A Mathematical Model for Linguistic Universals

Weinan E\textsuperscript{1,2}\textsuperscript{*}, Yajun Zhou\textsuperscript{2}\textsuperscript{*}

\textsuperscript{1}Department of Mathematics & Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA  
\textsuperscript{2}Beijing Institute of Big Data Research, Beijing 100871, P. R. China  

\textsuperscript{*}Corresponding authors. E-mail: weinan@math.princeton.edu (W.E), yajun.zhou.1982@pku.edu.cn (Y.Z.)

**Inspired by chemical kinetics and neurobiology, we propose a mathematical theory for pattern recurrence in text documents, applicable to a wide variety of languages. We present a Markov model at the discourse level for Steven Pinker’s “mentalese”, or chains of mental states that transcend the spoken/written forms. Such (potentially) universal temporal structures of textual patterns lead us to a language-independent semantic representation, or a translationally-invariant word embedding, thereby forming the common ground for both comprehensibility within a given language and translatability between different languages. Applying our model to documents of moderate lengths, without relying on external knowledge bases, we reconcile Noam Chomsky’s “poverty of stimulus” paradox with statistical learning of natural languages.**

We human beings distinguish ourselves from other animals (1–3), in that our brain development (4–6) enables us to convey sophisticated ideas and to share individual experiences, via languages (7–9). Texts written in natural languages constitute a major medium that perpetuates our civilizations (10), as a cumulative body of knowledge. The quantitative mechanism underlying the mental faculties of language has long been a difficult problem for anthropologists, linguists, neurobiologists and psychologists (11–15), before attracting the attention of computer and data scientists (16–21), in the recent wave of artificial intelligence. Instead of marveling at the partial success of data-hungry approaches (18–21) to machine learning, we still crave for a cost-effective, interpretable and universal algorithm for understanding natural languages—one that mimics language acquisition and knowledge accumulation during early childhood, based on limited resources, as in Chomsky’s “poverty of stimulus” scenario (11, 15). Without filling the gap of data sizes, one cannot satisfactorily answer nativists’ criticism (22) against empiricists’ statistical models for natural languages.

Rising to the challenges outlined above, we perform a detailed mathematical analysis for computable “linguistic universals”—statistical patterns common to a wide range of human languages. On the theoretical side, we will present a stochastic “mentalese” model that depicts the timecourse of Markov states behind individual concepts. On the practical side, we will demonstrate (through automated word translation and question answering) that word’s meaning can be numerically characterized by moderate-sized Markov neural networks, even when there is relatively scant data input.

Our Markov model explains, up to acceptably small error margins, how our innate language faculties (nature) may help us understand the world, by connecting dots of our past experiences (nurture), irrespective of our mother tongue. Bridging nature to nurture, our stochastic algorithm for Markov neural semantics reconciles the views of nativists and empiricists.

**Heuristic background**  Languages differ in their phonemic repertoires (“elementary particles” in Jakobson’s (24) terms), word morphologies (“atoms”) and syntactic structures (“molecules”), corresponding to the three short time scales (phonological processing level, lexical level, and sentence level) in the Friederici hierarchy (13), which are mapped to different brain regions in functional magnetic resonance imaging (fMRI). These three Friederici scales exhibit no universal linguistic patterns and bear no semantic significance. Ferdinand de Saussure’s foundational work (9) rules out semantic dependence on phonological representation (except for a limited set of onomatopoeias), while the inherent meaning of a word is affected by neither its morphological parameters (say, singular vs. plural, present vs. past) nor its syntactic rôles (say, subject vs. object, active vs. passive).

Based on the foregoing arguments, one might speculate that universal semantic content, or Pinker’s “mentalese” (12), may only exist at the discourse level (“bulk materials”, if we extrapolate Jakobson’s (24) metaphor), namely, on the longest time scale in Friederici’s neurobiological hierarchy (13). In this work, we turn such a qualitative speculation into a quantitative model (25). Concretely speaking, we observe the following statistical features of textual patterns (clusters of words that are morphologically related, see Fig. 1 and Fig. 2 for examples) shared by many languages in common:

1. The recurrence behavior of most textual patterns is consistent with time series generated by a certain Markov process, on the longest, as opposed to the shortest (26), neuro-linguistic time scale;
$W_i = \{\text{happier, happily, happiness, happy}\}, W_j = \{\text{marriage, married, marry}\}$

2. Recurrence kinetics of a given concept nearly remains independent of the language in which it is expressed;

3. Kinetic data quantify the semantic distance between different textual patterns, thus allowing us to construct semantic fields by statistical computations.

These long-range temporal features of documents written in various languages, in our opinion, point to a universal kinetic mechanism that defines the semantic roles of individual nodes in a web of words, mathematically and linguistically.

**Transition probability, spectral invariance, pattern recurrence** To begin, we show how to numerically construct a Markov matrix from a realistic text document, and how this Markov model enables us to interpret long-range temporal structures that are common to a wide variety of languages.

In our kinetic language model, we assume that (the gists of) texts are generated by a discrete Markov process on a semantic web with $N$ nodes, each of which represents a textual pattern $W_k$—a set of morphologically related content words (29), indexed by an integer $k \in \{1,2,\ldots,N\}$—occurring in a given document (30). The stochastic hoppings between the nodes are governed by an ergodic Markov transition matrix $P = (p_{ij})_{1 \leq i,j \leq N}$, which (putatively) caricatures the dynamics of mental activities (31) underlying a text, on the time scale of discourse level. We emphasize that our Markov model for the long-range behavior of human languages is independent of Chomsky’s transformational generative grammar (32), the latter of which characterizes short-range syntactic features as hierarchical trees without Markovian structures.

One can estimate transition probabilities between textual patterns on short time scales, by simply counting unigrams and bigrams in a large corpus (33). To estimate long-range transition probabilities from documents of moderate lengths (e.g. a literary piece, a Wikipedia page), i.e. to learn despite a “poverty of the stimulus”, we need some makeshift strategies.

Given a timecourse of molecular states in a biochemical reaction, we can partially reconstruct kinetic information (34) from the probability distribution for the waiting time between consecutive encounters of the same molecular state. Carrying this waiting time analysis a little further, we put a crude estimate of the transition probability $p_{ij}$ as

$$p_{ij} := \frac{n_{ij} e^{-(\log L_{ij})}}{\sum_{j'=1}^N n_{i j'} e^{-(\log L_{ij'})}},$$

where $n_{ij}$ counts the number of effective transitions from $W_i$ to $W_j$, and $L_{ij}$ is a statistic that measures the reduced fragment lengths of such transitions (Fig. 1). On the diagonal, the Gibbs weights $n_{ii} e^{-(\log L_i)}$ hearken back to the TF-IDF measure of word importance (35, 36). Off the diagonal, the ensemble average $\langle \log L_{ij} \rangle$ weighs the cost of biochemical activation energy required to jump from $W_i$ to $W_j$, so that the memorability factor $e^{-\langle \log L_{ij} \rangle}$ can be viewed as a naive estimate for the rate of associative learning per copy number $n_{ij}$, in Hebb’s fire-and-wire process (37). It is worth noting that our estimate of $\tilde{p}_{ij}$ was based on statistical analysis of the text in situ, without digesting a document (or small parts of it) as a scrambled bag of words, a procedure implemented in conventional algorithms (16, 18, 20).

The empirical Markov matrix $\hat{P} = (\hat{p}_{ij})_{1 \leq i,j \leq N}$ has some desirable properties.

First, the components of the dominant eigenvector $\vec{\pi}$ (satisfying $a \vec{P} = \vec{\pi}$) are nearly proportional to the word counts contributed by individual text patterns (Fig. 2A). This is consistent with the fact that the ergodic Markov matrix $P$ has a unique equilibrium state $\vec{\pi}$ (satisfying the invariant measure equation $\vec{\pi} \vec{P} = \vec{\pi}$), whose vector components represent the probabilities of encountering individual Markov states.

Second, the spectrum $\sigma(\hat{P})$ is approximately invariant against translations of texts (Fig. 2B). One can explain this
Figure 3: (A) Precipitous decays of \( r_n =: \frac{1}{2} \sum_{i,j \leq 100} |\pi_i P(n)_{ij} - \pi_j P(n)_{ji}| \) from the initial value \( r_1 \approx 0.07 \), for matrix powers \( P^n = (P(n)_{ij})_{i,j \leq 100} \) constructed from four versions of *Pride and Prejudice*. (In contrast, one has \( r_1 \approx 0.33 \) for a random 100×100 Markov matrix.) Such quick relaxations support our working hypothesis about detailed balance \( \pi_i P_{ij} = \pi_j P_{ji} \), with font size proportional to the square root of the memorability factor \( e^{-\log L_{ij}} \) (see scheme S2 for word stacking methods, and figs. S7–S20 for further examples). Topical (i.e. significantly non-Poissonian) patterns painted in red (resp. green) reside below (resp. above) the critical line of Poissonian banality (blue line in C), where the deviations exceed the error margin prescribed in [5]. (C) Recurrence statistics for textual patterns (gray, red and green dots with radii \( \sqrt{n} \)). Labels for proper names and some literary motifs are attached next to the corresponding colored dots. By Jensen’s inequality, all the data points must sit below the *green dashed line* with unit slope and zero intercept. Actually, almost all data points lie beneath [up to error margin prescribed in [5]] the blue line with unit slope and intercept \( -\gamma_0 \), a phenomenon that is predicted by detailed balance [cf. 4]. (D) Ružička similarities \( s_k(b^m_{en}, b^j_{en}) \) between selected topics (sorted by descending \( n_k \geq 20 \)) in English and French versions (see tables S1 and S2 for stylistic variations in translations) of *Pride and Prejudice*. Rows and columns with maximal \( s_k(b^m,e_{en}, b^j_{en}) \) less than 0.7 are not shown. Correct matchings are indicated by green cross-hairs.

through a thought experiment involving two monolingual subjects, Alice and Bob, whose native languages are A and B, respectively. Suppose that we ask Alice to measure the transition probability from words in language A to words in language B. Alice first processes the input in her native language by a Markov matrix \( P_A \), and then translates into language B, using a dictionary matrix \( T_{A-B} \). When Bob is assigned the same task, he needs to first translate the input into language B, using the same dictionary \( T_{A-B} \), before brainstorming in his own native language, using \( P_B \). Putatively, semantic content is universal (shared by native speakers of different languages), so we must have

\[
P_A T_{A-B} = T_{A-B} P_B. \tag{2}
\]

In the ideal scenario where translation is lossless (with invertible \( T_{A-B} \)), the Markov matrices \( P_A \) and \( P_B \) are indeed linked to each other by a similarity transformation that leaves their spectrum intact.

Third, the empirical matrix \( \hat{P} \) approaches detailed balance after a few iterations (Fig. 3A). Since textual patterns typically recur on the longest time scale in the Friederici hierarchy, we may further assume, for the sake of convenience, that the ergodic Markov transition matrix \( P \) honors the detailed balance condition \( \pi_i P_{ij} = \pi_j P_{ji} \) for \( 1 \leq i, j \leq N \).

On a Markov chain with detailed balance, the recurrence time \( \tau = \tau_{ij} \) of any individual state \( W_i \) is distributed as a weighted superposition of exponential decays (see fig. S1 for multi-exponential fits)

\[
P(\tau > t) = \sum_m c_m e^{-k_m t},
\]

where \( c_m, k_m > 0 \), and \( \sum_m c_m = 1 \),

a functional form that frequently crops up in dynamic studies of biological macromolecules (34). The multi-exponential decay laws impose an inequality constraint:

\[
\langle \log L_{ij} \rangle - \langle \log L_{ij} \rangle + \gamma_0 = \sum_m c_m \log \frac{1}{k_m} - \log \sum_m c_m k_m \leq 0, \tag{4}
\]

where \( \gamma_0 = 0.57721566 \ldots \) is the Euler–Mascheroni constant.
We postulate that a banal textual pattern tends to appear in a document in a memoryless manner, as in a Poisson process (fig. S1B). Therefore, the effective time elapse \( \tau = L_{ii} \) between consecutive encounters of a banal \( W_i \) obeys an exponential distribution \( P(L_{ii} > \tau) \sim e^{-\tau k} \) for a single kinetic rate constant \( k \) > 0. The equality \((\log L_{ii}) - \langle \log L_{ii} \rangle + \gamma_0 = 0 \) (blue line in fig. [2]) marks Poissonian balanlity when the text is infinitely long. If we have \( n_{ii} \) independent samples of exponentially distributed random variables \( L_{ii} \), then the statistic \( \hat{\delta}_i := \log(L_{ii}) - \langle \log L_{ii} \rangle - \gamma_0 + \frac{1}{2n_{ii}} \) satisfies an inequality

\[
|\hat{\delta}_i| < \frac{2}{\sqrt{n_{ii}}} \left( \sqrt{\frac{\pi^2}{6}} - 1 - \frac{1}{2n_{ii}} \right)
\]

with probability 95%. As a working definition, we consider a textual pattern topological, if its \( \hat{\delta}_i \) value violates the aforementioned inequality (Figs. 3B and C), corresponding to significantly non-Poissonian recurrence statistics.

If a textual pattern \( W_i \) qualifies as a topic by our definition, then the signals in its coarse-grained timecourse [say, a vector \( b := (b_{i1}, \ldots, b_{i61}) \)] representing word counts in each chapter of *Pride and Prejudice* are not overwhelmed by Poisson noise. This vectorization scheme, together with the Růžička similarity (38, 39) \( s_R(b^j, b^k) \) between two vectors with non-negative entries, allow us to align some topics found in parallel versions of the same document, in languages A and B (Fig. 3D). (See fig. S12B for a control experiment on non-topics.)

Local graphs, kinetic semantics, machine translation

We have just shown that diagonal kinetic data \( (n_{ii} \) and \( L_{ii} \) statistics) can be used to extract topics numerically, and a small fraction of such topics have vector representations on the scale of chapters, which are stable against translations. Our next goal is to computationally construct semantic fields from off-diagonal kinetic data \( (n_{ij} \) and \( L_{ij} \) statistics), and vectorize topics without coarse-graining the text document, in a language-independent manner, valid on all the temporal scales above the discourse level.

To achieve this, we first need to specify a local, directed, and weighted graph (corresponding to a localized Markov transition matrix \( P^{(ij)} \)) for each topical pattern \( W_i \). To localize, we remove edges between two vertices \( W_i \) and \( W_j \), when their average temporal distance \( \langle \log L_{ij} \rangle \) exceeds a statistical threshold to be defined below.

If we have a Markov process on a semantic web that honors detailed balance, then for each fixed textual pattern \( W_i \), we can determine the distribution of hitting times \( L_{ij}, 1 \leq j \leq N \) from that of the return times \( L_{ii} \) (40). Further assuming that the distribution of \( \langle \log L_{ij} \rangle, 1 \leq j \leq N \) is nearly Gaussian, we can use the following approximation:

\[
P(\log L_{ij} > \ell) \approx \alpha_{ij}(\ell) := \sqrt{\frac{n_{ij}}{2\pi b^j}} \int_\ell^{\infty} e^{-\frac{(x - \ell)^2}{2b^j}} \, d\ell,
\]

where

\[
\ell^*_j := \frac{\langle L_{ii} \log L_{ii} \rangle}{\langle L_{ii} \rangle} - 1, \quad b^j_* := \frac{\langle (L_{ij} - \log L_{ii})^2 \rangle}{\langle L_{ii} \rangle}.
\]

Here, the estimate “\( \approx \)” in (6) becomes asymptotically exact if distinct textual patterns are statistically independent (such as \( \alpha_{13}, \alpha_{24}, \alpha_{31}, \alpha_{34} \) in Fig. 3A).
Empirically, we find that higher $\alpha_{ij}(t)$ scores point to closer affinities between textual patterns (Fig. 4A), attributable to kinship (Elizabeth, Jane), courtship (Darcy, Elizabeth), disposition (Darcy, pride) and so on. Therefore, even in a “poverty of stimulus” scenario, without references other than the novel itself, we can use the $\alpha_{ij}(t)$ scores as guides to numerical approximations of semantic fields, hereafter referred to as semantic cliques.

We invite a topical pattern $\mathbb{W}_j$ to the semantic clique $\mathcal{J}_i$ (insets of Figs. 4A and B) surrounding $\mathbb{W}_i$, if $\text{min}(\alpha_{ii}(\log L_{ij}), \alpha_{ij}(\log L_{ji})) > \alpha_s$ for a threshold value $\alpha_s := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-x^2/2} \, dx = 0.8413...$ that is associated with the standard Gaussian distribution. On a local graph with vertices $\mathcal{J}_i = \{\mathbb{W}_i, \mathbb{W}_1, ..., \mathbb{W}_n\}$, we specify the connectivity of each directed edge by a localized Markov matrix $\tilde{\mathbf{P}}^{[i]}(\mathcal{J}_i) = (\tilde{p}_{ij}^{[i]}(\mathcal{J}_i))_{1 \leq j,k \leq N_i}$, which is row-renormalized from an $N_i \times N_i$ subblock of $\tilde{\mathbf{P}}$ with the same set of vertices as $\mathcal{J}_i$. Resetting the entries $\tilde{p}_{ik}^{[i]}$ and $\tilde{p}_{jk}^{[i]}$ as zero, one arrives at the localized recurrence matrix $\mathbf{R}^{[i]}$, which characterizes the distribution for recurrence times to the Markov state $\mathbb{W}_i$ in $\mathcal{J}_i$.

Experimentally, we resolve the connectivity of an individual pattern $\mathbb{W}_i$ through the recurrence spectrum $\sigma(\mathbf{R}^{[i]})$ (Fig. 4B). The dominant eigenvalues of $\mathbf{R}^{[i]}$ are concept-specific while remaining nearly language-independent. Such empirical evidence motivates us to quantify the kinetic profile of a textual pattern $\mathbb{W}_j$ by a descending list for the magnitudes of eigenvalues $v_i = (|\lambda_i(\mathbf{R}^{[i]}(\mathcal{J}_i))|, |\lambda_2(\mathbf{R}^{[i]}(\mathcal{J}_i))|, ... )$. We zero-pad this vector from the $i$th component onwards, where $\eta_i = -\sum_{j \leq i} \|\mathbf{p}^{[i]}_j\| \log \|\mathbf{p}^{[i]}_j\|$ is the entropy production rate (41) of the Markov matrix $\tilde{\mathbf{P}}^{[i]}$, measured in nats per word.

Given a textual pattern $\mathbb{W}_i^A$ in language A, the aforementioned vector embeddings allow us to numerically locate a semantically close pattern in a parallel text (42) written in another language B, in two steps: (1) Divide the document into $K$ chapters, and define the semantic similarity function as $s(\mathbb{W}_i^A, \mathbb{W}_j^B) := s_k(\mathbf{v}_i^A, \mathbf{v}_j^B)$ if $s_k(\mathbf{b}_i^A, \mathbf{b}_j^B) \geq \max \{1 - 0.07 \sqrt{K}, 1 - \sqrt{\|\mathbf{b}_i^A \wedge \mathbf{b}_j^B\| / (\|\mathbf{b}_i^A\| \vee \|\mathbf{b}_j^B\|)}\}$ [a ballpark screening (43) more robust than Fig. 3D] and $s_k(\mathbf{v}_i^A, \mathbf{v}_j^B) \geq 0.7$; $s(\mathbb{W}_i^A, \mathbb{W}_j^B) := 0$ otherwise. (2) Solve a bipartite matching problem (Fig. 5) that maximizes $\sum_{i,j} s(\mathbb{W}_i^A, \mathbb{W}_j^B)$, using the Hungarian Method (44) attributed to Jacobi–König–Egerváry–Kuhn (45).

Semantic clique, associative reasoning, query expansion
So far, our numerically constructed semantic cliques $\mathcal{J}_i$ incorporate (Figs. 4A and B)

![Figure 5: Semantic similarities $s(\mathbb{W}_i^A, \mathbb{W}_j^B)$ between selected topics (sorted in descending $n_i \geq 20$) in two versions of Pride and Prejudice. Rows and columns filled with zeros are not shown. Cross-hairs meet at optimal nodes that solve the bipartite matching problem (see Figs. S8–S20 for more examples). The thickness of each horizontal (resp. vertical) cross-hair is inversely proportional to the row-wise (resp. column-wise) ranking of the similarity score for the optimal node. Green (resp. amber) cross-hair indicates an exact (resp. a close but non-exact) match. At the same confidence level (0.7) for similarities, this experiment has better recall than Fig. 3D, without much cost of precision.](image-url)
Three-phase electric power is a common method of alternating current electric power generation, transmission, and distribution. A three-phase transformer with four wire output for 208Y/120 volt service: one wire for neutral; others for A, B, and C phases. A three-phase system is usually more economical than an equivalent single-phase or two-phase system at the same voltage because it uses less conductor material to transmit electrical power.

Figure 6: a A construction of semantic clique $\mathcal{D} \cup \mathcal{D}'$ weighted by $\pi$ and subsequent query expansion. Top 5 candidate answers shown with font sizes proportional to the entropy production score in (8). Here, the top answer (with highlighted background) is officially labeled as correct by the WikiQA team.

The three-phase system is usually more economical than an equivalent single-phase or two-phase system at the same voltage because it uses less conductor material to transmit electrical power.

Processes or knowledge feed.

| A | WikiQA-Q2955: What is three-phase electrical? |
|---|---|
| B | MAP | MRR |
| CNN | LCLR | EZ | LCLR | EZ | LCLR |
| 0.6040 | 0.6260 | 0.5670 | 0.5820 | 0.6090 | 0.6090 |
| 0.4400 | 0.4024 | 0.3720 | 0.3590 | 0.3970 | 0.3970 |

**Discussion and outlook**

In our current work, we develop a stochastic model that assigns meaning (semantic content) to words, via association and connectivity. Consistently using a single Markov framework, we are able to recover word frequencies (Fig. 2A), extract topics (Fig. 3B), interpret recurrence statistics (Figs. 2A and C), and ultimately demonstrate translatability (Figs. 2B, 3D, 4B, 5) as well as comprehensibility (Figs. 3A, 6A) through statistical learning of languages. Hopefully, our Markov model at the discourse level offers an approximate explanation for the universal hardware part in our language faculties, which are mechanisms common to all human minds. It is not our purpose to efface the diversity of the software aspects of human languages: the phonological encoding scheme adapts to geographical and climate conditions of native speakers (50–52), while lexical and syntactical features are highly susceptible to social, cognitive and cultural factors (53–56).

By analogy to the behavior of internet surfing (46, 47), we model the process of associative reasoning (48) as a navigation through the nodes $\mathcal{D} \cup \mathcal{D}'$ according to $\mathbf{P}$, which quantifies the click-through rate from one idea to another. The PageRank recursion (47) ensures a unique equilibrium state $\pi$ attached to $\mathbf{P}$. If our question $Q$ and a candidate answer $A$ contain, respectively, words from textual patterns $W_{Q_1}, \ldots, W_{Q_m}$ and $W_{A_1}, \ldots, W_{A_n}$, frequencies (counting multiplicities, but excluding function words and patterns with fewer than 3 occurrences in the reference document), then we assign the following entropy production score

$$\mathcal{F}[Q, A] := -\sum_{i=1}^{m} \sum_{j=1}^{n} p_{Q_i} p_{Q_i A_j} \log p_{Q_i A_j}$$


to this question-answer pair.

A sample work flow is shown in Fig. 6A, to illustrate how our machine answers a question, using a Wikipedia page (without infoboxes and other structural data) as its only reference document and training source. It can be seen that our numerical associative reasoning generates a weighted set of nodes $\mathcal{D} \cup \mathcal{D}'$ (presented graphically as a thought bubble in Fig. 6A) that is comparable to the output from a typical human reader of Wikipedia, without the help of external stimuli or knowledge feed.

We test our model (EZ in Fig. 6B) on all the 1242 questions in the WikiQA data set, each of which is accompanied by at least one correct answer located in a designated Wikipedia page. Our algorithm’s performance is roughly on par with LCLR and CNN benchmarks (49), improving upon the baseline by significant margin. This is perhaps remarkable, considering the relatively scant data at our disposal: unlike the LCLR approach, our numerical discovery of synonyms does not draw on the WordNet database or pre-existent corpora of question-answer pairs; unlike the CNN method, we do not need pre-trained word2vec embeddings (18) as semantic input.

Moreover, our algorithm (EZ* in Fig. 6B) performs slightly better on a subset of 990 questions that do not require quantitative cues (How large? How long? How many? How old? What became of? What happened to? What year? and so on). This indicates that our structural model fits associative reasoning better than rule-based reasoning (48), while imitating human behavior in the presence of limited data.

- Synonyms (pride and vanity in English, orgueil and fierté in French, etc.);
- Temperaments (Elizabeth, a delightful girl, often laughs, corresponding to French verbs sourire and rire);
- Co-references (e.g. Darcy as a personification of pride);
- Causality (such as pride based on fortune),

as well as other kinds of relations typically shared by words within a semantic field.

Now, we can exploit the connectivity of these semantic cliques $\mathcal{S}_i$ to build a rudimentary question-answering machine.

Concretely speaking, we are given a document (of moderate length) and a natural language question as input, and want to rate and rank the sentences within the document by their relevance to the question, containing topical patterns $\mathcal{D} = \{W_{Q_1}, \ldots, W_{Q_k}\}$. We expand the query into $\mathcal{D} \cup \mathcal{D}'$, a union of semantic cliques: $\mathcal{D} \cup \mathcal{D}' = \bigcup_{k=1}^{K} \mathcal{S}_k$. As before, we can construct a localized Markov matrix $\mathbf{P} = (p_{ij})_{1 \times N}$ on this subset of textual patterns $\mathcal{D} \cup \mathcal{D}'$. We further use the Brin–Page damping (46) to derive an ergodic Markov matrix $\tilde{\mathbf{P}} = (\tilde{p}_{ij})_{1 \times N}$, where $\tilde{p}_{ij} = 0.85 p_{ij} + 0.15 / N$.

By analogy to the behavior of internet surfing (46, 47), we model the process of associative reasoning (48) as a navigation through the nodes $\mathcal{D} \cup \mathcal{D}'$ according to $\mathbf{P}$, which quantifies the click-through rate from one idea to another. The PageRank recursion (47) ensures a unique equilibrium state $\pi$ attached to $\mathbf{P}$. If our question $Q$ and a candidate answer $A$ contain, respectively, words from textual patterns $W_{Q_1}, \ldots, W_{Q_m}$ and $W_{A_1}, \ldots, W_{A_n}$, frequencies (counting multiplicities, but excluding function words and patterns with fewer than 3 occurrences in the reference document), then we assign the following entropy production score

$$\mathcal{F}[Q, A] := -\sum_{i=1}^{m} \sum_{j=1}^{n} p_{Q_i} p_{Q_i A_j} \log p_{Q_i A_j}$$


to this question-answer pair.
tional complexity. On the practical side, our Markov model achieves some reasonable results in word translation and text comprehension. This suggests that we are probably close to a complete set of semantic invariants, after demystifying the long-range behavior of human languages. In the near future, we hope to extend our framework further, to incorporate both Markovian and non-Markovian structures across all the four Friederici neurobiological hierarchies—The Mathematical Principles of Natural Languages, as we envision, must and will combine the statistical analysis of a Markov model with neurolinguistic properties on shorter time scales (22).

References and Notes

1. A. Vouloumanos, M. D. Hauser, J. F. Werker, A. Martin, Child Development 81, 517 (2010).
2. A. R. Pfennig, et al., Science 346, 1256846 (2014).
3. R. C. Berwick, N. Chomsky, Why Only Us: Language and Evolution (MIT Press, Cambridge, MA, 2016).
4. C. S. L. Lai, S. E. Fisher, J. A. Hurst, F. Vargha-Khadem, A. P. Monaco, Nature 413, 519 (2001).
5. W. Enard, et al., Nature 418, 869 (2002).
6. R. Nudel, D. F. Newbury, Wiley Interdisciplinary Reviews: Cognitive Science 4, 547 (2013).
7. J. G. Herder, Abhandlung über den Ursprung der Sprache (Christian Friedrich Voß, Berlin, Prussia (now Germany), 1772).
8. J.-J. Rousseau, Essai sur l’origine des langues (Du Peyrou, Geneva, Switzerland, 1781).
9. F. de Saussure, Cours de linguistique générale (Payot, Paris, France, 1949), fifth edn.
10. P. T. Daniels, W. Bright, eds., The World’s Writing Systems (Oxford University Press, New York, NY, 1996).
11. N. Chomsky, The Behavioral and Brain Sciences 3, 1 (1980).
12. S. Pinker, The Language Instinct: How the Mind Creates Language (William Morrow and Company, New York, NY, 1994).
13. A. D. Friederici, Language Comprehension: A Biological Perspective, A. D. Friederici, ed. (Springer, Berlin, Germany, 1999), chap. 9, pp. 265–304.
14. M. D. Hauser, N. Chomsky, W. T. Fitch, Science 298, 1569 (2002).
15. N. Chomsky, Rules and representations (Columbia University Press, New York, 2005).
16. D. M. Blei, A. Y. Ng, M. I. Jordan, J. Mach. Learn. Res. 3, 993 (2003).
17. Y. Wilks, C. Brewster, Foundations and Trends in Web Science 1, 199 (2007).
18. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Advances in Neural Information Processing Systems 26 (2013), pp. 3111–3119.
19. A. Karpathy, J. Johnson, L. Fei-Fei, ICLR 2016 Workshop (2016).
20. S. Arora, Y. Li, Y. Liang, T. Ma, A. Risteski, Transactions of the Association for Computational Linguistics 4, 385 (2016).
21. S. Arora, R. Ge, F. Koehler, T. Ma, A. Moitra, PMLR 48, 2859 (2016).
22. S. Pinker, N. Chomsky, Q&A in Keynote Panel: The Golden Age — a Look at the Original Roots of Artificial Intelligence, Cognitive Science, and Neuroscience (2011). http://languageology.ldc.upenn.edu/myl/PinkerChomskyMIT.html
https://www.youtube.com/watch?v=1DkDme8nUUU
23. The most undesirable short-range “noise source” is reproduction of textual patterns. Typically, such behavior is not reproducible after translation into a different language. See Supplementary Materials for concrete examples.
24. R. Jakobson, Recherches structurales 1949. Interventions dans le débat glossématique (Nordisk Sprog- og Kulturforlag, Copenhagen, Denmark, 1949), vol. V of Travaux du Cercle Linguistique de Copenhague, pp. 205–213.
25. See Supplementary Materials for detailed linguistic heuristics, control experiments, and mathematical proofs pertaining to our model.
26. F. Jelinek, Proceedings of the IEEE 64, 532 (1976).
27. H. J. Sommers, A. Crisanti, H. Sompolinsky, Y. Stein, Phys. Rev. Lett. 60, 1895 (1988).
28. T. Tao, V. Vu, Ann. Probab. 38, 2023 (2010). With an appendix by Manjunath Krishnapur.
29. See Supplementary Methods for protocols of text data cleansing, word morphology clustering, as well as detailed criteria for content words, tailored for 12 European languages and 2 Asian languages exhibiting diverse typologies.
30. We exclude function words like a, the, of, whose rôles are confined to short time scales of the neurobiological hierarchy (13), as demonstrated by clinical investigations of Broca’s aphasia (58). Furthermore, recent fMRI–dMRI studies (59–61) of the human brain support the existence of separate pathways for syntactic and semantic networks (62). This justifies the removal of short-range language features in our Markov model for universal semantic content.
31. In the present work, we are only interested in mental activities that are responsible for semantic processing in the ventral stream, as opposed to syntactic computations in the dorsal tract.

32. N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, Cambridge, MA, 1965).

33. J.-B. Michel, *et al.*, *Science* **331**, 176 (2011).

34. Y. Zhou, X. Zhuang, *Biophys. J.* **91**, 4045 (2006).

35. H. P. Luhn, *IBM Journal of Research and Development* **1**, 309 (1957).

36. K. Spärck Jones, *Journal of Documentation* **28**, 11 (1972).

37. D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (John Wiley and Sons, New York, 1967).

38. M. Ružička, *Biológia (Bratislava)* **13**, 647 (1958).

39. E. Deza, M.-M. Deza, *Dictionary of distances* (Elsevier, Amsterdam, The Netherlands, 2006).

40. N. Haydn, Y. Lacroix, S. Vaienti, *Ann. Probab.* **33**, 2043 (2005).

41. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley Interscience, Hoboken, NJ, 2006), second edn.

42. It is worth mentioning that, Gale and Church have exploited the recurrence time of punctuation marks (i.e. sentence lengths) to align sentences in bilingual corpora. One may thus regard our kinetic machine translation as a type of Gale–Church alignment algorithm that operates at word level.

43. Here, ∧ (resp. ∨) denotes component-wise minimum (resp. maximum) of vectors; ∥b∥₀ counts the number of non-zero components in b; ∥b∥₁ sums over all the components in b.

44. H. W. Kuhn, *Naval Research Logistics Quarterly* **2**, 83 (1955).

45. H. W. Kuhn, *European Journal of Operational Research* **219**, 641 (2012).

46. S. Brin, L. Page, *Computer Networks and ISDN Systems* **30**, 107 (1998).

47. L. Page, S. Brin, R. Motwani, T. Winograd, *The PageRank citation ranking: Bringing order to the web*, Tech. rep., Stanford InfoLab (1999). [http://ilpubs.stanford.edu:8090/422/](http://ilpubs.stanford.edu:8090/422/)

48. S. A. Sloman, *Psychological Bulletin* **119**, 3 (1996).

49. Y. Yang, W.-t. Yih, C. Meek, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Lisbon, Portugal, 2015).

50. C. Everett, D. E. Blasi, S. G. Roberts, *Proc. Natl. Acad. Sci. USA* **112**, 1322 (2015).

51. C. R. Ember, *Journal of Language Evolution* **1**, 60 (2016).

52. C. Everett, *Frontiers in Psychology* **8**, Article 1285 (2017).

53. J. G. Frazer, *The Golden Bough: a study of magic and religion* (MacMillan, New York, NY, 1922), abridged edn.

54. W. Labov, *Principles of Linguistic Change: Social Factors* (Blackwell Publishers, Malden, MA, 2001).

55. K. Allan, K. Burridge, *Forbidden Words: Taboo and the Censoring of Language* (Cambridge University Press, Cambridge, UK, 2006).

56. W. Labov, *Principles of Linguistic Change: Cognitive and Cultural Factors* (Wiley-Blackwell, Malden, MA, 2010).

57. N. Chomsky, *Psychotherapy and Psychosomatics* **74**, 263 (2005).

58. H. Gardner, *The Shattered Mind: the Person after Brain Damage* (Knopf Distributed by Random House, New York, 1975).

59. S. M. Wilson, *et al.*, *Neuron* **72**, 397 (2011).

60. J. D. Griffiths, W. D. Marslen-Wilson, E. A. Stamatakis, L. K. Tyler, *Cerebral Cortex* **23**, 139 (2012).

61. A. D. Friederici, J. Bahlmann, S. Heim, R. I. Schubotz, A. Anwander, *Proc. Natl. Acad. Sci. USA* **103**, 2458 (2006).

62. A. D. Friederici, *Neurobiology of Language*, G. Hickok, S. L. Small, eds. (Elsevier, Amsterdam, The Netherlands, 2016), chap. 29, pp. 349–356.

63. W. A. Gale, K. W. Church, *Computational Linguistics* **19**, 75 (1993).

**Acknowledgments**

We thank X. Wan, R. Yan and D. Zhao for their suggestions on experimental design, during the early stages of this work. We thank X. Sun for fruitful discussions on neural networks.

**Author contributions:** W. E, Y. Z. designed and conducted the research, and wrote the paper.

**Data and materials availability:** Source codes (in Mathematica) are available upon request. Provenances of multilingual texts are described in Supplementary Materials.
Supplementary Materials

Materials and Methods
Figs. S1–S20
Tables S1–S5
Schemes S1–S2
References 64–128