



Research



Cite this article: Guo X, Verstyuk S, Chen H, Zhou B, Skiena S. 2026 Statistical structure and the evolution of languages. *Proc. R. Soc. B* **293**: 20252374.

<https://doi.org/10.1098/rspb.2025.2374>

Received: 31 January 2025

Accepted: 26 January 2026

Subject Category:

Evolution

Subject Areas:

bioinformatics, cognition, evolution

Keywords:

language evolution, word embeddings, generative processes, spatial correlation, Taylor's law

Author for correspondence:

Steven Skiena

e-mail: skiena@cs.stonybrook.edu

[†]These authors contributed equally to the study.

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.8383184>.

THE ROYAL SOCIETY
PUBLISHING

Statistical structure and the evolution of languages

Xingzhi Guo^{1,†}, Sergiy Verstyuk^{2,†}, Haochen Chen¹, Baojian Zhou³ and Steven Skiena¹

¹Stony Brook University, Stony Brook, NY, USA

²Harvard University, Cambridge, MA, USA

³Fudan University, Shanghai, People's Republic of China

SV, 0009-0006-9288-3315; SS, 0000-0003-0397-7514

Human cultural development is marked by the emergence of new words and ideas, reflecting societal changes. But how does this evolution proceed? We use modern methods in natural language processing (namely, word embeddings) to measure statistical traces of cultural development, providing a testing ground to compare different models as to how this process works. We show that real embeddings of English and 21 other languages exhibit a series of previously unrecognized regularities. Specifically, these are: (i) *frequency assortativity*, where entities of high popularity cluster near other high-popularity entities; (ii) *characteristic clustering velocity profiles* due to aggregation into hierarchical structures; (iii) *persistent temporal dynamics*, where newly created entities appear disproportionately near other recent entries; and (iv) *Taylor's law*, implying that over time and across empirical semantic space the variance in new entity counts scales as a power of the mean, which helps systematize and quantify large historical fluctuations of neologisms. To explain these facts, we propose a class of generative models (specifically, directed preferential placement) that construct synthetic embeddings exhibiting similar regularities. We show that analogous regularities also occur in other datasets, suggesting that such generating models may shed light on new aspects of language and cultural evolution.

1. Introduction

Language marks an important stage in human cultural evolution [1,2], with the regular emergence of new words and ideas reflecting societal developments in every period of time. New layers of history, science and popular culture shape this discussion, preserving traces in the way language is used and constructed. But how does this process happen? Do new concepts arise in the shadow of the most popular or influential concepts, or do they emerge to fill empty regions of the concept landscape? Are new words disproportionately created in particular semantic regions, or diffused uniformly across the landscape? By what mechanisms are observed clusters formed in the concept space?

Here, we use modern methods in natural language processing to measure statistical traces of cultural evolution, providing a testing ground to compare different models as to how this complex process operates. Information as diverse as vocabularies of multiple languages and the linkages between Wikipedia pages exhibit several common but non-trivial regularities, serving as universals across datasets. Literature that is now classic discovered such regularities as power laws in the distribution of frequencies and node degrees [3–6] as well as the power law distribution of distinct words [7–9] in text documents and real networks.

We extend these findings in several directions, showing that word embeddings of different human languages exhibit the following four regularities:

© 2026 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/4.0/>, which permits unrestricted use, provided the original author and source are credited.

- (1) A kind of spatial stratification implying that neighbour popularity concentration follows rank, meaning that higher popularity entities (words and pages) tend to have high popularity counterparts as their close neighbours, while low-popularity entities live in poorer neighbourhoods. Such *frequency (or degree) assortativity* arises both in language and network data.
- (2) A systematic scaling of community stratas, whereby entities progressively concentrate around cluster centres, revealing consistent *clustering velocity* profiles.
- (3) A spatially concentrated and temporally lasting growth of new entities exhibiting *persistent local dynamics*, so that new entities arise where other entities were recently created.
- (4) Variability of the newly created entities scales with the mean in a power-law manner, revealing a systematic pattern that contradicts uniform randomness, known in natural sciences as *Taylor's law*, and seen here in the amplitudes of fluctuations of neologisms over time.

Moreover, the first two results could be additionally verified for 21 languages other than English, which helps insure against systematic biases inherent in the former (as discussed in, e.g. [10]).

Thus, we identify several properties of the fundamental structure underlying language and cultural development. These properties are important, as they not only improve our theoretical understanding but also provide grounds for comparing sets of unlabelled synthetic embeddings from abstract models to those produced from real datasets, and thus validating which explanatory models can generate similar structures. We show how such complex structures can be produced by fairly minimalistic generating models, particularly a class of directed preferential placement models.

Our methodology rests on vector representations and embedding spaces: efficient low-dimensional mappings of high-dimensional data that represent language vocabularies [11,12], networks [13,14] and other products of learning and creativity [15]. Embeddings are real-valued vectors that aim to reflect semantic similarities within the original high-dimensional space in terms of geometric distances within the approximating low-dimensional space. Additional details on this methodology can be found in electronic supplementary material, §A.

The present work postulates that an embedding methodology uncovers a latent space that is primitive to representations in the form of a graph or text. Hence, our generating models act directly on this space. This knowledge, or epistemic, space (perhaps more intuitively understood as semantic space for vocabulary lists or concept space for, say, academic citation networks) can be thought of as a geometric structure with dimensions being epistemic qualities (e.g. size, temperature) and elements (points, regions) being epistemic entities (semes, words, concepts). For the relevant background, see electronic supplementary material, §B.

Previous work has used static word embeddings to broadly study social and linguistic phenomena. These include studies of semantic change in language over time, the influence of culture, history and geography on semantic alignment of different languages [16], identifying human biases in language corpora [17,18], tracking the evolution of gender and ethnic stereotypes in US published sources [19,20], inspecting social class perceptions in books [21], statistical regularities in semantic change [22] and in the emergence of new words [23], understanding the structure of concepts in natural language [24,25], political polarization and linguistic divergence in social media with a particular focus on dynamics [26].

Importantly, linguists are extremely interested in dynamic, evolutionary developments, albeit with a different methodology from ours: e.g. relying on phylogenetic methods [27–32], various frequency-focused approaches [33–42] and computational techniques with evolutionary features [43–45]. (Of course, period-specific snapshots are also of interest, e.g. see [46].)

(a) Stochastic processes behind language evolution

Understanding ‘empirical laws’ as emergent, predictable from primitive deterministic or random generating processes within a consistent theoretical causal model is central to the modern scientific method; the argument through the lens of linguistics can be found in, e.g. [47]. To reproduce known and novel regularities, we propose generating models that range from the completely structureless to more and more complex ones. Our generating models are rooted in the idea of cumulative advantage, which can be traced back to Eggenberger & Pólya [48] and Yule [49]. Its accepted modern formulations are due to [5] and [50] for set data (e.g. in terms of urns and balls), while Barabási & Albert [3] produced a leading application for graph data (a network of nodes and links). The core of this notion involves adding measures of popularity to existing entities (e.g. frequencies to words or referrals to websites) in a way that favours entities that are already popular. This approach gained prominence and numerous extensions, with Pitman & Yor [51] providing an influential generalization that allows greater control of the generating process via additional parameters, which was subsequently integrated with basic Markov [52] or topic [53] language models, as well as used for artificial language generation [54].

Linguists have proposed their own generating or empirical models for language and its vocabulary in particular: game-theoretic [55,56], learning-focused [57,58], graph-theoretic [31,44,45], statistical [38,42] and evolutionary [43]. However, they are focusing on different aspects from us. For instance, Karjus *et al.* [38] developed a statistical model that quantifies baseline changes and trends in lexical time series with the motivation of controlling and adjusting for topical fluctuations in corpus data. On the other hand, Ramiro *et al.* [45] modelled the emergence of new word senses in English historical lexicon as semantic chaining of different senses on a graph using taxonomical distance functions, while Montero *et al.* [43] analysed language change through the lens of the Wright–Fisher model of evolution, distinguishing between competitive effects and random factors. Indeed, a very influential theoretical framework in linguistics is explicitly based on evolutionary principles [59,60].

In this paper, we rely on cumulative-advantage processes (the preferential placement model) to capture the cardinal patterns of popularity. To allow for a temporal dimension, we consider history-dependent processes (the preferential placement with recency model). To account for geometric patterns of spatial clustering, we also evaluate context-dependent processes (the directed preferential placement model). Finally, as a placebo control, we invoke a selection of quasi-uniform, randomly dispersed processes. We discuss these models more concretely at the time of their first usage.¹

(b) Datasets and methodology

This study uses four popular datasets of static word embeddings (FastText, GloVe, Polyglot and Word2vec/GoogleNews) to represent the vocabulary structure of English and multiple other languages. There is considerable academic precedent here, where many prior works have used the structure of fixed word embeddings as proxies to represent the state of a language, including studying changes in word meanings over time [18,61] and assessing the degree of gender and racial bias reflected in large language corpora [62,63].

The available static word embedding datasets we use all differ substantially in the size/composition of their training corpora, the optimization criteria and tokenization strategies employed. But our experiments on all four classes of embeddings yield generally consistent results, indicating that these design decisions are irrelevant at the scale on which we evaluate our theoretical models.

Many linguistic frameworks, such as contemporary large language models, are based on the technology of contextual word embeddings, where the vector representation of a word in a text depends on the context around it. For example, this allows them to naturally distinguish word senses in vocabulary words having multiple meanings (polysemy) and to capture the related semantic changes [64]. In contrast, static embeddings have a single representation per vocabulary word, which presumably sits somewhere in the middle of the representations of each of its distinct senses. The creation of new senses of existing words in addition to the *de novo* creation of new vocabulary is not explicitly formalized in our theoretical models here (although they can be extended to do so). Our choice was guided by theoretical considerations as well as by data limitations. For more details, see electronic supplementary material, §C.

We focus on a broad selection of linguistic datasets that cover English and up to 21 other languages (Bengali, Catalan, Danish, German, Esperanto, Spanish, Finnish, French, Hindi, Hungarian, Indonesian, Italian, Japanese, Dutch, Norwegian, Polish, Portuguese, Russian, Swedish, Thai, Chinese), and span as much as eight centuries of history (electronic supplementary material, §D, describes our data).

Throughout the paper, we standardize the embedding space to a unit hyperball, centring its mass to lie at **0**. In relevant cases, we define local neighbourhoods of different entities in terms of hypercones partitioning the d -dimensional unit hyperball into $K = 100$ equal-volumed segments (see electronic supplementary material, §E, for more details). In the interest of completeness, the formal notation covering our setup is presented in electronic supplementary material, §F.

Finally, a few clarifications about this study, its scope and limitations are worth emphasizing. We focus on word creation and usage, but not modification. We abstract away from word polysemy. The static word embeddings we use are trained only once over the whole time span.

2. Results

We now demonstrate the existence of four distinct, non-trivial structural regularities in embeddings that hold across multiple language and network datasets. We test for the presence of each such regularity in synthetic embeddings constructed from a collection of different generating models, demonstrating that one particular model—directed preferential placement (DPP), see §2a—best captures the features observed in embeddings trained on our data.

(a) Frequency/degree assortativity

We observe that entities of high popularity tend to concentrate in embedding space near other high-popularity entities (as suggested by the cartoon in figure 1, left panel). In particular, we can measure the popularity of vocabulary words by the frequency with which they appear in a representative text corpus and compare the observed frequency of each word to that of its closest neighbours in embedding space. Visually, the frequencies of embedding points are positively associated with those of their neighbours (red curve in figure 1, right panel) across a wide range of frequencies. We name this phenomenon *frequency assortativity*.

Moreover, a similar pattern of assortativity by popularity, in this case meaning degree assortativity, may be observed within a Wikipedia network (see electronic supplementary material, figure S9 in §G), suggesting a broader question of how entities in knowledge space clump by popularity.

Now, relying on the postulate that knowledge (in practice, embedding) space is the right primitive object for semantic data, the models we designed generate elements directly as points in this space rather than text phrases or graph nodes/edges. Specifically, the preferential placement (PP) model generates elements using a preferential attachment mechanism that either increases the advantage of the already popular entities or places new entities uniformly around such popular elements. A variant designed to capture history dependence, the preferential placement with recency (PPR) model, increases the radius in the case of older entities,

¹Evolutionary linguists may think about our structureless placebo control models as neutral/stochastic drift null models, and about our more complex models with accumulating popularities as alternatives with systematic biases/directional forces/selection (e.g. see [41,43,59]).

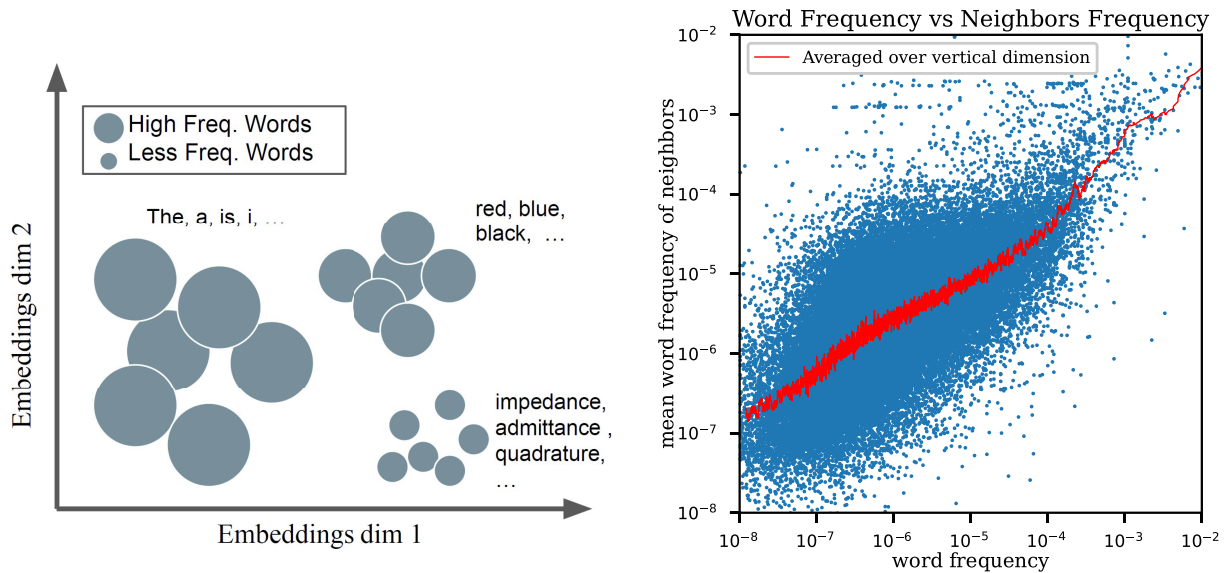


Figure 1. Highly used words stay closer to other highly used words, defining semantically popular regions. Left: intuitive cartoon. Right: scatter plots calculated using *Word2vec* and *wordfreq* datasets.

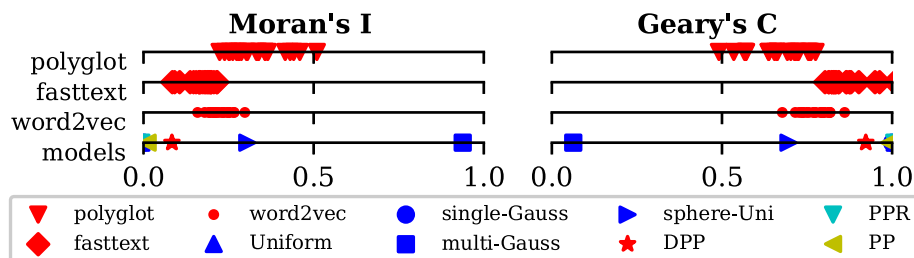


Figure 2. Spatial auto-correlation is consistent across languages and embedding algorithms—some models are able to replicate it. Top three rows: word embedding algorithms applied to 22 languages (mean values). Bottom row: generating models, with most of them counterfactually turning out at the extremes of 0 or 1. The null hypothesis of no spatial correlation is associated with the value of 0.

rather than placing new entities at the same radius around all existing entities. The DPP model, being aware of the spatial context, pushes new entities away from dense areas, instead of placing new entities uniformly around old entities. See electronic supplementary material, §H, for more details. When fitting these parametric models, we calibrate their parameters by trying to match the values of selected statistics as measured in real-life datasets. Also note that all our generating models replicate the familiar power-law distribution in frequencies (see electronic supplementary material, figure S10 in §I).

Borrowing techniques from mathematical geography [65,66], we can quantify the strength of frequency assortativity with standard spatial correlation statistics: (i) Moran's I , g_I , which may be thought of as a spatial analogue of Pearson's correlation coefficient for linear associations, and which lies between 1 (when the variable in question exhibits similar values in nearby areas) and -1 (when the nearby values are inversely related, as in a chess board), as well as (ii) Geary's C , g_C , which is suitable for measuring nonlinear associations, and here is also mapped to $[-1, 1]$ range.

Indeed, the vocabulary data demonstrate strongly positive spatial correlations, a regularity observed across multiple languages (see figure 2, as well as a more detailed electronic supplementary material, table S1 in §J), even within specific parts of speech (see electronic supplementary material, table S2 in §J). Among our generating models, the uniform and PP models do not capture this (the uniform does not capture even the regular 'rich get richer' effect), but DPP fares fairly well.

These observations suggest that a principal feature of empirical semantic spaces is their non-homogeneity, that there are certain regions that attract multiple high-popularity entities, and this stratification is robust and not blurred out over time (thus exhibiting 'hysteresis' [67]). For example, such high-frequency clusters may collect words reflecting cultural developments that are important now or were important in the past, while low-frequency clusters constitute nascent semantic regions that have already experienced some new word creation but not yet massive word usage.

(b) Clustering velocity profiles

Entities in real embedding spaces naturally organize in structures called clusters, themselves composed of natural sub-clusters. We can characterize these cluster structures by measuring how quickly clusters aggregate when we modulate an appropriate sensitivity parameter, defining a *clustering velocity profile*. Focusing on the density of empirical semantic spaces, we inspect our knowledge datasets for the presence of clusters that become noticeable at a high enough resolution (figure 3 demonstrates how CV helps quantify and visualize where we are on the continuum of possible semantic densities). Observed semantic spaces demonstrate surprisingly regular clustering velocity profiles.

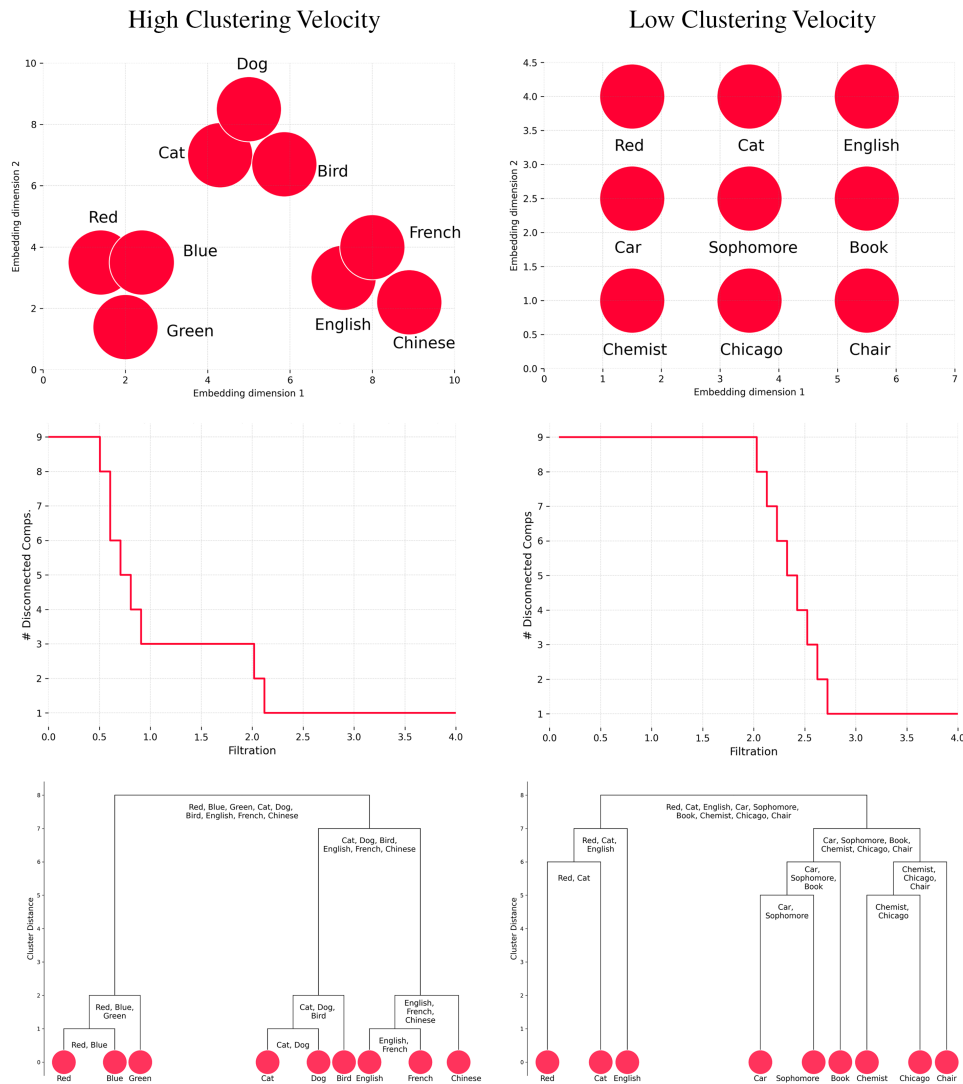


Figure 3. Semantic spaces with differing densities have distinct CV profiles as well as disparate hierarchical clustering dendrograms. Intuitive cartoon. Left: words are close variants of each other. Right: words have very different meanings. Top: embedding spaces. Middle: CV profiles. Bottom: dendrograms.

Specifically, we adapt an algorithm that detects clusters relying on the differences in within- and between-cluster densities (a graph-partitioning algorithm based on graph modularity due to Clauset *et al.* [68] and Newman [69]). What we see strongly resembles a power-law relationship between the number of clusters and the resolution parameter (figure 4, top panel). We repeated this experiment with a popular single-linkage clustering algorithm based on topological considerations, plotting the number of components (the Betti-0 number for simplicial complexes, see [70,71]) against the maximum distance threshold for points to be co-clustered. The resulting filtration curves produce a profile that is no less regular with respect to diverse embeddings (figure 4, bottom). For both algorithms, a more uniform space would produce a more rectangular profile with a more abrupt transition to a single cluster, which is evidently not true in either case. Although the specific shape of the profile depends upon the algorithm used—but, importantly, not on the language identity—our preferred generating model demonstrates a behaviour consistent with both of these regularities, while more simplistic models fail (in addition to figure 4, also see electronic supplementary material, figure S11 and table S3 in SK).

A natural interpretation is that empirical semantic space consists of hierarchical clusters that continuously merge together as we increase the level of aggregation, taking on a higher level of abstraction. The speed of such aggregation is an inherently meaningful property characterizing the richness of the semantic space, and it can be measured and compared across different real embeddings.

(c) Dynamics

This brings us to looking more closely at the dynamics of human knowledge expansion. Static embeddings derived from training data do not explicitly reflect the date when each word entered the language, but the first appearance of vocabulary words in a given language is a topic of long interest in historical linguistics, so independent data sources are readily available. Before proceeding further, we must clarify that the timing of words' origins available to us exists only for the English language, so the results reported next apply only to it.

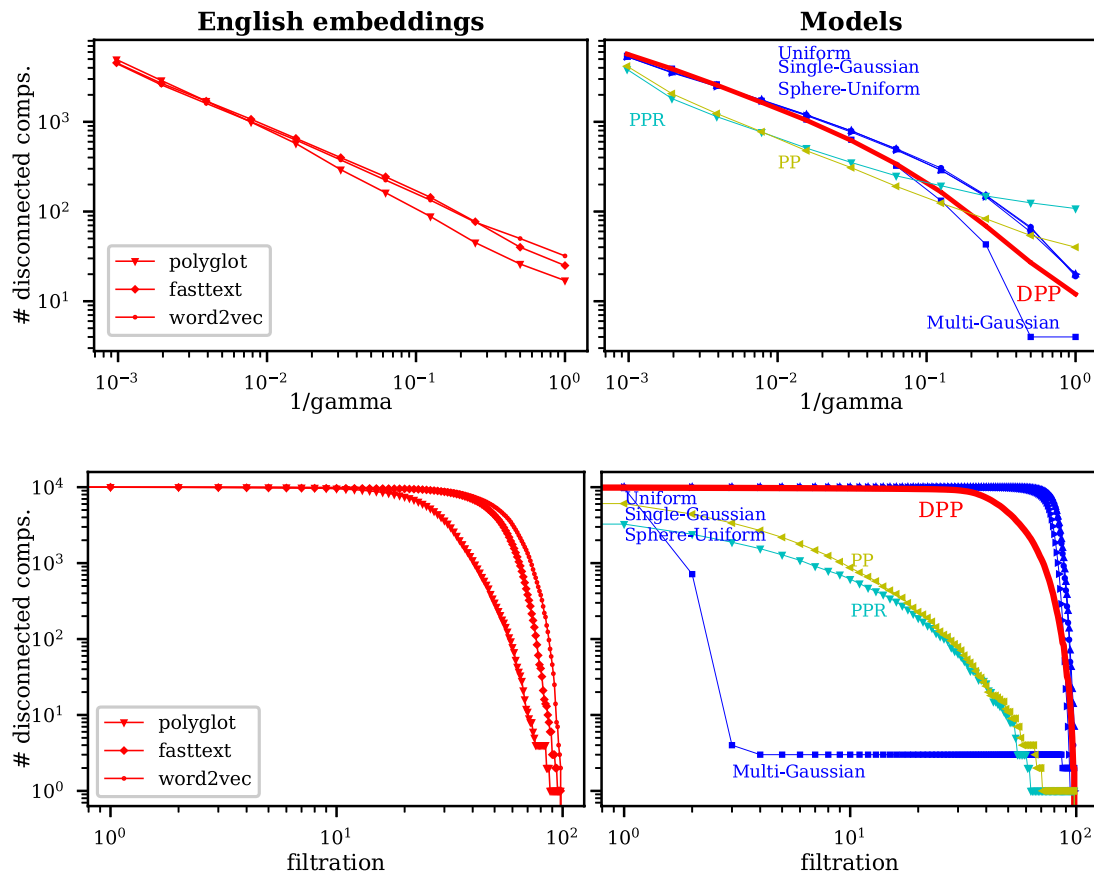


Figure 4. Distinguishable clustering velocity profiles. Number of clusters depending on clustering algorithms' hyperparameters for different knowledge datasets, as well as generating models. Top: graph-theoretic approach. Bottom: topological approach. Left: word embedding algorithms for the English language. Right: generating models.

We can demonstrate that the number of new vocabulary words increases linearly over time, even as respective usage frequencies rise exponentially (see electronic supplementary material, figure S12 in §L). However, as in §2b, here we focus solely on word creation rather than usage (although there were previous studies that explored word usage frequencies from a dynamic perspective, and which also utilize cumulative advantage mechanisms, e.g. [72]).

Using the word creation data, we explore neighbourhood generation dynamics in embedding space. To define local neighbourhoods for analysis, we partition the d -dimensional unit hypersphere of embedding space into $K = 100$ equal-volumed hypercone segments. See electronic supplementary material, §E, for more details. We can then study the dynamics of word creation in terms of the relative activity of such regions at different historical points in time.

We order the words in our dataset by their first-known usage date based on the Stanford histWord dataset (electronic supplementary material, §D), and we count the number of new additions to different cones in the embedding space over time. We find that new words do not appear uniformly but rather in the relatively active regions of the language space, following a linear trend that turns out to be biased in favour of active regions (figure 5). This finding is represented by the curve with a slope of 0.5, which is consistent with the process where every new point is placed in the neighbourhood where its neighbours' age is uniformly distributed between the current year and the oldest available year. The result resonates with the insight that less frequently used words exhibit faster lexical replacement with semantically related new words [42] (in addition to exhibiting faster semantic change, see [22]); as well as with the findings that new words appear in neighbourhoods characterized by the presence of relatively few other words, with these existing words growing in terms of usage frequencies [23].

Finally, by taking the birth of new words as an activity, we investigate the activity level of the word embedding space over time. Focusing on the dependence over time, we visualize the *temporal persistence of activity* in local areas of the embedding space by highlighting consecutive spans of high activity ('hot spots') and low activity ('cold spots'); see figure 6. The persistence pattern is different, in a statistical sense, from the random placement used as a null hypothesis: for hot spots, the average span length is 1.404 (1% confidence interval is [1.330, 1.489]) in real data versus 1.258 (CI [1.195, 1.329]) under the null hypothesis; and for cold spots, it is 1.928 (CI [1.787, 2.103]) versus 1.606 (CI [1.473, 1.761]). (The data source here is Merriam-Webster; the embedding algorithm used is Word2vec.)

To evaluate the performance of our generating models, we define a statistic that measures the co-movement in terms of newly created words within a neighbourhood for the whole embedding space over time. Comparing the results from real datasets and from our generating models, only DPP approaches the empirically observed values, while other generating models exhibit either too strong co-movement or its complete absence (see the description of our statistical measure and the results in electronic supplementary material, table S4 in §M).

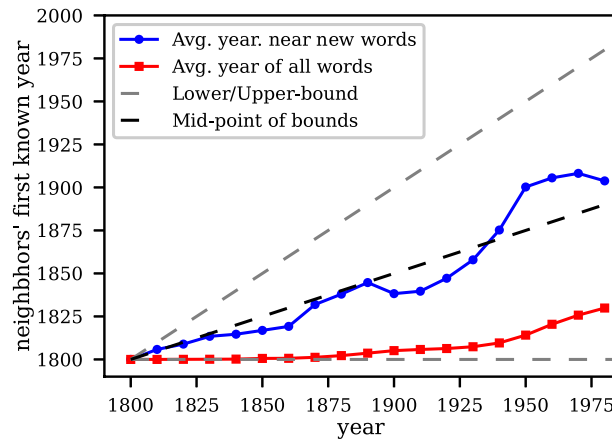


Figure 5. Neighbours of newly inserted words are newer. Grey dashed lines represent min/max possible neighbourhood ages; the black dashed line represents the age of a hypothetical neighbourhood whose elements are chosen uniformly over feasible ages.

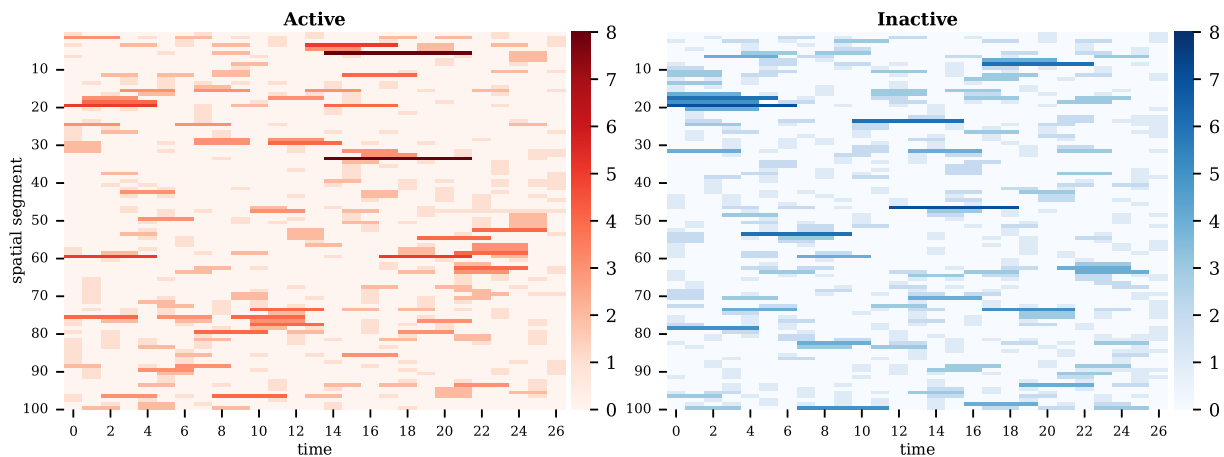


Figure 6. Activity persistence in real-world data. Consecutive spans of high/low activity levels highlighted, with longer spans visually brighter. Time is in decades/years; each spatial segment represents a hypercone in the semantic space. Left: ‘hot spots’ (activity defined as ‘hot’ if more than 50% above the uniform baseline). Right: ‘cold spots’ (‘cold’ if more than 20% below the uniform baseline).

This result is consistent with the finding that language evolves in ‘punctuational bursts’, which may signify the critical times of cultural development [73]. It is also consistent with the findings of Karjus *et al.* [38], who model the dynamics of different lexical topics statistically over time and identify temporarily trending topics that they show to be associated with the emergence of new words. Directed preferential placement processes generate empirically plausible levels of such trends, bursts and lulls.

(d) Taylor’s law

Taking a different perspective on point densities within regions, we document in our knowledge datasets what is known as *Taylor’s power law* of fluctuation scaling (TL) [74]. This is a statement about the distribution of heterogeneous sub-samples of a population in time and space, asserting that variability grows with the average activity. This regularity has been observed before in various domains, including ecology, biology, physics, communications, finance, number theory, etc., and remains open to fundamental causal explanations (such as in [75,76]) or mechanistic statistical interpretations (as in [77–79]).

There has been some prior work on the relevance of Taylor’s law to word frequencies. Kobayashi & Tanaka-Ishii [80] partition long texts into short, equal-sized windows where, for each vocabulary word, they compute the mean and standard deviation of the number of windows in which it appears. Each vocabulary word then defines a point in mean and standard deviation space, which they fit with a line to demonstrate Taylor’s law. Gerlach & Altmann [81] are concerned with laws associated with vocabulary size as a function of the document length. Zipf’s law naturally implies that vocabulary size should grow sublinearly with document length, but they show that the variance of vocabulary size observes Taylor’s law, attributing this fluctuation to differences in topic distribution as measured by probabilistic topic models [82].

We, however, are interested in Taylor’s law as defined in semantic space rather than individual word vocabulary distributions and model the process of new vocabulary creation in languages as opposed to the usage distributions in documents. Specifically, we partition the empirical semantic space into equal-volumed regions (cones) and time into different historical periods. The population of a cone is the set of words whose embeddings semantically fall into it. Then, we study the fluctuations of such populations, focusing on their means and variances as measured over either time or space.

Formally, the so-called spatial Taylor’s law implies that for counts N_{kt} of certain entities x_i living in the embedding space $\mathcal{E}$ and distributed in time and space, where we define $N_{kt} := \#\{i : x_i \in C(k) \text{ \& } y_{it} > 0\}$, with $C(k)$ denoting a region (cone) $k \in \{1, \dots, K\}$,

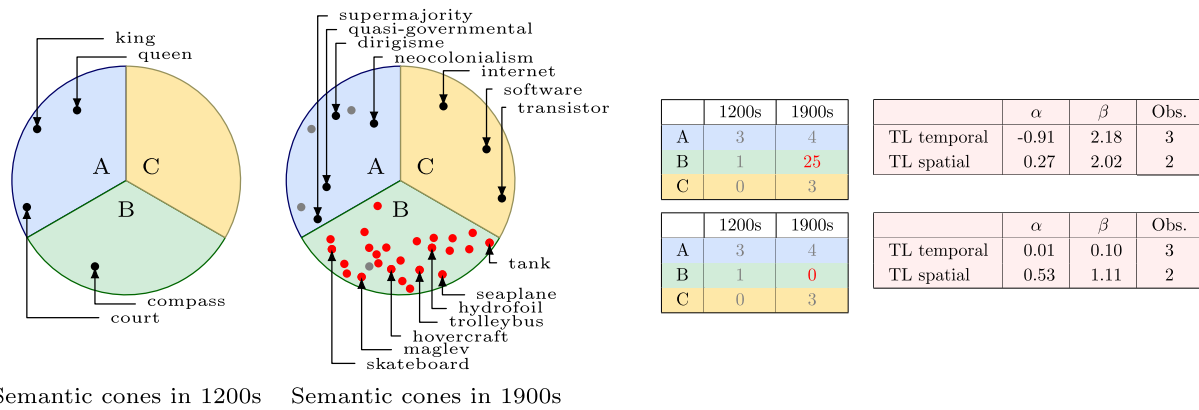


Figure 7. New words are created in different regions of semantic space at different times, in an uneven manner. Intuitive cartoon. Pie charts: across different semantic cones, the addition of new words over time. Polycoloured tables: raw data for TL calculation, with one segment perturbed (in red). Pink tables: estimates for temporal (row-wise regression) and spatial (column-wise) TL, multiplier/offset α and exponent/slope β , as well as the number of observations in regression, with and without data perturbation.

their time period-specific means and variances scale as $V_t[N_{kt}] \approx \alpha(E_t[N_{kt}])^\beta$ with some constants α and β for $t = 1, \dots, T$, where the moments are first computed over the spatial dimension (see figure 7).² The temporal Taylor's law applies in a symmetric fashion to the scaling of spatial unit-specific moments first computed over the temporal dimension.

To provide some intuition about the interpretation of the TL, the exponent β can be viewed as a measure of aggregation, with higher values indicating more heterogeneity and stronger clustering of individuals. Then the spatial TL captures clumping of entities in space, and a higher exponent can be interpreted as greater patchiness, while the temporal TL reflects aggregation through time, with a higher exponent associated with larger fluctuations.

For the analysis of the dynamics of word embeddings, we associate each noun with the year that it first entered the English language, as reported by the Merriam-Webster dictionary, dropping those having no date in the dictionary (see electronic supplementary material, §D, for data source details).

Focusing on the temporal Taylor's law, the implied linear-in-logarithms relationship holds in all our language data: (i) the exponent (slope in the linear regression in logarithms) ranges from 1.8 to 2.1, which is in line with much of the existing literature (a random allocation of points in the embedding space following a quasi-uniform null distribution would produce a slope of 1.0); (ii) the demonstrated goodness-of-fit measures are relatively moderate R^2 levels of 0.69–0.88. Both the structured generating models and the structureless placebo generating models have exponents that are too low (with the exception of multi-Gaussian, but its deficiency is counterfactually high fit), while the DPP falls within the broad range of empirical Taylor's law exponents, and its goodness-of-fit measure is also comparable to that of empirical embeddings (see figure 8 left panel for visualization, also see electronic supplementary material, table S5 in §N, for more detailed numerical results). Turning to the spatial Taylor's law, the estimated empirical exponents of 1.1–1.4 reveal a pattern that is close to what would have been produced by structureless random processes (again, a quasi-uniform null would produce 1.0); at the same time, R^2 levels are not really discriminating in this exercise. Our structureless models produce either a little too low or way too high (multi-Gaussian) exponents; all the structured generating models also demonstrate slopes that are excessively steep, with the only exception being DPP (see figure 8 right, as well as electronic supplementary material, table S6 in §N).

Our analysis of empirical embeddings shows a robust pattern of heterogeneity in vocabulary creation: semantic activity is not happening in a random, haphazard manner, but there often are large outliers above and below the observed average. This is particularly true in the temporal sense, as the heterogeneity manifested in moments calculated across time is higher than that in moments calculated across spatial cones. In other words, ebbs and flows over historic time are more pronounced than the valleys and hills across semantic space. In practice, these quantitative mean–variance relationships can help identify when deviations of activity from the mean are within the usual ranges and when we see truly abnormal outliers. Such regularities can also be used to test alternative generating models, as we do here.

3. Conclusion

Word embeddings provide a geometric formalism that effectively captures the semantics of natural languages. We have demonstrated that such word embeddings display several interesting properties.

Frequency assortativity implies that certain regions of the semantic space get many popular words, which manifests as clusters of high-frequency words (and on the flip side, there are regions of unpopular words). Interestingly, while we know that word embeddings self-organize around parts of speech [83], as we also observe in our empirical embeddings, they are not the sole driver of such clustering: indeed, frequency assortativity is also observed within specific parts of speech. In future research, it would be useful to inspect the characteristics beyond (trivial) semantics that unify the members of the same cluster. For instance, in line with Zipf [84], some of the higher frequency words have many meanings, which suggests that polysemy is an important aspect to consider here. Interestingly, our DPP model can be easily extended to produce not only entirely new words but also, with positive

²Since in our application counts N_{kt} are monotonically increasing over time, to focus on two-way fluctuations, we actually use their increments $\tilde{N}_{kt} = N_{kt} - N_{kt-1}$ instead.

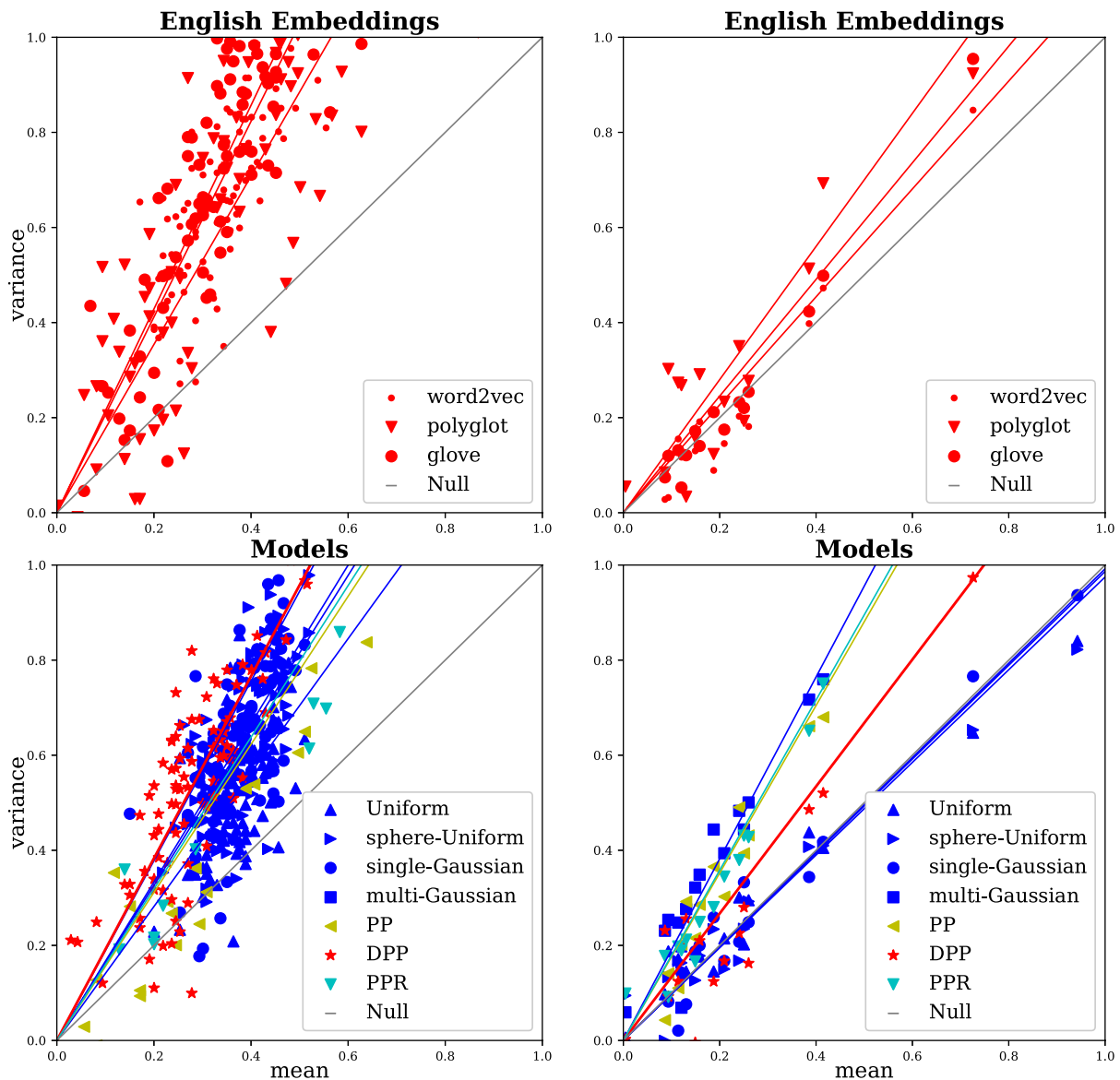


Figure 8. Mean versus variance of word counts over time and across regions of semantic space. Logarithms of the data and linear regression fits are shown. Left: temporal Taylor's law from row-wise (in the sense of figure 7) regression. Right: spatial Taylor's law from column-wise regression. Top: embedding algorithms. Bottom: generating models. Datasets and models (coloured marks); regressions (pale lines); under unconditional quasi-uniform H_0 benchmark (grey line labelled 'null').

probability, new word senses that inherit the old word labels of their polysemous parents (cf. [45]); though appealing, testing such a model would require addressing the limitations of existing curated datasets.

Taking a more global perspective, clustering velocity profiles reflect the degree of non-uniformness of the semantic space in terms of the density of words. For instance, genuinely new words appear in low-density regions and make the semantic space globally more uniform, while new synonyms of old words do so only locally in their respective clusters by not wandering far from their parents, which actually reinforces the global clustering structure. In future work, it would be interesting to verify whether the semantic space becomes more uniform over time, and if so, how fast the speed of such evolution is.

Moving from a static to a dynamic perspective, we find that the creation of new words is not dispersed evenly across the semantic space but is concentrated in certain active regions. Moreover, such regions exhibit persistence over time: active regions in a given period are likely to remain active in the next period (and vice versa for dormant regions). It is fruitful to correlate the bursts of such semantic activity with shifts and trends in the cultural, demographic and technological areas over human history (cf. [38]).

Finally, we inspected and quantified the heterogeneity in the distribution of word creation over the semantic space as well as across time through the lens of Taylor's law (i.e. in the corresponding spatio-temporal log-linear framework). We have identified substantial temporal heterogeneity, with large fluctuations in the appearances of new words over time. From our earlier results, we already knew that over time, semantic cones fluctuate between abundance and scarcity, but using the TL-based approach allows us to quantify this. Conducting a comparative quantitative analysis for languages other than English would be useful, as the differences or their lack would be an interesting result in itself.

Stochastic processes that generate embedding-like point sets with analogous properties define mechanisms that may help explain fundamental aspects of language evolution. Our preferred generating process, directed preferential placement, is one such

process that broadly matches the properties of natural human languages. It pushes new words away from high-density and high-frequency regions, always trying to ‘seed’ new semantic regions. Our simulation results are consistent with the following empirical pattern. Whenever a semantic region has a stable period, words in dense subregions are re-used and their frequencies are growing. This shows up as a ‘cold’ spot of frequency-rich neighbourhoods getting ‘richer’. Whenever a region goes through bursts of creation (like punctuated equilibria in biological evolution), new words emerge in empty subregions to reflect the interest in new cultural phenomena. This shows up as a ‘hot’ spot of ‘poor’ neighbourhoods staying ‘poor’ (initially, at least). And each semantic region oscillates between ‘hot’ and ‘cold’ states, accumulating frequencies along the way. As possible testable hypotheses for future studies, we ponder what happens in between the above two extremes. Perhaps the evolution of a semantic area goes through such stages as (i) little creation and little use, (ii) high creation and limited use, (iii) a lot of use following previous creation, and (iv) gradual cooling of use and creation.

Similar embeddings (vector representations) can be constructed for other cultural objects, including social networks and information sources such as Wikipedia. We believe that the directed preferential placement mechanism may model developments in domains far beyond natural languages.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. All data and code have been archived and are available on Dryad [85].

Supplementary material is available online [86].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. X.G.: conceptualization, data curation, investigation, methodology, validation, visualization, writing—original draft; S.V.: conceptualization, formal analysis, methodology, writing—original draft; H.C.: conceptualization, data curation, investigation; B.Z.: conceptualization, data curation, investigation, methodology, visualization; S.S.: conceptualization, funding acquisition, methodology, resources, writing—original draft.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. S.V. was supported by the Center of Mathematical Sciences and Applications at Harvard University. S.S. was partially supported by NSF grants IIS-1926781, IIS-1927227, IIS-1546113, OAC-191952 and a New York State Empire Innovation grant.

References

- Binder P, Smith K. 2013 *The language phenomenon: human communication from milliseconds to millennia*. Heidelberg, Germany: Springer.
- Smith JM, Szathmáry E. 1995 *The major transitions in evolution*. Oxford, UK: Freeman.
- Barabási AL, Albert R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512. (doi:10.1126/science.286.5439.509)
- Merton RK. 1968 The Matthew effect in science. *Science* **159**, 56–63. (doi:10.1126/science.159.3810.56)
- Simon HA. 1955 On a class of skew distribution functions. *Biometrika* **42**, 425–440. (doi:10.1093/biomet/42.3-4.425)
- Zipf GK. 1949 *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge, MA: Addison-Wesley Press.
- Egghe L. 2007 Untangling Herdan's law and Heaps' law: mathematical and informetric arguments. *J. Am. Soc. Inf. Sci.* **58**, 702–709. (doi:10.1002/asi.20524)
- Heaps HS. 1978 *Information retrieval: computational and theoretical aspects*. New York, NY: Academic Press.
- Herdan G. 1960 *Type-token mathematics*. The Hague, Netherlands: Mouton.
- Blasi D, Henrich J, Adamou E, Kemmerer D, Majid A. 2022 Over-reliance on English hinders cognitive science. *Trends Cogn. Sci.* **26**, 1153–1170.
- Mikolov T, Chen K, Corrado G, Dean J. 2013 Efficient estimation of word representations in vector space. arXiv 1301.3781.
- Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013 Distributed representations of words and phrases and their compositionality. In *Proc. 26th International Conf. Neural Information Processing Systems*, vol. 2, pp. 3111–3119, Curran Associates.
- Goyal P, Ferrara E. 2018 Graph embedding techniques, applications, and performance: a survey. *Knowl. Based Syst.* **151**, 78–94. (doi:10.1016/j.knsys.2018.03.022)
- Hamilton WL. 2020 Graph representation learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **14**, 1–159. (doi:10.1007/978-3-031-01588-5_1)
- Baevski A, Hsu WN, Xu Q, Babu A, Gu J, Auli M. 2022 data2vec: a general framework for self-supervised learning in speech, vision and language. In *International Conf. Machine Learning*, pp. 1298–1312. Proceedings of Machine Learning Research.
- Thompson B, Roberts SG, Lupyan G. 2020 Cultural influences on word meanings revealed through large-scale semantic alignment. *Nat. Hum. Behav.* **4**, 1–10. (doi:10.1038/s41562-020-0924-8)
- Caliskan A, Bryson JJ, Narayanan A. 2017 Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186. (doi:10.1126/science.aal4230)
- Jones J, Amin M, Kim J, Skiena S. 2020 Stereotypical gender associations in language have decreased over time. *Sociol. Sci.* **7**, 1–35. (doi:10.15195/v7.a1)
- Adukia A, Chiril P, Christ C, Das A, Eble A, Harrison E, Runesha HB. 2021 Tales and tropes: gender roles from word embeddings in a century of children's books. *BFI Working Paper 2021-128*. (doi:10.2139/ssrn.4221213)
- Garg N, Schiebinger L, Jurafsky D, Zou J. 2018 Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl Acad. Sci. USA* **115**, 3635–3644. (doi:10.1073/pnas.1720347115)
- Kozłowski AC, Taddy M, Evans JA. 2019 The geometry of culture: analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* **84**, 905–949. (doi:10.1177/0003122419877135)
- Hamilton WL, Leskovec J, Jurafsky D. 2016 Diachronic word embeddings reveal statistical laws of semantic change. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, pp. 1489–1501. Stroudsburg, PA, USA: Association for Computational Linguistics. (doi:10.18653/v1/P16-1141)
- Ryskina M, Rabinovich E, Berg-Kirkpatrick T, Mortensen D, Tsvetkov Y. 2020 Where new words are born: distributional semantic analysis of neologisms and their semantic neighborhoods (eds A Ettinger, G Jarosz, J Pater). In *Proc. Society for Computation in Linguistics 2020*, pp. 367–376. New York, NY, USA: Association for Computational Linguistics.
- Grand G, Blank IA, Pereira F, Fedorenko E. 2022 Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat. Hum. Behav.* **6**, 975–987. (doi:10.1038/s41562-022-01316-8)
- van Loon A, Freese J. 2023 Word embeddings reveal how fundamental sentiments structure natural language. *Am. Behav. Sci.* **67**, 175–200. (doi:10.31235/osf.io/r7ewx)

26. Karjus A, Cuskley C. 2024 Evolving linguistic divergence on polarizing social media. *Humanit. Soc. Sci. Commun* **11**, 1–14. (doi:10.1057/s41599-024-02922-9)
27. Gray RD, Atkinson QD, Greenhill SJ. 2011 Language evolution and human history: what a difference a date makes. *Phil. Trans. R. Soc. B Biol. Sci.* **366**, 1090–1100. (doi:10.1098/rstb.2010.0378)
28. Greenhill SJ, Atkinson QD, Meade A, Gray RD. 2010 The shape and tempo of language evolution. *Proc. R. Soc. B* **277**, 2443–2450. (doi:10.1098/rspb.2010.0051)
29. Greenhill SJ, Currie TE, Gray RD. 2009 Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. B* **276**, 2299–2306. (doi:10.1098/rspb.2008.1944)
30. Greenhill SJ, Wu CH, Hua X, Dunn M, Levinson SC, Gray RD. 2017 Evolutionary dynamics of language systems. *Proc. Natl Acad. Sci. USA* **114**, E8822–E8829. (doi:10.1073/pnas.1700388114)
31. Nicholls GK, Gray RD. 2006 Quantifying uncertainty in a stochastic model of vocabulary evolution. In *Phylogenetic methods and the prehistory of languages* (eds P Forster, C Renfrew). Cambridge, UK: McDonald Institute for Archaeological Research.
32. Shcherbakova O, Gast V, Blasi DE, Skirgård H, Gray RD, Greenhill SJ. 2023 A quantitative global test of the complexity trade-off hypothesis: the case of nominal and verbal grammatical marking. *Linguist. Vanguard* **9**, 155–167. (doi:10.1515/lingvan-2021-0011)
33. Bochkarev V, Shevlyakova A, Solov'yev V. 2015 Average word length dynamics as indicator of cultural changes in society. *Soc. Evol* **14**, 153–175.
34. Bochkarev V, Solov'yev V, Wichmann S. 2014 Universals versus historical contingencies in lexical evolution. *J. R. Soc. Interface* **11**, 20140841. (doi:10.1098/rsif.2014.0841)
35. Calude AS, Miller S, Pagel M. 2020 Modelling loanword success – a sociolinguistic quantitative study of Māori loanwords in New Zealand English. *Corpus Linguist. Linguist. Theory* **16**, 29–66. (doi:10.1515/clt-2017-0010)
36. Feltgen Q, Fagard B, Nadal JP. 2017 Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change. *R. Soc. Open Sci.* **4**, 170830. (doi:10.1098/rsos.170830)
37. Grieve J. 2018 Natural selection in the modern english lexicon. In *Proc. 12th International Conf. Evolution of Language*, pp. 153–157. Torun, Poland: Evolang 12 Organizing Committee. (doi:10.12775/3991-1.037)
38. Karjus A, Blythe RA, Kirby S, Smith K. 2020 Quantifying the dynamics of topical fluctuations in language. *Lang. Dyn. Change* **10**, 86–125. (doi:10.1163/22105832-01001200)
39. Lieberman E, Michel JB, Jackson J, Tang T, Nowak MA. 2007 Quantifying the evolutionary dynamics of language. *Nature* **449**, 713–716. (doi:10.1038/nature06137)
40. Michel JB *et al.* 2011 Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182. (doi:10.1126/science.1199644)
41. Newberry MG, Ahern CA, Clark R, Plotkin JB. 2017 Detecting evolutionary forces in language change. *Nature* **551**, 223–226. (doi:10.1038/nature24455)
42. Pagel M, Atkinson QD, Meade A. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720. (doi:10.1038/nature06176)
43. Montero JG, Karjus A, Smith K, Blythe RA. 2025 Reliable detection and quantification of selective forces in language change. *Corpus Linguist. Linguist. Theory* **21**, 31–73. (doi:10.1515/clt-2023-0064)
44. Pasquini M, Serva M, Vergni D. 2023 Gradual modifications and abrupt replacements: two stochastic lexical ingredients of language evolution. *Comput. Linguist.* **49**, 301–323. (doi:10.1162/coli_a_00471)
45. Ramiro C, Srinivasan M, Malt BC, Xu Y. 2018 Algorithms in the historical emergence of word senses. *Proc. Natl Acad. Sci. USA* **115**, 2323–2328. (doi:10.1073/pnas.1714730115)
46. Jackson JC, Watts J, Henry TR, List JM, Forkel R, Mucha PJ, Greenhill SJ, Gray RD, Lindquist KA. 2019 Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522. (doi:10.1126/science.aaw8160)
47. Altmann EG, Gerlach M. 2016 Statistical laws in linguistics. In *Lecture notes in morphogenesis*, pp. 7–26. Heidelberg, Germany: Springer. (doi:10.1007/978-3-319-24403-7_2)
48. Eggenberger F, Pólya G. 1923 Über die Statistik verketteter Vorgänge. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift Für Angewandte Mathematik Und Mechanik* **3**, 279–289. (doi:10.1002/zamm.19230030407)
49. Yule GU. 1925 II. —A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Phil. Trans. R. Soc. B Contain. Pap. Biol. Character* **213**, 21–87. (doi:10.1098/rstb.1925.0002)
50. Price DDS. 1976 A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inf. Sci.* **27**, 292–306. (doi:10.1002/asi.4630270505)
51. Pitman J, Yor M. 1997 The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900. (doi:10.1214/aop/1024404422)
52. Teh YW. 2006 A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proc. 21st International Conf. Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp. 985–992. Morristown, NJ, USA: Association for Computational Linguistics. (doi:10.3115/1220175.1220299)
53. Sato I, Nakagawa H. 2010 Topic models with power-law using Pitman-Yor process. In *Proc. 16th ACM SIGKDD International Conf. Knowledge Discovery and Data Mining, KDD '10*, Washington, DC, pp. 673–682. New York, NY: Association for Computing Machinery. (doi:10.1145/1835804.1835890)
54. Hopkins M. 2022 Towards more natural artificial languages. In *Proc. 26th Conf. Computational Natural Language Learning (CoNLL)*, Abu Dhabi, United Arab Emirates (Hybrid), pp. 85–94. Stroudsburg, PA: Association for Computational Linguistics. (doi:10.18653/v1/2022.conll-1.7)
55. Nowak MA, Krakauer DC. 1999 The evolution of language. *Proc. Natl Acad. Sci. USA* **96**, 8028–8033. (doi:10.1073/pnas.96.14.8028)
56. Smith K. 2004 The evolution of vocabulary. *J. Theor. Biol.* **228**, 127–142. (doi:10.1016/j.jtbi.2003.12.016)
57. Kirby S, Dowman M, Griffiths TL. 2007 Innateness and culture in the evolution of language. *Proc. Natl Acad. Sci. USA* **104**, 5241–5245.
58. Real F, Griffiths TL. 2010 Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proc. R. Soc. B* **277**, 429–436. (doi:10.1098/rspb.2009.1513)
59. Blythe RA. 2012 Neutral evolution: a null model for language dynamics. *Adv. Complex Syst.* **15**, 1150015. (doi:10.1142/S0219525911003414)
60. Croft W. 2000 *Explaining language change: an evolutionary approach*. London, UK: Longman.
61. Kulkarni V, Al-Rfou R, Perozzi B, Skiena S. 2015 Statistically significant detection of linguistic change. In *Proc. 24th International Conf. World Wide Web*, Florence, Italy, pp. 625–635. Republic and Canton of Geneva, Switzerland. (doi:10.1145/2736277.2741627)
62. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. 2016 Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. <https://proceedings.neurips.cc/paper/2016/file/a486cd07e4a3d270571622f4f316ec5-Paper.pdf>.
63. Manzini T, Lim YC, Tsvetkov Y, Black AW. 2019 Black is to criminal as caucasian is to police: detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers)*, Minneapolis, Minnesota. Stroudsburg, PA: Association for Computational Linguistics. (doi:10.18653/v1/N19-1062)
64. Periti F, Montanelli S. 2024 Lexical semantic change through large language models: a survey. *ACM Comput. Surv.* **56**, 1–38. (doi:10.1145/3672393)
65. Anselin L. 1995 Local indicators of spatial association—LISA. *Geogr. Anal.* **27**, 93–115. (doi:10.1111/j.1538-4632.1995.tb00338.x)
66. O'Sullivan D, Unwin DJ. 2010 *Geographic information analysis*. Hoboken, NJ: John Wiley & Sons.
67. Ewing JA, Jenkin HCF. 1882 VII. On the production of transient electric currents in iron and steel conductors by twisting them when magnetised or by magnetising them when twisted. *Proc. R. Soc. Lond.* **33**, 21–23. (doi:10.1098/rsp1.1881.0067)

68. Clauset A, Newman ME, Moore C. 2004 Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111.
69. Newman MEJ. 2016 Equivalence between modularity optimization and maximum likelihood methods for community detection. *Phys. Rev. E* **94**, 052315. (doi:10.1103/PhysRevE.94.052315)
70. Carlsson G, Vejdemo-Johansson M. 2021 *Topological data analysis with applications*. Cambridge, UK: Cambridge University Press.
71. Rabadan R, Blumberg AJ. 2019 *Topological data analysis for genomics and evolution: topology in biology*. Cambridge, UK: Cambridge University Press.
72. Perc M. 2012 Evolution of the most common English words and phrases over the centuries. *J. R. Soc. Interface* **9**, 3323–3328. (doi:10.1098/rsif.2012.0491)
73. Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M. 2008 Languages evolve in punctuational bursts. *Science* **319**, 588–588. (doi:10.1126/science.1149683)
74. Taylor L. 1961 Aggregation, variance and the mean. *Nature* **189**, 732–735. (doi:10.1038/189732a0)
75. Anderson RM, Gordon DM, Crawley MJ, Hassell MP. 1982 Variability in the abundance of animal and plant species. *Nature* **296**, 245–248. (doi:10.1038/296245a0)
76. Cohen JE, Xu M, Schuster WSF. 2013 Stochastic multiplicative population growth predicts and interprets Taylor's power law of fluctuation scaling. *Proc. R. Soc. B Biol. Sci.* **280**, 20122955. (doi:10.1098/rspb.2012.2955)
77. Eisler Z, Bartos I, Kertész J. 2008 Fluctuation scaling in complex systems: Taylor's law and beyond1. *Adv. Phys.* **57**, 89–142. (doi:10.1080/00018730801893043)
78. Cohen JE, Xu M. 2015 Random sampling of skewed distributions implies Taylor's power law of fluctuation scaling. *Proc. Natl Acad. Sci. USA* **112**, 7749–7754. (doi:10.1073/pnas.1503824112)
79. Cohen JE. 2020 Every variance function, including Taylor's power law of fluctuation scaling, can be produced by any location-scale family of distributions with positive mean and variance. *Theor. Ecol.* **13**, 1–5. (doi:10.1007/s12080-019-00445-7)
80. Kobayashi T, Tanaka-Ishii K. 2018 Taylor's law for human linguistic sequences. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, pp. 1138–1148. Stroudsburg, PA, USA: Association for Computational Linguistics. (doi:10.18653/v1/P18-1105)
81. Gerlach M, Altmann EG. 2014 Scaling laws and fluctuations in the statistics of word frequencies. *New J. Phys.* **16**, 113010. (doi:10.1088/1367-2630/16/11/113010)
82. Blei DM. 2012 Probabilistic topic models. *Commun. ACM* **55**, 77–84. (doi:10.1145/2133806.2133826)
83. Perozzi B, Al-Rfou R, Skiena S. 2014 Deepwalk: online learning of social representations. In *Proc. 20th ACM SIGKDD International Conf. Knowledge Discovery and Data Mining*, pp. 701–710. New York, NY: Association for Computing Machinery. (doi:10.1145/2623330.2623732)
84. Zipf GK. 1945 The meaning-frequency relationship of words. *J. Gen. Psychol.* **33**, 251–256. (doi:10.1080/00221309.1945.10544509)
85. Guo X *et al.* 2026 Data and code from: statistical structure and the evolution of languages. Dryad Digital Repository. (doi:10.5061/dryad.pg4f4qs44)
86. Guo X, Verstyuk S, Chen H, Zhou B, Skiena S. 2026 Supplementary material from: Statistical Structure and the Evolution of Languages. Figshare. (doi:10.6084/m9.figshare.c.8383184)