Supporting Information

Ramiro et al. 10.1073/pnas.1714730115

Historical Analyses of Word Form Reuse Versus Innovation

To investigate the relative frequencies of cases where new meanings are labeled with existing word forms (i.e., resulting in polysemy) and cases where novel word forms are created (i.e., resulting in new lexemes), we conducted a set of analyses using data from the HTE described in the main text. Our analyses quantified the proportion of cases over the history of English in which new meanings were lexicalized using existing word forms vs. morphological derivations of existing word forms vs. new lexemes.

We analyzed lexeme and meaning records over the past 1,000 y (1001–2000) in English (the HTE does not provide dates for word and meaning records earlier than the Old English period, i.e., before 1000). For each new sense entering the dictionary at a given time point, we classified how it was lexicalized into one of three categories. We considered a historical case polysemous if the novel sense was labeled by an existing word form, that is, if a word is reused exactly, with no change in the existing form. We considered the case innovative if the novel sense was labeled by a new lexeme that was not present in the existing lexicon before that time. To take into account cases of morphological derivation in English, where words share a common root via suffixing (e.g., cost $\rightarrow$ costly), we considered these as derived but not exact reuses. Other forms of morphological derivation are possible in English, e.g., prefixing, but we chose to focus on suffixing due to the availability of automatic stemming procedures that detect cases of suffixing. Our analyses thus provided a conservative estimate of cases of morphological derivation, such that some of the cases we classified as "innovation" likely involved some forms of derivation (e.g., prefixing) or compounding (e.g., "smartwatch").

At each year where novel senses were recorded, we counted the cases of reused (exact and derived) and new words. To identify cases of derived reuse automatically, we first used the Natural Language Toolkit Python package of Snowball stemmer (www.nltk.org/_modules/nltk/stem/snowball.html) to stem the words. We then treated a case as derived if a novel sense was labeled by a word that shares a root with the existing words in the lexicon. We also excluded all phrase records (i.e., compound words) in the thesaurus that include a space.

We summarize the total counts of the cases we considered in Fig. S1, *Left*, collapsing data across the 1,000-y period we analyzed. The result shows that cases of reuse, regardless of whether they are exact only or include derived uses, are substantially more prevalent than the cases of innovation (new). In particular, we obtained 425,817 total counts of reuse, 345,285 of which are strict polysemy (exact reuse) and 80,532 of which are derived reuses. In comparison, we obtained half as many counts (193,444) in the case of innovation. The difference between exact reuse and new cases was highly significant by a binomial test ($p < 0.0001$), suggesting that polysemy dominates word form innovation in the history of English.

To evaluate whether this trend is persistent over the course of history, we compared the count of reused cases to that of new cases across all available years in the period we considered. This difference should be positive if polysemy dominates innovation as a strategy of labeling emerging meanings for a given year. Fig. S1 shows this is the case, where the values of difference are dominantly positive, with 669 times in which cases of reused (exact) exceed those of innovation (new), 129 times in which the opposite is true, and 205 cases of tie (almost all tied cases occurred in the first 200-y period, where sense records are relatively sparse). Together, these analyses suggest that polysemy has been a dominant strategy throughout the history of English.

Example Calculation of Conceptual Proximity

We provide an example calculation of the conceptual proximity measure we used for the analyses described in the main text. Table S1 shows two senses sampled from the sense records of the word game (top panel) and the taxonomic representation of these senses (bottom panel), following the hierarchical definitions provided in the HTE described in the main text:

Because the two senses share two parent tiers in the taxonomic hierarchy (i.e., The social world $\rightarrow$ Leisure), their conceptual proximity is calculated as follows:

$$c(\bullet, \star) = \frac{2 \times |parent|}{l(\bullet) + l(\star)} = \frac{2 \times 2}{5 + 6} = \frac{4}{11}. \quad \textbf{[S1]}$$

Illustration of Verb Taxonomy

Verbs are classified under the same taxonomy as nouns in the HTE. Table S2 illustrates an example pair of senses for the verb "play."

Nearest-Neighbor Chaining and Minimal Spanning Tree

We demonstrate that the nearest-neighbor chaining model we proposed in the main text approximates the process of constructing a minimal spanning tree over time. Specifically, we show that it resembles Prim's classic algorithm that yields a minimal spanning tree, with a certain probability. We then present empirical evidence that this probability is substantially higher than other competing models we have considered.

Theoretical Connection. We first show that the nearest-neighbor chaining model is closely related to Prim's greedy algorithm of constructing a minimal spanning tree. Assuming a graph with vertices $v \in V$, edges e, and edge weights (or costs) c, Prim's algorithm runs as follows:

Algorithm 1. Prim's algorithm that gives a minimal spanning tree with lowest edge costs.

Initialize: Choose an arbitrary vertex $v \in V$; let $S = \{v\}$ and T be an empty set
while $S \neq V$ **do**
  Select an edge e such that only one of its endpoints is in S and c is minimal;
  Add e to T;
  Add endpoints of e to S;
end
Return: T

The nearest-neighbor chaining model we proposed is a probabilistic version of the Prim's algorithm with a fixed initial vertex s_0 that corresponds to the initial sense of a word. Assuming all possible senses of a word as vertices of a graph $s \in S$ (e.g., in a taxonomy-based similarity space or a Euclidean space), all pairs of these senses forming edges e, and costs of sense extension as distance between an existing sense and a novel sense d, the nearest-neighbor chaining algorithm constructs a tree probabilistically that favors the minimally distant sense to chain to at each step:

Algorithm 2. Nearest-neighbor chaining algorithm that approximates the Prim's algorithm.

Initialize: Choose initial sense $s_0 \in S$; let $C = \{s_0\}$ and T be an empty set
while $C \neq S$ **do**
  Sample a sense-sense edge e such that only one of its endpoints is in C and d is minimal, with probability proportional to $\exp(-d)$;
  Add e or its alternative candidate (with probability no greater than that of e) to T;
  Add endpoints of e (that includes the new sampled sense) to S;
end
Return: T

In the case where the nearest-neighbor chaining model maximizes the probability of choice at each step, the algorithm converges to Prim's algorithm except for a fixed starting point, therefore yielding a globally minimally distant sense network over time:

Algorithm 3. Nearest-neighbor chaining algorithm that maximizes probability at each step.

Initialize: Choose initial sense $s_0 \in S$; let $C = \{s_0\}$ and T be an empty set
while $C \neq S$ **do**
  Select an edge e such that only one of its endpoints is in C and d is minimal;
  Add e to T;
  Add endpoints of e (that includes the new sense with the globally minimal d) to S;
end
Return: T

Because this special case only occurs with a certain probability (e.g., see case study in *Model Cost*), it follows that the nearest-neighbor model is guaranteed to produce a minimal spanning tree with probability no less than p, where p is the probability that the algorithm goes with the best choice at each step, among all other possible tree structures. Given this theoretical connection, we next validate empirically whether the nearest-neighbor chaining model indeed dominates other competing candidate models in producing low-cost sense extensional paths.

Empirical Validation. We provide a comprehensive simulation to compare models in terms of cost, complementing the illustrative simulation that we described in the main text.

We aimed to investigate whether the nearest-neighbor chaining model produces the lowest empirical cost empirically, considering four variable parameters in a simulation (which we did not vary systematically in the simulation described in the main text): (*i*) number of hypothetical senses of a word; (*ii*) typology or relative positions of the senses in 2D space; (*iii*) initial sense, or seeding position of the model; and (*iv*) randomness due to the probabilistic nature of each algorithm.

To account for these factors, we performed the simulation just as we introduced in the main text by further (step *i*) varying the number of senses in a 2D Euclidean space, from 5 to 30 in increments of 5; (step *ii*) generating 10 different randomized typologies by sampling senses (or points) in a $[0 - 1{,}0 - 1]$ grid, for each level we varied in step *i*; (step *iii*) iterating through every available sense and treating it as the initial position (and hence exhausting the possible starting positions for each model); and (step *iv*) probabilistically sampling according to Luce's choice rule (which we have defined separately for each model) at each step, in 20 different iterations.

Fig. S2 shows the tallies for which model has yielded the minimal-cost probabilistic path at all varying conditions. We observed that, in almost all cases, the nearest-neighbor chaining model was dominant in producing the lowest-cost sense network over time, which provides strong empirical support for our theoretical proposal that the algorithm constructs a near–minimal-cost sense network.

Model Cost and Likelihood

We illustrate how to calculate cost and likelihood of the models we have proposed by a simple case study. We show how cost and likelihood are dissociable, such that a model that yields a higher likelihood does not imply that it is necessarily lower in cost. Their correspondence in the data is an empirical result, not an a priori outcome.

Fig. S3 shows the semantic space that we used for this case study, similar to that in *Computational Formulation of Theory*. Despite the simplicity of this setting, the principles we demonstrate here should generalize to more complex scenarios. Specifically, we have constructed a hypothetical word that includes four senses, labeled A to D, following their temporal orders of emergence (also indicated in the figure). We took Euclidean distance $d(\cdot,\cdot)$ between any pair of senses as a proxy of semantic relatedness, where similarity between two senses is $sim(\cdot,\cdot) = e^{-d(\cdot,\cdot)}$ as described in the main text. In this specific case, sense pairs AB and BD bear equal distances of 1—slightly shorter than distance between senses A and C, which is 1.1. We will see later how such a construction yields different temporal predictions from the progenitor model and the nearest-neighbor chaining model, which we focus on comparing for the purpose of illustration.

Space of Possible Paths. Given the configuration described in Fig. S3, there will be $3! = 3 \times 2 \times 1 = 6$ possible historical orders in which senses B, C, and D could have been derived over time: ABCD, ABDC, ACBD, ACDB, ADBC, and ADCB. In general, for a word with n senses, the number of possible historical orders is $(n-1)!$. However, it is important to note that each model specifies a potentially different way in which senses may be derived over time, given a specific historical order of emergence.

In theory, each model specifies a full probability distribution over all possible orders of sense emergence for a given word. Such a distribution reflects the degree to which a model favors one possible extensional mechanism over other alternatives in accounting for the historical order of senses. For example, given a single emerging order of ABCD, a nearest-neighbor chaining model might assign a relatively high probability to the extensional path A→B, B→C, C→D (e.g., a chain-based mechanism of extension), as opposed to A→B, A→C, A→D (e.g., a radially structured mechanism of extension), which may, in turn, be assigned with a high probability by the progenitor model. The models we have proposed effectively explore the space of possible ways in which new senses could have been derived, even though there exists only one true historical order via which they have actually emerged. This fact allows us to distinguish between the cost of a model (i.e., a cognitive metric that defines how efficient or cost-effective models are), which depends on the aggregated semantic distance that a specific algorithm traverses over time, and the likelihood of a model (i.e., a statistical metric for evaluating models against truth), which depends on the true historical order of sense emergence.

Model Cost. Because the space of possible sense emerging paths scales rapidly with number of senses (e.g., for 13 senses, $(n-1)! = 12!$ yields 479,001,600 possibilities), it is intractable to estimate probabilities and costs exhaustively. In this constrained case study, however, we can perform exhaustive calculations.

We first show the calculation of cost with a single-path prediction made by a model, which chooses the candidate sense with maximal probability at each step. We then generalize and show the cost of each possible path.

Fig. S4 illustrates the predicted paths and associated probabilities from two models, progenitor and nearest-neighbor chaining, in the hypothetical space we described. At each step, each model infers which sense is likely to be the next emerging sense, extended from existing senses of that word. Specifically, the progenitor model infers an emerging sense with probability in proportion to its similarity to the earliest progenitor sense A (i.e., marked in red in the figure). For example, initially in time, it assigns a probability to each of the candidate senses B, C, and D, with B carrying the most probability weight because it is the closest to A among other candidates. The nearest-neighbor chaining model makes similar predictions and yields the same probability weighting in this step, because there is only one existing sense to chain from. However, these two models differ in their predicted path in the next step(s): The progenitor model infers with greater probability that C should emerge next, because C is closer to A than D; the nearest-neighbor chaining model instead infers with greater probability that D should emerge next, because D is the closest candidate to existing sense B—just as C is to existing sense A, but distance DB is shorter than distance AC by construction (and hence chain DB is considered semantically more related than chain AC).

We can calculate cost of these models based on their predicted paths, assuming they maximize probability at each step. Specifically, we define cost as the aggregated sum of distances in the extension path. In this case, the nearest-neighbor chaining model yields a lower cost (as expected),

$$c_{nnchaining} = d(A,B) + d(B,D) + d(A,C) = 1 + 1 + 1.1 = 3.10, \quad \textbf{[S2]}$$

in comparison with the progenitor model,

$$c_{progenitor} = d(A,B) + d(A,C) + d(A,D) = 1 + 1.1 + 1.85 = 3.95. \quad \textbf{[S3]}$$

We can compute the probabilities of these paths based on Luce's choice rule (we abbreviate similarity $sim(\cdot,\cdot)$ as $s\cdot\cdot$ in the following),

$$p_{nnchaining} \propto p(B|A) \times p(D|A,B) \times p(C|A,B,D) \quad \textbf{[S4]}$$

$$\propto \frac{sAB}{sAB + sAC + sAD} \times \frac{sBD}{sAC + sBD} \times \frac{sAC}{sAC} \quad \textbf{[S5]}$$

$$\propto 0.428... \times 0.524... \times 1 = 0.23, \quad \textbf{[S6]}$$

$$p_{progenitor} \propto p(B|A) \times p(C|A,B) \times p(D|A,B,C) \quad \textbf{[S7]}$$

$$\propto \frac{sAB}{sAB + sAC + sAD} \times \frac{sAC}{sAC + sAD} \times \frac{sAD}{sAD} \quad \textbf{[S8]}$$

$$\propto 0.428... \times 0.679... \times 1 = 0.29. \quad \textbf{[S9]}$$

Similarly, we can also compute cost and probability of each of the six possible paths under these two models separately, as summarized in Table S3.

In this specific case, we can compute the average cost under each model m,

$$E[cost_m] = \sum p(path_m)c(path_m), \quad \textbf{[S10]}$$

which yields 3.36 for the chaining model and 3.95 for the progenitor model. In the main text, we estimate cost under a single prediction from each model that maximizes stepwise probability, which yields 3.10 for the chaining model and 3.95 for the progenitor model. Regardless whether we consider the average or the single-shot cost, the chaining model yields a lower overall cost.

Model Likelihood. To evaluate the models, we estimate model likelihood as the log probability at which each model would predict the true historical order of emerging senses, i.e., $\mathcal{L} = p(path_{true}) = p(A = t0) \times p(B = t1|A = t0) \times p(C = t2|B = t1, A = t0) \times p(D = t3|C = t2, B = t1, A = t0)$ for the case scenario we described. This metric determines the likelihood under which senses would emerge as observed, based on the specific extensional mechanism postulated by a given model. Note that the value of this metric is determined by how probable a model considers the true order of sense emergence, not how cost-effective a model is in extending senses from one another.

Table S4 shows the stepwise calculations of likelihood for both models, given the true emerging order of senses: ABCD. In this case, the progenitor model yielded a higher overall likelihood, because it assigned the highest probability to the true emerging sense at each step. However, this is not so for the nearest-neighbor chaining model, because it assigned more probability to D (i.e., false candidate) as opposed to C (i.e., true candidate) when C emerged at $t2$. The likelihood of the random baseline model is (i.e., $1/6 = 1/3 \times 1/2 \times 1$), and it is lower than both models examined here.

Table S5 summarizes the cost and likelihood for both models. These results suggest how cost and likelihood of models may be dissociated, such that a cost-optimized model may not always yield the optimal likelihood in predicting temporal emergence of word senses.

Model Comparison Controlling for Age of Words

We summarize the results from additional model comparisons, based on the second word set that controls for age of word described in the main text.

Fig. S5 summarizes the mean log likelihood ratios and the winner-take-all results for all models on this word set. Consistent with the BNC word set, the nearest-neighbor chaining model yields the best performance in these tests. We assessed the significance of this result by performing paired t tests between the chaining model and each of the competitors ($p < 0.001$ from all tests ($n = 2{,}648$) with Bonferroni correction for multiple tests: against exemplar ($t = 17.0$), prototype ($t = 17.7$), progenitor ($t = 20.2$), and local ($t = 11.9$)).

Analyses of Conditions That Favor Chaining

We describe the measure we used to explore the relative superiority of the chaining model in the main text, along with example words given this measure.

We defined a superiority score of chaining S_{nn} by the extent to which the nearest-neighbor chaining model outperforms other competing models we have considered, in terms of their ability to predict the historical order of emergence of a word's senses. Specifically, we took the expected value of difference in log likelihoods between the chaining model and each of the remaining four models,

$$S_{nn} = E[\mathcal{L}_{nn} - \mathcal{L}_m], \text{where } m \in \{\text{exemplar, prototype, progenitor, local}\}. \quad \textbf{[S11]}$$

Intuitively, this score determines the degree to which the nearest-neighbor chaining model predicts historical ordering of a word's senses better than each of the alternative models. Fig. S6 visualizes the distribution of this score for all words in the BNC set, along with some example words. We observed that this distribution is strongly positively skewed (skewness = 2.56), indicating that nearest-neighbor chaining generally outperforms other models in predicting the orders of sense emergence for most words.

Table S6 further summarizes information regarding words that fall in the left and right tails of this distribution. As the table indicates, the words with the highest superiority scores of chaining also tended to be highly polysemous words, carrying many different senses. Importantly, the relative superiority of the chaining model is also high for words such as over and game (and face, which we used to illustrate chaining in the main text), which are paradigm examples highlighted by Lakoff (6) and Wittgenstein (3). In contrast, words with fewer senses tended to have lower scores, such that the historical order of development of their senses was better explained by other models such as the prototype or local models. As noted in the main text, it is interesting that the chaining model performed better on words that have developed more senses over history, since it is these words that could have the most costly sense extensional paths. This finding supports our proposal that nearest-neighbor chaining has been a preferred mechanism for minimizing cost in historical word sense extension.

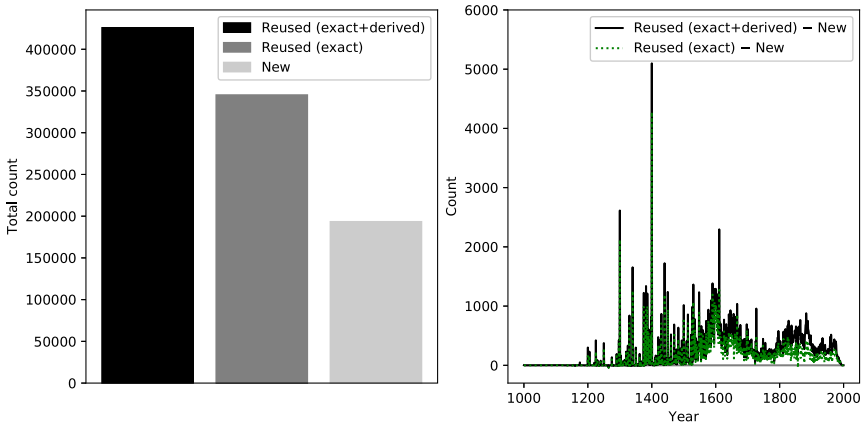

Fig. S1. Summary of historical analyses on reuse and innovation of word forms. Reused cases are tabulated for exact cases (black bar) and exact+derived, i.e., including those with root-sharing via suffixing (gray bar); new cases correspond to those that do not fall under either of the reused categories. (*Left*) The overall counts across the past 1,000 y of English. (*Right*) The differences in counts over the course of history.

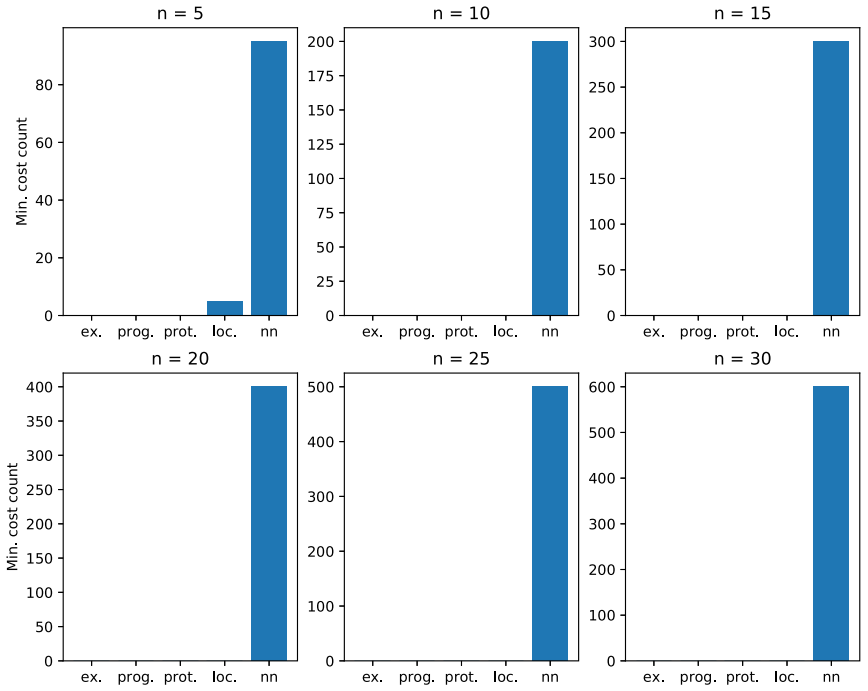

Fig. S2. Comparison of model-generated costs in the simulation. Number of senses is increased from 5 to 30 in incremental steps of 5 in the simulation. Vertical axis indicates the number of runs at which a model produced the lowest overall cost among all competing models. The abbreviations on the horizontal axis (from left to right) correspond to exemplar, progenitor, prototype, local, and nearest-neighbor chaining models.

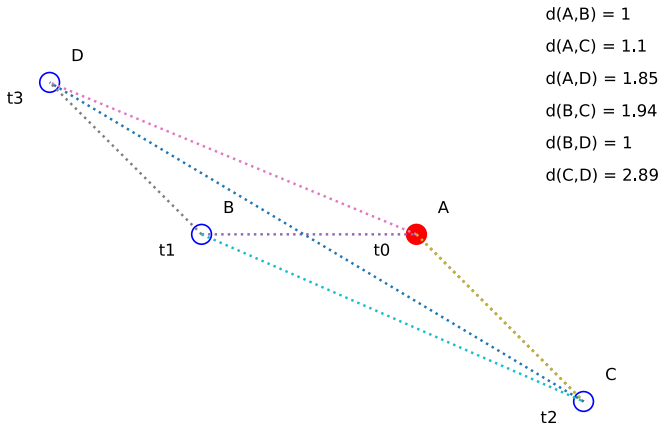

Fig. S3. Semantic space of a hypothetical word. Dots marked with A, B, C, and D represent emerging senses for that word. The red dot represents the earliest sense. The true emerging order for this word is as follows: A (appearing at time t0), B (appearing at t1), C (appearing at t2), and D (appearing at t3). Distance between senses $d(\cdot, \cdot)$ is indicated by a dotted line, based on Euclidean distance.

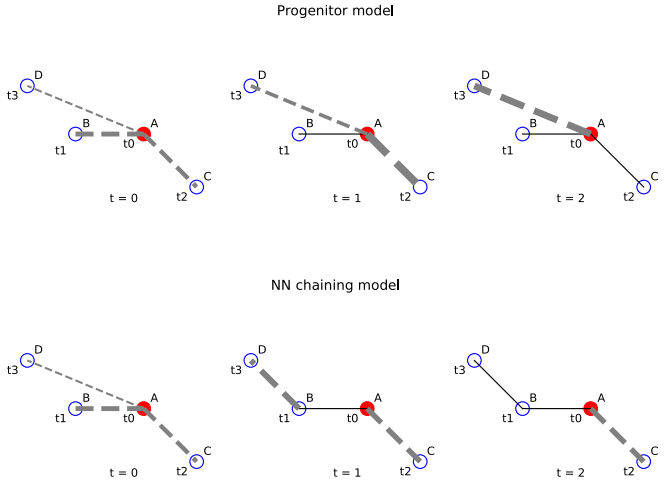

Fig. S4. Illustration of sense extensional paths for the hypothetical word as predicted by the progenitor model (*Top*) and the nearest-neighbor chaining model (*Bottom*). Gray dashed lines indicate possible choices of extension given an existing sense. Width of dashed lines is proportional to the probability of any candidate sense being chosen at a given time point, based on Luce's choice rule. Black solid lines indicate chosen paths. Cost of a model corresponds to the total distance traversed in sense extension, or the sum of graph edges at t2.

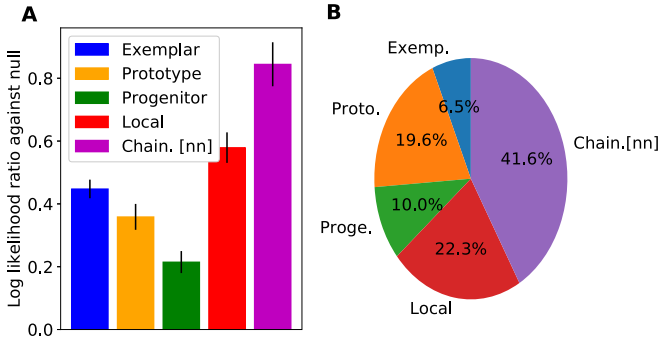

Fig. S5. Summary of model performances on the word set that controls for word age. (*A*) Likelihood ratio test; 0.0 on the *y* axis indicates performance of the null model. Bar height indicates the mean log likelihood ratio averaged over the pool of most common words from the BNC corpus. Error bars indicate 95% confidence intervals. (*B*) Visualization of winner-take-all percentage breakdown among the proposed models from the same test. "Chain. [nn]" refers to the nearest-neighbor chaining model.

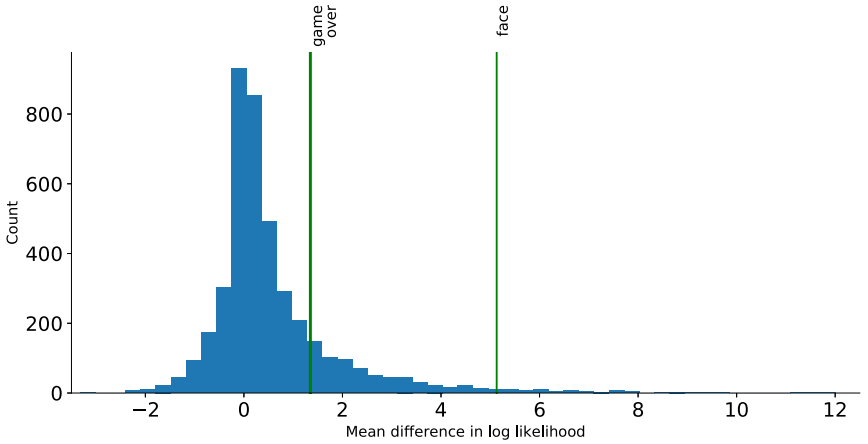

Fig. S6. Histogram of superiority scores of chaining in the BNC word set. Positive and negative values on the *x* axis indicate relative superiority of chaining above and below its competitive models, with locations of example words marked in green (scores of game and "over" are very similar, and hence these words are located close to each other).

Table S1. Two example senses of game recorded in the HTE

Definition of meaning	HTE code	Symbol
Celebratory social event	03.13.02.02|04	•
Ancient match/competition	03.13.04.01|02.02	⋆

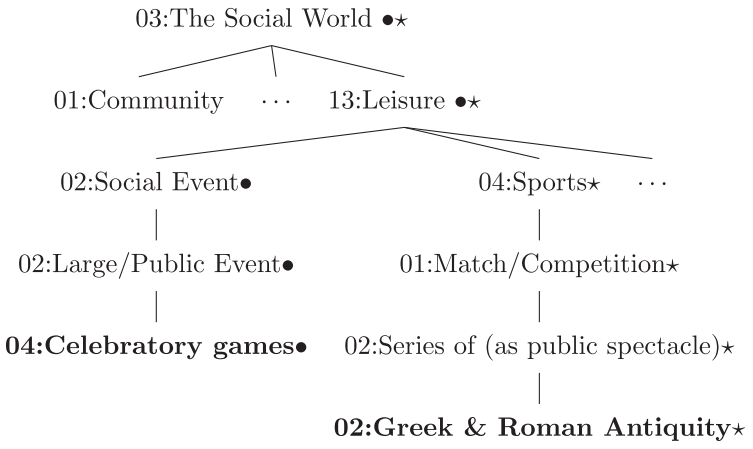

Table S2. Two example senses of play recorded in the HTE

Definition of meaning	HTE code	Symbol
Play a card	03.13.01.05.02.07	•
Play instrument	03.13.03.02.06.03	⋆

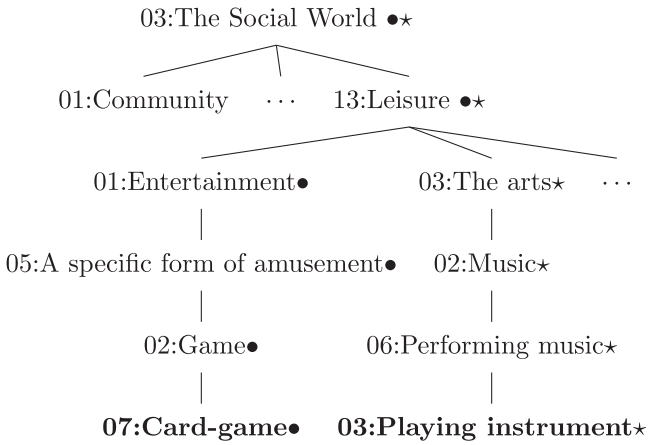

Table S3. Summary of probability (*p*) and cost (*c*) of all possible paths from the progenitor and nearest-neighbor chaining models

Path	$p_{progenitor}$	$c_{progenitor}$	$p_{nnchaining}$	$c_{nnchaining}$
ABCD	0.29	3.95	0.20	3.10
ABDC	0.14	3.95	0.23	3.10
ACBD	0.27	3.95	0.27	3.10
ACDB	0.12	3.95	0.12	3.95
ADBC	0.10	3.95	0.10	3.95
ADCB	0.20	3.95	0.10	3.95

Table S4. Calculation of model likelihoods

Model	$p(B = t1\|A)$	$p(C = t2\|B, A)$	$p(D = t3\|C, B, A)$	Likelihood
Progenitor				
Formula	$\frac{sAB}{sAB+sAC+sAD}$	$\frac{sAC}{sAC+sAD}$	$\frac{sAD}{sAD}$	
Value	0.428...	0.679...	1	0.29
NN chaining				
Formula	$\frac{sAB}{sAB+sAC+sAD}$	$\frac{sAC}{sAC+sBD}$	$\frac{sBD}{sBD}$	
Value	0.428...	0.475...	1	0.20

Here, sAB denotes *sim*(*A*, *B*).

Table S5. Summary of cost and likelihood of the two models in the simulation

Model	Cost	Likelihood
Progenitor	3.95	**0.29**
NN Chaining	**3.10**	0.20

Bold-faced quantities reflect better values along the variables of interest.

Table S6. Example words that showed the best and worst performance of chaining relative to other models chaining in the BNC word set

Word	Best model	Number of senses
Words best predicted by chaining		
Turn	nn	159
Round	nn	154
In	nn	217
Show	nn	140
Shift	nn	102
Order	nn	81
Hold	nn	138
List	nn	59
Strip	nn	79
Sharp	nn	106
Soft	nn	119
Walk	nn	104
Set	nn	192
Point	nn	147
Pull	nn	92
Rank	nn	71
Cross	nn	99
Cut	nn	125
Pitch	nn	114
Miss	nn	67
Words worst predicted by chaining		
Thread	prog.	37
Dark	loc.	53
Standing	loc.	67
Cancer	loc.	14
Bitch	prog.	14
Heart	loc.	46
Tongue	loc.	42
Entertain	prot.	28
Chair	prot.	31
Faculty	loc.	21
Last	loc.	59
Wit	prot.	42
Necessity	prog.	14
Wealth	loc.	13
Striker	prot.	28
Descend	loc.	16
Way	ex.	46
Super	prot.	38
Difference	prot.	18
Eye	prot.	29

"Best model" shows the model that yielded the highest log likelihood ratio score for a given word. Abbreviations "ex.," "prog.," "prot.," "loc.," and "nn" correspond to exemplar, progenitor, prototype, local, and nearest-neighbor chaining models, respectively. "Number of senses" records the number of total senses for a given word from the HTE database.