% HPDI with the ROPE (Fig. 3). While this apparent resistance could be explained by insufficient or noisy data, it is also possible that these languages lack word-initial lengthening. Language-specific factors that could affect the degree of lengthening and deserve further attention in future research include

the phoneme inventory of the language, the distribution of segments with variable pronunciations (in particular glottal stops), phonological length distinctions (singletons versus geminates) and lexical stress.

Some inevitable limitations might influence the interpretation and generalizability of our findings. First, one limitation of this study lies in the corpus-based approach using aggregated language documentation data and recordings of natural speech. While these data sources provide an ecologically valid and rich set of linguistic samples, they are susceptible to noise and variability inherent in natural speech recordings. They were created over several decades, using different recording equipment and protocols, leading to potential inconsistencies in audio quality. Despite efforts in preselecting high-quality audio for the corpus[30], the inherent variation in recording conditions remains a concern. However, the corpus-based approach offers the advantage of observing effects in spontaneously produced speech, outside of a strict experimental setting with a less varied sample of texts and speakers.

Second, the sample size, although comprising 51 diverse languages from 30 different language families, still poses a limitation. For some of these languages, we have data from only one (Kamas, Texistepec Popoluca and Yongning Na) or two speakers (Tabasaran, Northern Alta, Kurmanji and Southern British English), while for many other languages, we have data from more than ten speakers. In an ideal scenario, a larger sample size would enhance the study's generalizability across an even broader spectrum of languages and language families, as well as speakers[64,65]. However, while other multilingual speech corpora are available[66–68], none of these corpora, in our view, achieve the necessary balance between corpus size, detailed annotation of relevant features and metadata, and expert-informed processing allowing for reliable alignments across a multitude of low-resourced languages that are offered by DoReCo.

A third limitation of the present study lies in its simplistic view of consonant duration. Consonant duration is a multifaceted phenomenon encompassing various acoustic components such as burst, frication, voice onset time and formant transition periods. This also resonates with previous calls for acknowledging the importance of fine phonetic detail for social aspects of communication[69]. Complementing the study with a detailed articulatory perspective that includes annotation of articulatory gestures in the production of consonants could add more depth to our understanding of the underlying principles of word-initial consonant lengthening for specific languages. However, recording and annotating this kind of complex articulatory data is outside the scope of this study. Another limitation related to the previous one is the lack of accounting for word-level prominence in our analysis. Our corpus data are not annotated for suprasegmental features such as stress or tone. However, on the basis of available phonological descriptions, only 4 of the 51 languages can with some certainty be considered to have fixed initial word stress, while most other languages are either tone languages or stress languages with non-initial stress (Supplementary Information section B). In our model, those four languages do not seem to show any patterns for initial-lengthening effects that distinguish them from the other languages. It therefore seems unlikely that our overall results are skewed by not taking word-initial prominence into account.

Despite these limitations, the evidence across a worldwide sample of languages suggests that the lengthening of word-initial consonants is a potentially fundamental process structuring human speech. This strong effect emerges while carefully controlling for between-speaker variability and variability across segments, which adds additional credence to this conclusion. Given the diverse sample of languages in our study, we predict that this effect is replicable for other languages and datasets.

Methods

Language sample

Our study uses data from the DoReCo corpus (v.1.2)[29]. The corpus contains time-aligned transcriptions and annotations that mostly originated from language documentation collections covering a wide range of typologically diverse languages. In total, DoReCo v.1.2 contains corpora from 51 languages from 30 language families. All corpora are comparable in size and include at least 10,000 phones (before filtering). A detailed account of the individual corpora and their sources are presented in Extended Data Table 1. Word units in our data were defined and annotated by the language experts who contributed data to DoReCo (Extended Data Table 1), on the basis of current standards in descriptive linguistics. Within DoReCo, the heterogeneous documentation data were processed using a combination of automatic and manual techniques. Forced time alignments were created using the WebMAUS service[70] first for start and end times of words, which were then corrected manually for the whole corpus[30]. Following this, the updated alignments were used as input to create automatic alignments at the segment level.

We have converted the corpus data to the Cross-Linguistic Data Format (CLDF)[71,72] to facilitate the reuse of the data and replication of our results. A detailed description of using the corpus as a CLDF dataset is provided as Supplementary Information section A. All preprocessing steps were handled using an SQLite query that is based on the CLDF dataset. Before fitting the models, we cleaned the data by excluding certain observations. Since we are interested only in the lengthening of initial consonants, we removed all vowels from the data. We also removed geminates (that is, phonologically long consonants) due to their intrinsic lengthening. Utterance-initial stops have been excluded because their initial closure period following a pause is unmeasurable[73]. We excluded sounds with a duration equal to or below 30 ms, which was set as the minimum duration by the MAUS aligner, with shorter durations being indicative of imprecise last-resort alignments[30]. Lastly, we excluded outliers beyond three standard deviations of the mean for each speaker. For most speakers, this resulted in an upper threshold of around 300 ms, which is a very conservative threshold concerning the expected duration of individual segments. Random samples of excluded segments showed that these cases are mostly transcription or alignment errors and have been correctly excluded.

Causal effects on segment duration

In our model, we controlled for several known causal effects on the duration of phones. We controlled for inter- and intra-speaker variation in speech rate through the proxy variable 'local speech rate', which is equal to the average duration of phones per utterance. We also controlled for the number of phones per word and the word-form frequency as fixed effects. The word-form frequency is computed as the frequency of each form within the DoReCo corpus core set of each language. Both parameters are predicted to be highly correlated. For frequency of occurrence, more frequent words are known to be shorter (Zipf's law of abbreviation)[74–77]. Longer words have been shown to have shorter components, most crucially shorter affixes and shorter phones in specific conditions such as under phrasal accent (Menzerath's law or polysyllabic shortening)[26,39,41,78,79]. In our model, these three variables were log-scaled and standardized for each language.

The effect for word- and utterance-initial position was modelled with varying intercepts and slopes across all languages. This ensured that we could assess the effects in all languages, instead of interpreting the effect on the population level as being true for all languages[80,81]. We also included 'speaker' as a varying effect in our model, as there are huge amounts of variation between speakers in all linguistic domains[82–84]. It is necessary to control for this kind of variation to make valid generalizations about language[64,65]. Finally, we controlled for variation of the effect across different segments since there might be variation in the elasticity of segments depending on their place and manner of articulation. In total, the corpus includes 191 different segment types, which are mapped from their X-Sampa representation in DoReCo to the Cross-Linguistic Transcription Systems standard[85,86].

Model fitting and evaluation

The reason for choosing a Bayesian approach is the wide range of tools to include prior knowledge of the world in the model and to develop

a transparent and reliable model output that is explicit about any uncertainty involved in the inference[87,88]. The goal of our analysis is to determine the effect size of the word-initial position of phones in speech. Given that we know quite a lot about speech sounds in general, such as expected duration and known causal influences, we can add this prior knowledge directly into the model. Bayesian regression offers several well-designed measures for enabling transparency of the workflow[89,90]. We report on all relevant points of the Bayesian Analysis Reporting Guidelines[42] either in the main text or in the Supplementary Information. We did not include a large-scale sensitivity analysis for our prior distributions, due to the large and energy-intensive computing times. We hope that the prior predictive checks provide sufficient information for the credibility of our prior distributions. We further excluded the points that relate to hypothesis testing with Bayes factors since no model comparison was done in our study. Instead of doing a model comparison or null-hypothesis significance test, we analysed the effect size of our target parameter while controlling for known causal factors.

The model was fit using brms[91,92], a package in R[93] that uses cmdstanR as a backend. The model was run with 4,000 MCMC iterations (2,500 for warm-up) on four parallel chains. A computational and visual confirmation of model convergence as well as prior and posterior predictive checks are presented in Supplementary Information section B.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

For this study, we used data from the DoReCo corpus (v.1.2) and converted them to a CLDF dataset (v.1.2.1)[29,94]. While the data are available as Open Access, some files come with a non-derivative restriction. We have therefore added instructions for an automated workflow of downloading the data and converting it to an SQlite database via CLDF instead of providing the data directly[71,72], thereby adhering to the non-derivative restrictions. To reproduce the exact steps, please follow the instructions provided in our GitHub repository (https://github.com/FredericBlum/initial_lengthening/blob/v1.0/README.md).

Code availability

The current version of the code (v.1.0) is available via Zenodo at https://doi.org/10.5281/zenodo.13141902 (ref. 95) and curated on GitHub (https://github.com/FredericBlum/initial_lengthening/tree/v1.0). We provide full instructions to reproduce our results in a README.md in the shared repository. The models have been uploaded to an OSF directory (https://doi.org/10.17605/OSF.IO/TC9ZX) since we could not upload them to GitHub due to their large file size.

References

1. Cutler, A. in *Lexical Representation and Process* (ed. Marslen-Wilson, W.) 342–356 (MIT Press, 1989).
2. Brent, M. R. Speech segmentation and word discovery: a computational perspective. *Trends Cogn. Sci.* **3**, 294–301 (1999).
3. Mattys, S. L., White, L. & Melhorn, J. F. Integration of multiple speech segmentation cues: a hierarchical framework. *J. Exp. Psychol. Gen.* **134**, 477–500 (2005).
4. Gong, X. L. et al. Phonemic segmentation of narrative speech in human cerebral cortex. *Nat. Commun.* **14**, 4309 (2023).
5. Henrich, J., Heine, S. J. & Norenzayan, A. Most people are not WEIRD. *Nature* **466**, 29–29 (2010).
6. Blasi, D. E., Henrich, J., Adamou, E. & Kemmerer, D. Over-reliance on English hinders cognitive science. *Trends Cogn. Sci.* **26**, 1153–1170 (2022).
7. Ladefoged, P. & Maddieson, I. *The Sounds of the World's Languages* (Blackwell, 1996).
8. Evans, N. & Levinson, S. C. The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.* **32**, 429–448 (2009).
9. Skirgård, H. et al. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Sci. Adv.* **9**, 6175 (2023).
10. Wedel, A., Ussishkin, A. & King, A. Incremental word processing influences the evolution of phonotactic patterns. *Folia Linguist.* **53**, 231–248 (2019).
11. Norris, D., Mcqueen, J. M., Cutler, A. & Butterfield, S. The possible-word constraint in the segmentation of continuous speech. *Cogn. Psychol.* **34**, 191–243 (1997).
12. Kingston, J. Lenition. In *Selected Proc. 3rd Conference on Laboratory Approaches to Spanish Phonology* (eds Colantoni, L. & Steele, J.) 1–31 (Cascadilla Proceedings Project, 2008).
13. Lavoie, L. M. *Consonant Strength: Phonological Patterns and Phonetic Manifestations* (Routledge, 2015); https://doi.org/10.4324/9780203826423
14. Katz, J. Lenition, perception and neutralisation. *Phonology* **33**, 43–85 (2016).
15. Topintzi, N. *Onsets: Suprasegmental and Prosodic Behaviour* Cambridge Studies in Linguistics Vol. 125 (Cambridge Univ. Press, 2010); https://doi.org/10.1017/CBO9780511750700
16. Easterday, S. *Highly Complex Syllable Structure: A Typological and Diachronic Study* (Language Science Press, 2019); https://doi.org/10.5281/zenodo.3268721
17. Paschen, L., Fuchs, S. & Seifart, F. Final lengthening and vowel length in 25 languages. *J. Phon.* **94**, 101179 (2022).
18. Kentner, G., Franz, I., Knoop, C. A. & Menninghaus, W. The final lengthening of pre-boundary syllables turns into final shortening as boundary strength levels increase. *J. Phon.* **97**, 101225 (2023).
19. Fletcher, J. in *The Handbook of Phonetic Sciences* 2nd edn (eds Hardcastle, W. J. et al.) 521–602 (Blackwell, 2010); https://doi.org/10.1002/9781444317251.ch15
20. Klatt, D. H. Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *J. Acoust. Soc. Am.* **59**, 1208–1221 (1976).
21. White, L., Benavides-Varela, S. & Mády, K. Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *J. Phon.* **81**, 100982 (2020).
22. Quené, H. Durational cues for word segmentation Dutch. *J. Phon.* **20**, 331–350 (1992).
23. Fougeron, C. & Keating, P. A. Articulatory strengthening at edges of prosodic domains. *J. Acoust. Soc. Am.* **101**, 3728–3740 (1997).
24. Cho, T. Prosodic boundary strengthening in the phonetics–prosody interface. *Lang. Linguist. Compass* **10**, 120–141 (2016).
25. Cho, T. & McQueen, J. M. Prosodic influences on consonant production in Dutch: effects of prosodic boundaries, phrasal accent and lexical stress. *J. Phon.* **33**, 121–157 (2005).
26. White, L. Communicative function and prosodic form in speech timing. *Speech Commun.* **63-64**, 38–54 (2014).
27. Souza, R. in *Prosodic Boundary Phenomena* (eds Schubö, F. et al.) 35–86 (Language Science Press, 2023); https://doi.org/10.5281/zenodo.7777469
28. White, L. *English Speech Timing: A Domain and Locus Approach*. PhD thesis, Univ. Edinburgh (2002); https://era.ed.ac.uk/handle/1842/23256
29. Seifart, F., Paschen, L. & Stave, M. *Language Documentation Reference Corpus (DoReCo)* (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/NKL.7CBFQ779

30. Paschen, L. et al. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proc. 12th Language Resources and Evaluation Conference* (eds Calzolari, N. et al.) 2657–2666 (European Language Resources Association, 2020); https://aclanthology.org/2020.lrec-1.324
31. Winter, B. *Statistics for Linguists: An Introduction Using R* (Routledge, 2019); https://doi.org/10.4324/9781315165547
32. Vasishth, S. & Nicenboim, B. Statistical methods for linguistic research: foundational ideas—part I. *Lang. Linguist. Compass* **10**, 349–369 (2016).
33. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Chapman and Hall/CRC, 2020); https://doi.org/10.1201/9780429029608
34. Gelman, A. et al. *Bayesian Data Analysis* (Chapman and Hall/CRC, 2013); https://doi.org/10.1201/b16018
35. Gelman, A. & Tuerlinckx, F. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Comput. Stat.* **15**, 373–390 (2000).
36. Kruschke, J. K. Rejecting or accepting parameter values in Bayesian estimation. *Adv. Methods Pract. Psychol. Sci.* **1**, 270–280 (2018).
37. Makowski, D., Ben-Shachar, M. S., Chen, S. H. A. & Lüdecke, D. Indices of effect existence and significance in the Bayesian framework. *Front. Psychol.* **10**, 2767 (2019).
38. Bentz, C. & Ferrer-i-Cancho, R. Zipf's law of abbreviation as a language universal. In *Proc. Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics* (eds Bentz, C., Jäger, G. & Yanovich, I.) 1–4 (Univ. Tübingen, 2016); https://doi.org/10.15496/publikation-10057
39. Kanwal, J., Smith, K., Culbertson, J. & Kirby, S. Zipf's law of abbreviation and the principle of least effort: language users optimise a miniature lexicon for efficient communication. *Cognition* **165**, 45–52 (2017).
40. Strunk, J. et al. Determinants of phonetic word duration in ten language documentation corpora: word frequency, complexity, position, and part of speech. *Lang. Doc. Conserv.* **14**, 423–461 (2020).
41. Stave, M., Paschen, L., Pellegrino, F. & Seifart, F. Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzerath's laws. *Linguist. Vanguard* **7**, 20190076 (2021).
42. Kruschke, J. K. Bayesian analysis reporting guidelines. *Nat. Hum. Behav.* **5**, 1282–1291 (2021).
43. Guzmán Naranjo, M. & Becker, L. Statistical bias control in typology. *Linguist. Typol.* **26**, 605–670 (2021).
44. Guzmán Naranjo, M. & Mertner, M. Estimating areal effects in typology: a case study of African phoneme inventories. *Linguist. Typol.* **27**, 455–480 (2022).
45. Chun, Y. & Griffith, D. A. *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology* (Sage, 2013).
46. Donegan, C. geostan: an R package for Bayesian spatial analysis. *J. Open Source Softw.* **7**, 4716 (2022).
47. Tiefelsdorf, M. & Boots, B. The exact distribution of Moran's *I*. *Environ. Plan. A* **27**, 985–999 (1995).
48. Griffith, D. A. A linear regression solution to the spatial autocorrelation problem. *J. Geogr. Syst.* **2**, 141–156 (2000).
49. Griffith, D. A. & Chun, Y. Some useful details about the Moran coefficient, the Geary ratio, and the join count indices of spatial autocorrelation. *J. Spat. Econom.* **3**, 12 (2022).
50. Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. *Glottolog* v.5.0 (Max Planck Institute for Evolutionary Anthropology, 2024); https://doi.org/10.5281/zenodo.10804357
51. Guzmán Naranjo, M. & Jäger, G. Euclide, the crow, the wolf and the pedestrian: distance metrics for linguistic typology. *Open Res. Eur.* **3**, 104 (2023).
52. Byrd, D. & Krivokapić, J. Cracking prosody in articulatory phonology. *Annu. Rev. Linguist.* **7**, 31–53 (2021).
53. Norris, D. & McQueen, J. M. Shortlist B: a Bayesian model of continuous speech recognition. *Psychol. Rev.* **115**, 357–395 (2008).
54. Byrd, D. & Saltzman, E. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *J. Phon.* **31**, 149–180 (2003).
55. Zec, D. in *The Cambridge Handbook of Phonology* (ed. Lacy, P.) 161–194 (Cambridge Univ. Press, 2007); https://doi.org/10.1017/CBO9780511486371.009
56. Blevins, J. in *Forty Years On: Ken Hale and Australian Languages* (eds Simpson, J. et al.) 481–492 (Pacific Linguistics, 2001); https://doi.org/10.15144/PL-512.481
57. Green, A. D. in *The Syllable in Optimality Theory* (eds Féry, C. & van de Vijver, R.) 238–253 (Cambridge Univ. Press, 2003); https://doi.org/10.1017/CBO9780511497926.010
58. Miceli, L. & Round, E. Where have all the sound changes gone? Examining the scarcity of evidence for regular sound change in Australian languages. *Linguist. Vanguard* **8**, 509–518 (2022).
59. Marley, A. H. Sound change in Aboriginal Australia: word-initial engma deletion in Kunwok. *Linguist. Vanguard* **8**, 645–659 (2022).
60. Blevins, J. in *The Oxford Handbook of Historical Phonology* (eds Honeybone, P. & Salmons, J.) 485–500 (Oxford Univ. Press, 2015); https://doi.org/10.1093/oxfordhb/9780199232819.013.006
61. Sun, Y. & Poeppel, D. Syllables and their beginnings have a special role in the mental lexicon. *Proc. Natl Acad. Sci. USA* **120**, 2215710120 (2023).
62. Wedel, A., Kaplan, A. & Jackson, S. High functional load inhibits phonological contrast loss: a corpus study. *Cognition* **128**, 179–186 (2013).
63. Wedel, A., Ussishkin, A. & King, A. Crosslinguistic evidence for a strong statistical universal: phonological neutralization targets word-ends over beginnings. *Language* **95**, 428–446 (2019).
64. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2020).
65. Winter, B. & Grice, M. Independence and generalizability in linguistics. *Linguistics* **59**, 1251–1277 (2021).
66. Salesky, E. et al. A corpus for large-scale phonetic typology. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* (eds Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J.) 4526–4546 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.acl-main.415
67. *Lingua Libri* (Wikimédia France, 2020–2023); https://lingualibre.org/wiki/LinguaLibre:Main_Page
68. Ardila, R. et al. Common voice: a massively-multilingual speech corpus. In *Proc. 12th Language Resources and Evaluation Conference* (eds Calzolari, N. et al.) 4218–4222 (European Language Resources Association, 2020); https://aclanthology.org/2020.lrec-1.520
69. Hawkins, S. Roles and representations of systematic fine phonetic detail in speech understanding. *J. Phon.* **31**, 375–405 (2003).
70. Kisler, T., Schiel, F. & Sloetjes, H. Signal processing via web services: the use case WebMAUS. In *Proc. Digital Humanities* (ed. Meister, J. C.) 30–34 (Hamburg University Press, 2012).
71. Forkel, R. et al. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Sci. Data* **5**, 180205 (2018).
72. Forkel, R. & List, J.-M. CLDFBench: give your cross-linguistic data a lift. In *Proc. 12th Language Resources and Evaluation Conference* (eds Calzolari, N. et al.) 6995–7002 (European Language Resources Association, 2020); https://aclanthology.org/2020.lrec-1.864

73. Turk, A., Nakai, S. & Sugahara, M. in *Methods in Empirical Prosody Research* (eds Sudhoff, S. et al.) 1–28 (De Gruyter, 2006); https://doi.org/10.1515/9783110914641.1
74. Zipf, G. K. *The Psycho-biology of Language: An Introduction to Dynamic Philology* (George Routledge & Sons, Houghton, Mifflin, 1935).
75. Zipf, G. K. *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949).
76. Sigurd, B., Eeg-Olofsson, M. & Weijer, J. Word length, sentence length and frequency—Zipf revisited. *Stud. Linguist.* **58**, 37–52 (2004).
77. Jurafsky, D., Bell, A., Gregory, M. & Raymond, W. D. in *Frequency and the Emergence of Linguistic Structure* (eds Bybee, J. & Hopper, P.) 229 (John Benjamins, 2001); https://doi.org/10.1075/tsl.45.13jur
78. Gahl, S., Yao, Y. & Johnson, K. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *J. Mem. Lang.* **66**, 789–806 (2012).
79. Piantadosi, S. T., Tily, H. & Gibson, E. Word lengths are optimized for efficient communication. *Proc. Natl Acad. Sci. USA* **108**, 3526–3529 (2011).
80. Evans, N. & Levinson, S. C. The myth of language universals. *Behav. Brain Sci.* **32**, 429–448 (2009).
81. Bickel, B. Statistical modeling of language universals. *Linguist. Typol.* **15**, 401–413 (2011).
82. Baayen, H., Davidson, D. J. & Bates, D. M. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **59**, 390–412 (2008).
83. Yu, A. C. L. & Zellou, G. Individual differences in language processing. *Annu. Rev. Linguist.* **5**, 131–150 (2019).
84. Barth, D. et al. in *Doing Corpus-Based Typology with Spoken Language Data: State of the Art* (eds Haig, G. et al.) 179–232 (Univ. Hawai'i Press, 2021); http://hdl.handle.net/10125/74661
85. Anderson, C. et al. A cross-linguistic database of phonetic transcription systems. *Yearb. Poznan Linguist. Meet.* **4**, 21–53 (2018).
86. List, J.-M., Anderson, C., Tresoldi, T., Rzymski, C. & Forkel, R. CLTS: Cross-Linguistic Transcription Systems. *Zenodo* https://doi.org/10.5281/zenodo.10997741 (2024).
87. Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F. & Kong, E. J. Bayesian data analysis in the phonetic sciences. *J. Phon.* **71**, 147–161 (2018).
88. Vasishth, S. & Gelman, A. How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics* **59**, 1311–1342 (2021).
89. Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. & Gelman, A. Visualization in Bayesian workflow. *J. R. Stat. Soc. A* **182**, 389–402 (2019).
90. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2016).
91. Bürkner, P.-C. brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80**, 1–28 (2017).
92. Bürkner, P.-C. Advanced Bayesian multilevel modeling with the R package brms. *R J.* **10**, 395–411 (2018).
93. R Core Team. *R: A Language and Environment for Statistical Computing* https://www.R-project.org/ (R Foundation for Statistical Computing, 2018).
94. Seifart, F., Paschen, L., Stave, M., Forkel, R. & Blum, F. CLDF dataset derived from the DoReCo core corpus v1.2.1. *Zenodo* https://doi.org/10.5281/zenodo.10990565 (2024).
95. Blum, F., Paschen, L., Forkel, R., Fuchs, S. & Seifart, F. Code accompanying the submission for 'Consonant lengthening marks the beginning of words across a diverse sample of languages'. *Zenodo* https://doi.org/10.5281/zenodo.11198843 (2024).
96. Rose, F. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.cbc3b4xr
97. Ozerov, P. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.0dbazp8m
98. Cowell, A. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.36f5r1b6
99. Griscom, R. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.f77c7m72
100. Cobbinah, A. Y. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.a332abw8
101. Vanhove, M. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.edd011t1
102. Seifart, F. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.6eaf5laq
103. Quesada, J. D., Skopeteas, S., Pasamonik, C., Brokmann, C. & Fischer, F. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.ebc4ra22
104. Reiter, S. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.a8f9q2f1
105. Krifka, M. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.efeav5l9
106. Ponsonnet, M. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.fae299ug
107. Däbritz, C. L., Kudryakova, N., Stapert, E. & Arkhipov, A. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.f09eikq3
108. Schiborr, N. N. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.9c271u5g
109. Kazakevich, O. & Klyachko, E. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.5e0d27cu

110. Franjieh, M. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.02084446
111. Avanzi, M., Béguelin, M.-J., Corminboeuf, G., Diémoz, F. & Johnsen, L. A. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.3520l685
112. Hellwig, B. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.b93664ml
113. Harvey, A. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.a4b4ijj2
114. Hartmann, I. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.b57f5065
115. Burenhult, N. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.6a71xp0p
116. Kim, S.-U. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.06ebrk38
117. Vydrina, A. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.d5aeu9t6
118. Gusev, V., Klooster, T., Wagner-Nagy, B. & Arkhipov, A. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.cdd8177b
119. Döhler, C. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.c5e6dudv
120. O'Shannessy, C. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.7452803q
121. Bartels, H. & Szczepański, M. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.6c6e4e9k
122. Haude, K. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.da42xf67
123. Thieberger, N. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.ba4f760l
124. Aznar, J. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.2801565f
125. Garcia-Laguia, A. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.efea0b36
126. Haig, G., Vollmer, M. & Thiele, H. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.ca10ez5t
127. Güldemann, T., Ernszt, M., Siegmund, S. & Witzlack-Makarevich, A. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.f6c37fi0
128. Ring, H. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.5ba1062k
129. Seifart, F. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.ffb96lo8
130. Witzlack-Makarevich, A., Namyalo, S., Kiriggwajjo, A. & Molochieva, Z. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.fde4pp1u
131. Xu, X. & Bai, B. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.3db4u59d
132. Forker, D. & Schiborr, N. N. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.81934177
133. Wegener, C. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.b74d1b33
134. Gippert, J. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.9ba054c3
135. Teo, A. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.5ad4t01p

136. Hellwig, B., Schneider-Blum, G. & Ismail, K. B. K. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.eea8144j
137. Bogomolova, N., Ganenkov, D. & Schiborr, N. N. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.ad7f97xr
138. Mosel, U. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.9322sdf2
139. Wichmann, S. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.c50ck58f
140. Skopeteas, S., Moisidi, V., Tsetereli, N., Lorenz, J. & Schröter, S. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.ac166n10
141. Schnell, S. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.3e2cu8c4
142. O'Shannessy, C. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.042dv614
143. Riesberg, S. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.9d91nkq2
144. Michaud, A. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.abe65p95
145. Skopeteas, S. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.9cbb3619
146. Gipper, S. & Ballivián Torrico, J. in *Language Documentation Reference Corpus (DoReCo)* v.1.2 (eds Seifart, F. et al.) (Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon 2, 2022); https://doi.org/10.34847/nkl.7ca412wg

Acknowledgements

We thank L. Dees, J. Krivokapić, J. Mansfield, A. Wedel and S. Wichmann for their helpful comments. We thank C. Rzymski for an extensive review of our code and for providing support with the HPC cluster. We thank M. Mertner for suggestions on analysing possible spatial dependencies. All remaining errors are our responsibility. This study was partially supported by the Max Planck Society Research Grant 'Beyond CALC: Computer-Assisted Approaches to Human Prehistory, Linguistic Typology, and Human Cognition (CALC^3)' (F.B.), awarded to J.-M. List (2022–2024), and DFG grants SE 1949/3-1 and SE 1949/5-1 awarded to F.S. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

F.B. conceptualized the study under the supervision of F.S. and S.F. F.B. designed and analysed the statistical model. R.F. provided the conversion of the raw data to CLDF as well as the preprocessing of the data. F.B., S.F. and F.S. wrote the initial draft of the Introduction. F.B. wrote the initial draft of the Results. L.P. wrote the initial draft of the Discussion. F.B. and L.P. wrote the initial draft of the Methods. R.F. wrote the usage guide (Supplementary Information section A). F.B. and L.P. wrote Supplementary Information section B. All authors have read, commented on and approved the manuscript.

Funding

Open access funding provided by Max Planck Society.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-024-01988-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-01988-4.

Correspondence and requests for materials should be addressed to Frederic Blum.

Peer review information *Nature Human Behaviour* thanks Gerrit Kentner, Laurence White and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41562-024-01988-4

Extended Data Table 1 | Source Table

Name	Glottocode	Speakers	Phones	Utterances	Words	Source
Anal	anal1239	10	21114	928	7779	[97]
Arapaho	arap1274	4	14496	1237	2393	[98]
Asimjeeg Datooga	tsim1256	12	17476	244	4053	[99]
Baïnounk Gubëeher	bain1259	9	22036	581	3314	[100]
Beja	beja1238	5	32291	919	6303	[101]
Bora	bora1263	6	17910	310	3741	[102]
Cabécar	cabe1245	10	13942	757	6707	[103]
Cashinahua	cash1254	3	15336	1146	5296	[104]
Daakie	port1286	10	17176	1020	7561	[105]
Dalabon	ngal1292	4	10417	471	1892	[106]
Dolgan	dolg1241	6	18013	168	4092	[107]
English (Southern England)	sout3282	2	11630	497	4297	[108]
Evenki	even1259	23	23692	533	3358	[109]
Fanbyak	orko1234	10	15730	1019	6909	[110]
French (Swiss)	stan1290	8	15046	548	6853	[111]
Goemai	goem1240	5	12125	875	6168	[112]
Gorwaa	goro1270	8	15366	441	5145	[113]
Hoocąk	hoch1243	11	15132	1226	3791	[114]
Jahai	jeha1242	3	16000	679	4706	[115]
Jejuan	jeju1234	5	12600	464	3599	[116]
Kakabe	kaka1265	4	11879	885	6873	[117]
Kamas	kama1351	1	23054	883	5498	[118]
Komnzo	komn1238	11	22976	1479	6956	[119]
Light Warlpiri	ligh1234	6	15871	520	4869	[120]
Lower Sorbian	lowe1385	4	18314	770	7551	[121]
Mojeño Trinitario	trin1278	6	18421	643	5138	[29]
Movima	movi1243	5	23545	1089	5648	[122]
Nafsan (South Efate)	sout2856	12	18285	841	5867	[123]
Nisvai	nisv1234	8	22592	643	5701	[124]
Northern Alta	nort2875	2	17410	421	5658	[125]
Northern Kurdish (Kurmanji)	nort2641	2	13629	379	5681	[126]
N‖ng	nngg1234	6	8421	937	6151	[127]
Pnar	pnar1238	6	12104	432	4674	[128]
Resígaro	resi1247	3	17116	1071	4914	[129]
Ruuli	ruul1235	7	16822	706	3774	[130]
Sadu	sadu1234	6	10395	608	8039	[131]
Sanzhi Dargwa	sanz1248	5	10032	249	2603	[132]
Savosavo	savo1255	7	16241	1069	8703	[133]
Svan	svan1243	9	22448	626	4689	[134]
Sümi	sumi1235	23	11778	228	5328	[135]
Tabaq (Karko)	kark1256	4	14017	539	4468	[136]
Tabasaran	taba1259	2	9826	471	2808	[137]
Teop	teop1238	11	15450	640	7938	[138]
Texistepec Popoluca	texi1237	1	17162	504	3972	[139]
Urum	urum1249	30	24959	545	4903	[140]
Vera'a	vera1241	7	18414	850	7324	[141]
Warlpiri	warl1254	18	20026	1686	5058	[142]
Yali (Apahapsili)	apah1238	10	11872	471	2735	[143]
Yongning Na	yong1270	1	12867	575	4790	[144]
Yucatec Maya	yuca1254	6	19006	477	6459	[145]
Yurakaré	yura1255	16	42167	2283	9152	[146]

Extended Data table with Glottolog language identification code ('Glottocode'[50]), number of speakers, phones, words, and utterances in our dataset, and source for all languages in the corpus[97–146].

nature portfolio

Corresponding author(s): Frederic Blum

Last updated by author(s): 25.07.2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data was collected via an automated download using cldfbench (v1.14.0), presenting the dataset in the Cross-Linguistic Data Formats (CLDF). The full data collection can be reproduced with the code provided in the Zenodo repository: https://doi.org/10.5281/zenodo.10990565. The code is maintained on GitHub: https://github.com/cldf-datasets/doreco
Data analysis	All code is available on the Zenodo repository: 10.5281/zenodo.11198843. For the analysis, brms (v2.21.0) has been used with R (v.4.4.1) and cmdstanr (v0.7.1). The code is maintained on GitHub: https://github.com/FredericBlum/initial_lengthening

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

For this study, we used data from the DoReCo corpus (v1.2) and converted it to a CLDF dataset (v1.2.1). While the data is available as Open Access, some files come

with an ND (`Non-Derivative') restriction, which is why we have added instructions for an automated workflow of downloading the data and converting it to an SQlite database via cldfbench, thereby adhering to the non-derivative restrictions. To reproduce the exact steps, please follow the instructions provided in our GitHub repository (https://github.com/FredericBlum/initial_lengthening/blob/v0.2/README.md, https://doi.org/10.5281/zenodo.10990565). The code to reproduce the data preparation is stored on Zenodo.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study is a quantitative analysis of time-aligned speech segments. We use Bayesian statistics to derive the posterior distribution of the word-position variable and simulate draws from the posterior distribution.
Research sample	The sample consists of 51 typologically diverse languages from 35 different language families from all other the world. The dataset is unique in a sense that it is based on non-WEIRD Languages and provides time-alignments for individual segments.
Sampling strategy	We used the convenience sample of the whole corpus available up to date.
Data collection	The data was fetched automatically from the repository using the cldfbench software. No blinding was involved.
Timing	Raw data downloaded 2023-12-05T15:10:11.867347
Data exclusions	We have removed vowels and geminates from the data due to their inherently different mechanisms of articulations. We have also removed all segments of length <30ms due to DoReCO specific annotations mechanisms which pools all erroneous values at 30ms. Outliers were removed at 3xSD*Mean for each speakers. The results are the same with outliers included. In total, we have removed 999.552 speech segments, the grand majority of them being vowels. All data exclusion criteria are made explicit within the paper as well.
Non-participation	No participants were directly involved in the study.
Randomization	Randomization was not applicable. The analysis consists of a convencience sample of observational data. Varying effects have been included to account for confounds.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used — *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation — *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s) — *State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*

Authentication — *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination — *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*

Commonly misidentified lines (See ICLAC register) — *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance — *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*

Specimen deposition — *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods — *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight — *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals — *For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*

Wild animals — *Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*

Reporting on sex — *Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall*

	numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☒	☐	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks — *Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*

Novel plant genotypes — *Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*

Authentication — *Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication. — *For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*

Files in database submission — *Provide a list of all files available in the database submission.*

Genome browser session
(e.g. UCSC) — *Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates — *Describe the experimental replicates, specifying number, type and replicate agreement.*

Sequencing depth — *Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*

Antibodies — *Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Peak calling parameters — *Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*

Data quality — *Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*

Software — *Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation — *Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*

Instrument — *Identify the instrument used for data collection, specifying make and model number.*

Software — *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

(See Eklund et al. 2016)

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.