Mental Algorithms in the Historical Emergence of Word Meanings

Christian Ramiro (chrisram@berkeley.edu)
Cognitive Science Program
University of California, Berkeley

Barbara C. Malt (barbara.malt@lehigh.edu)
Department of Psychology
Lehigh University

Mahesh Srinivasan (srinivasan@berkeley.edu)
Department of Psychology, Cognitive Science Program
University of California, Berkeley

Yang Xu (yang_xu_ch@berkeley.edu)
Department of Linguistics, Cognitive Science Program
University of California, Berkeley

Abstract

Words frequently acquire new senses, but the mental process that underlies the historical emergence of these senses is often opaque. Many have suggested that word meanings develop in non-arbitrary ways, but no attempt has been made to formalize these proposals and test them against historical data at scale. We propose that word meaning extension should reflect a drive towards *cognitive economy*. We test this proposal by exploring a family of computational models that predict the evolution of word senses, evaluated against a large digitized lexicon that dates back 1000 years in English language history. Our findings suggest that word meanings not only extend in predictable ways, but also that they do so following an historical path that tends to minimize cognitive cost - through a process of nearest-neighbor chaining. Our work contributes a formal approach to reverse-engineering mental algorithms of the human lexicon.

Keywords: Word meaning; semantic change; polysemy; chaining; nearest neighbor algorithm; lexicon

Over history, words have frequently acquired new senses, and become *polysemous* (Bréal, 1897). But the mental process that underlies the historical emergence of word senses is often opaque. Wittgenstein's notion of *family resemblance* (Wittgenstein, 1953, p31-32) highlights the challenge for researchers, showing that the many senses of the word *game* form "a complicated network of similarities overlapping and criss-crossing" with nothing identifiably in common (as for board games, card games, ball games, Olympic games, and so on). The network is presumably a reflection of the complex path the word *game* took in the historical development of its meaning. Decades of research have suggested possible ways that word meanings might be mentally structured or extended over time, but none has been tested formally against historical data at scale. We propose that word meanings should develop historically in ways that minimize cognitive effort, hence reflecting a drive towards *cognitive economy* (Zipf, 1949; Rosch, 1975). We test this proposal by formalizing previous theories in computational models that predict how word senses might emerge over time, contributing a principled approach to reverse-engineering mental algorithms of the human lexicon.

Our starting point is a set of influential ideas from cognitive science and linguistics suggesting that word meanings or categories might be structured in non-arbitrary ways. For example, pioneering work by Rosch (Rosch, 1975) showed that common semantic categories signified by words such as *bird* and *furniture* tend to exhibit a prototype structure, such that certain members of a category are more representative than others (e.g., robins and sparrows are more representative as birds than penguins or bats are). Although this theory has since been adapted to describe how word meanings might be structured (Lakoff, 1987) or extended over time (Geeraerts, 1997), it has not been computationally specified or evaluated broadly in accounting for historical patterns in how word senses emerge. A prominent alternative proposal is exemplar theory (e.g., Medin & Schaffer, 1978; Nosofsky, 1986), which suggests that all encountered members of the category are stored and used in categorization judgments, although different members may be weighted differently. This proposal has also been used to describe how language might change over time, particularly concerning phonological and semantic representation (Bybee, 2006). To our knowledge, however, there has been no formal comparison of prototype and exemplar theories with respect to their ability to explain the historical emergence of word senses.

A critical addition to this theoretical terrain is the idea of *chaining* - popularized by Lakoff and other scholars (Lakoff, 1987; Malt, Sloman, Gennari, Shi, & Wang, 1999) - as a possible mechanism that constrains word meaning extension. Chaining operates by linking an emerging idea (an incipient word sense) to a highly-related, already lexicalized word sense. When this process repeats over time, a chained structure in meaning space results. Recent work by Xu et al. (2016) has explored a preliminary version of this proposal via a nearest-neighbor model in a single semantic domain – household containers – but no systematic formalization or evaluation of chaining has been applied to explain the historical emergence of word senses more broadly. Further, al-

though chaining seems plausible as a mechanism, its theoretical value has been limited in two respects: 1) No work has formally specified why chaining might be a preferred mechanism for the development of word meanings; 2) No large-scale assessment of chaining vs. alternative mechanisms has been performed against historical records of word sense extension, leaving open how chaining fares with respect to alternatives. These issues leave open the question of whether the evolution of word meanings follows a cognitively predictable path, and if so, what principles explain this process.

In the current work, we hypothesize that the emergence of word meanings should follow an historical path that minimizes collective cognitive effort. In particular, we propose that chaining should be a preferred algorithm for extending word meanings across history because it tends to minimize the cognitive cost of linking novel ideas with existing ones - a critical property not previously considered with regard to historical sense extension. To test the validity of this argument, we motivate nearest-neighbor chaining with tree-based computer algorithms that minimize edge lengths in a graph. We then formalize the process of chaining as a cognitively economical strategy for encoding novel ideas into an existing lexicon (cf. Xu, Malt, & Srinivasan, 2016).

We critically assess our proposal by developing a family of computational algorithms of word meaning extension - inspired by the previous literature that described above - and evaluate them against a large historical database of word-meaning records in English, spanning over 1,000 years. Our research extends a growing body of work which suggests that structures of language conform to efficient design principles (Zipf, 1949; Rosch, 1975; Piantadosi, Tily, & Gibson, 2011; Kemp & Regier, 2012; Kirby, Tamariz, Cornish, & Smith, 2015), by bringing the perspective of cognitive economy to bear on the evolution of polysemy.

Modeling the emergence of word meanings

Computational formulation

We present here a formulation of five cognitive algorithms that might predict the historical emergence of word meanings, along with a null model. Given the initial, *progenitor* meaning of a word, each non-null algorithm postulates a distinct chaining mechanism by which novel word senses might emerge over time by "attaching to" existing meanings. Each algorithm generates a prediction of the historical order through which the meanings for any given word should emerge, which we then test against the historical record. In effect, we reverse-engineer the mental mechanisms of sense extension.

Table 1 summarizes the full set of proposed algorithms. Here m stands for meaning or word sense, and t stands for time. Each algorithm infers the word sense that emerges at time $t+1$ (m_{t+1}), based on existing senses of a word up to time t ($m_1, ..., m_t$). The inferred sense is drawn from the candidate pool of senses (denoted by m^*) that appear after t for a given word. A perfect model would fully recapitulate the historical emerging order of all senses of a word. All of our models are parameter-free and thus make minimal assumptions in the computational formulation.

Table 1: Proposed models of word meaning extension.

Name	**Description**
Random (null)	$m_{t+1} \sim random\ draw\ m^*$
Exemplar	$m_{t+1} \sim E_{m_i}[sim(m^*, m_i)]$
Prototype	$m_{t+1} \sim sim(m^*, prototype(m_1, ..., m_t))$
Progenitor	$m_{t+1} \sim sim(m^*, m_1)$
Local	$m_{t+1} \sim sim(m^*, m_t)$
Chaining	$m_{t+1} \sim \max_{i=1}^{t} sim(m^*, m_i)$

1. The *random algorithm* – or null model – predicts the historical emergence of a word's senses to be random. This would only be plausible if word senses emerge purely based on immediate communicative needs with no further cognitive constraints.

2. The *exemplar algorithm* adapts from work by Nosofsky (1986), whereby the emerging sense at $t+1$ is predicted to be the one that bears the highest semantic similarity on average (or the highest sum of semantic similarities, which is equivalent in our case) with existing senses of a word at time t. We define semantic similarity identically in all algorithms, and we defer its formal definition to a later section.

3. The *prototype algorithm* is adapted from work by Rosch (1975) and Geeraerts (1997) and predicts the emerging sense at $t+1$ to be the one that bears the highest semantic similarity with the prototypical sense at time t. The prototype at t is defined as the sense that bears the highest semantic similarity with existing senses of a word $prototype(m_1, ..., m_t) \leftarrow \max_i \sum_{j \neq i} sim(m_j, m_i)$. Thus, this algorithm allows the most representative sense of a word to change as a function of time, as more word senses develop.

4. The *progenitor algorithm* is a variant of the prototype model that assumes a fixed prototype that is always the initial, progenitor word sense (i.e., the earliest sense recorded in history). It predicts the emerging sense at $t+1$ to be the one that bears the highest semantic similarity (among all candidate senses) with respect to the progenitor sense.

5. The *local algorithm* assumes that word meanings emerge in a temporal linear chain, where the emerging sense at $t+1$ is the one that bears the highest semantic similarity with the sense that appears at time t. Critically, senses that appear prior to t have no influence on the emerging sense at $t+1$ on this model. This algorithm posits that sense extension will yield minimal cost locally between consecutive time points, as opposed to yielding globally minimal cost (described below).

6. The *chaining (or nearest-neighbor) algorithm* is closely related to Prim's algorithm for constructing a minimal spanning tree (Prim, 1957) - but with a fixed (as opposed to random) starting point, i.e., it always begins with the progenitor sense of a word. In essence, this algorithm predicts the

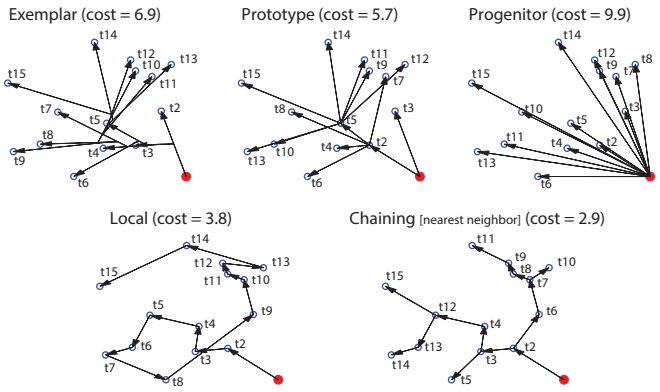

Figure 1: Simulation of the proposed algorithms of word sense extension. The solid red circle symbolizes the progenitor sense of a word. The blue circles represent emerging word senses, and the arrows indicate the predicted path that each algorithm makes about order of emergence. The time labels indicate the predicted sequence of emergence. The cost is the aggregated Euclidean distances traversed by the arrows.

emerging sense at $t+1$ to be the one that bears the highest semantic similarity to any of the existing senses up to t, hence rendering a chain that connects nearest-neighboring senses over time. In contrast with the other algorithms described above, this chaining algorithm is also similar to single linkage clustering (Gower & Ross, 1969) which tends to yield a tree (i.e., each tree node is a sense in this case) with minimal edge lengths among nodes of a graph (i.e., the graph is a network of senses of a word, developed in history). Due to this property, the chaining algorithm assumes the least cumulative historical cognitive effort for the extension of word senses (where effort is inverse to the degree of association between emerging and existing senses of a word), providing the computational implementation of our hypothesis.

Simulation of sense extension algorithms

To illustrate how nearest-neighbor chaining would yield a near-minimal-cost historical path, we provide a simulation for the proposed algorithms of sense extension as follows.

We generated 15 randomly placed points in a two-dimensional plane that represents the meaning space for a hypothetical word (see Figure 1). We took Euclidean distance between-points as a proxy for semantic distance (or inverse semantic similarity) between two senses. We also designated the bottom-right point in the space as the progenitor sense, i.e., it is the earliest seeding sense for the word that is a given to any algorithm. We then applied the family of sense extension algorithms to the remaining data points and visualized the path of emerging senses predicted by each algorithm. Figure 1 shows that these algorithms yield distinct typologies and paths in the simulated meaning space. Specifically, the exemplar algorithm links novels senses to all existing senses based on average distances between them (illustrated by chains that develop from spaces between senses as opposed to those that stem off directly from senses). The prototype algorithm predicts a dynamic radial structure (Lakoff, 1987), where temporal chains are established by linking novel senses to prototype senses, while allowing the prototype to change over time. The progenitor algorithm predicts a strict radial structure where all senses stem from the earliest progenitor meaning. The local algorithm predicts a linear temporal chain of senses by attaching each emerging sense to the existing sense that appears one time point earlier. Finally, the chaining algorithm renders a tree structure that branches off as needed to preserve nearest-neighbor relations between emerging and existing senses. Importantly, the chaining algorithm yields the minimal aggregated edge lengths, hence rendering a minimal cost in semantic space. This result is robust to variations in simulation parameters and is a consequence of the close link between the nearest-neighbor chaining algorithm and the concept of a minimal spanning tree.

Below, we test the extent to which these algorithms can recapitulate the emergence of word senses, as recorded in a large historical lexicon of English.

Treatment of data

Historical lexicon

To evaluate our proposed algorithms, we used the Historical Thesaurus of English (HTE) (Kay, Roberts, Samuels, Wotherspoon, & Alexander, 2015) - a unique large-scale historical lexicon constructed from the Oxford English Dictionary. This database includes approximately 800,000 word forms and their senses, dated and recorded over a span of over 1,000 years - from Old English to the present day. Each word sense in the HTE is annotated with the date of its emergence (and, where applicable, obsolescence) and part of speech, and is structured in a fine-grained semantic hierarchy that features about a quarter of a million concepts. Consecutive tiers of the hierarchy typically follow a `IsA` or `PartOf` relation. For example, one sense of the word *game* under the HTE code "01.07.04.04" is defined in terms of a four-tier hierarchy: `The world (01)`→`Food and drink (01.07)`→`Hunting (01.07.04)`→`Thing hunted/game (01.07.04.04)`.

Measure and validation of semantic similarity

To quantify similarity between word senses, we defined a measure using the semantic hierarchy in the HTE and then validated it against human judgments. Specifically, we approximated psychological similarity between a pair of word senses $sim(m_i, m_j)$ by a common measure of similarity used in psychology that is bounded in the range of (0,1) (Nosofsky, 1986; Shepard, 1987):

$$sim(m_i, m_j) = e^{-d(m_i, m_j)}. \quad (1)$$

Here $d(m_i, m_j)$ represents thesaurus-based conceptual distance between two meanings, which we defined by the inverse of a conceptual similarity measure ($s(\cdot,\cdot)$), commonly used in

natural language processing (Wu & Palmer, 1994; Jurafsky & Martin, 2009):

$$d(m_i, m_j) = 1 - s(m_i, m_j) = 1 - \frac{2 \times |p|}{l(m_i) + l(m_j)}. \quad (2)$$

Here $|p|$ is the number of parent tiers shared by senses m_i and m_j, and $l(\cdot)$ is the depth of a meaning in the semantic hierarchy. This measure gives 1 if two meanings are identical, and 0 if they have nothing in common. Table 2 illustrates the calculation of this measure with a concrete example.

Table 2: Illustration of conceptual similarity based on two senses of *game* recorded in the HTE. Since the two senses share two parent tiers (i.e., The social world→Leisure) in the hierarchy, the conceptual similarity is $s(\bullet, \star) = \frac{2\times 2}{5+6} = \frac{4}{11}$.

Description of sense	HTE code	Symbol
Celebratory social event	03.13.02.02|04	$\bullet$
Ancient match/competition	03.13.04.01|02.02	$\star$

We validated this measure of semantic similarity via standard techniques in natural language processing, by evaluating its performance in predicting human judgments of word similarities. Following Resnik (1995), we approximated word similarity by using the pair of senses for the two words that results in maximum sense similarity, defined as follows: $wordsim(w_i, w_j) = \max_{m_i \in senses(w_i), m_j \in senses(w_j)} s(m_i, m_j)$.

Our measure of semantic similarity yielded a Spearman's correlation of 0.43 ($p < 0.001$) on Lex-999 (Hill, Reichart, & Korhonen, 2015), which provides a well-known data set of human word similarity judgments. The performance of our measure of semantic similarity is better than the corpus-based skipgram (Word2Vec) model, which has been trained on 1 billion words of Wikipedia text (Mikolov, Chen, Corrado, & Dean, 2013) and roughly on par with the same model trained on 300 billion words (Faruqui & Dyer, 2015). In addition, our measure of semantic similarity obtained a Spearman's correlation of 0.52 ($p < .001$) on Sim-353 (Finkelstein et al., 2001), another common data set of human word relatedness judgments, which is comparable to the state-of-the-art GLOVE word vector model, which has been trained on 6 billion words (Faruqui & Dyer, 2015; Pennington, Socher, & Manning, 2014).

Having validated our measure of semantic similarity, we used it to assess the mental algorithms described above.

Choices of words

We focused our analyses on explaining word sense extension in a set of the most common English words. Specifically, we worked with the most frequent 6318 words in the British National Corpus (BNC). Some of the word forms are duplicated in this set because one word can function in multiple part-of-speech categories. However, our results were robust regardless of whether we collapsed these words by form or distinguished them by part-of-speech.

Model evaluation and results

We used model likelihood to assess the performance of each proposed algorithm.[1] We defined likelihood as a probability function that specifies the degree to which a model accounts for the entire sequence of senses that historically emerged for a given word. To be concrete, for a sequencce of senses $m_1, m_2, m_3, ..., m_t$, the likelihood $\mathcal{L}$ is the joint probability of observing such a sequence under a certain model $\mathcal{M}$:

$$\mathcal{L}_{\mathcal{M}} = p(m_1)p(m_2|m_1)p(m_3|m_1, m_2)...p(m_t|m_1, ..., m_{t-1}). \quad (3)$$

We assumed that the progenitor sense is always given, so $p(m_1) = 1$. For all remaining emerging senses, the set of algorithms can be evaluated by calculating likelihood based on their specifications in Table 1. For example, the progenitor model would yield a likelihood for the emerging sense at $t = 2$ (conditioned on that appeared at $t = 1$) as follows:

$$p(m_2|m_1) = \frac{sim(m^*, m_1)}{\sum_{m^* \in \{m_2, ..., m_t\}} sim(m^*, m_1)}. \quad (4)$$

The algorithm then steps through each point in time and the likelihood correspondingly calculates the degree to which the algorithm predicts the true emerging sense at that point, among a candidate pool of senses that appear after.

Because our null hypothesis is that there exists no predictability in how word senses develop in history, we evaluated each cognitive algorithm against the random null algorithm, using the log likelihood ratio (*LLR*) - a standard metric for model comparison in statistics: $LLR = \log(\mathcal{L}_{\mathcal{M}}/\mathcal{L}_{null})$. This quantity should be greater than 0 if a given model accounts for word sense extension better than the null, and the converse if the null does better. For any given word, the likelihood function of the null can be determined theoretically, and it is simply the inverse of factorial of $N-1$ for a word with N senses: $\mathcal{L}_{null} = 1 \times \frac{1}{N-1} \times \frac{1}{N-2} \times ... \times \frac{1}{1} = \frac{1}{(N-1)!}$. Thus the log likelihood ratio indicates whether a model predicts the sequence of emerging word senses better than chance.

[1] Because each of the models we examined is parameter-free, metrics that take into account model complexity such as the Bayesian Information Criterion would give identical results to those only taking into account likelihood.

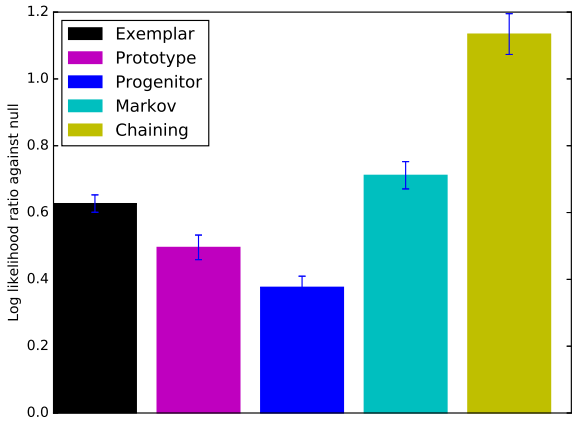

Figure 2: Summary of model performances against the null. "0.0" on the y-axis indicates performance of the null model. Bar height indicates the mean log likelihood ratio averaged over the entire pool of most common words from the BNC corpus. Error bars indicate 95% confidence intervals.

Figure 2 summarizes the results. The bar plot shows that each of the proposed algorithms accounts for the historical data that we examined significantly better chance ($p < 0.001$ from 1-tailed t-tests), reflected in the positive log likelihood ratios. This observation suggests that the null hypothesis can be rejected: The emerging order of word senses in the English lexicon is not purely random.

Critically, the nearest-neighbor chaining algorithm yielded the highest overall likelihood among all models, and this result was statistically significant according to paired t-tests between the chaining model and each of the remaining models ($p < 0.001$ in all four comparisons). This observation provides evidence that word senses emerge in cognitively efficient ways by approximating a minimal spanning tree over the course of history. As such, these data support our hypothesis about nearest-neighbor chaining as the dominant mental algorithm for the historical emergence of word senses.

To illustrate the nearest-neighbor chaining process, we visualized the predicted chaining path for the English word *game*. Figure 3 shows a low-dimensional projection (via multi-dimensional scaling with a random starting point) of all senses of *game* as a noun, taken from the HTE database. As can be seen, the chaining algorithm forms a minimal spanning tree among the senses of *game*, by linking neighboring nodes that are semantically close. Importantly, this process of meaning extension tends to support branching and the formation of local clusters, identified roughly in this case by the three sense groups of "hunting game" (upper-left cluster), "scheme" (middle cluster), and "sports and entertainment" (upper-right cluster) in Figure 3. This offers a computational basis for *family resemblance* (Wittgenstein, 1953) and polysemy, by allowing words to develop both related and distinct senses over time.

Conclusions

We presented the first large-scale computational investigation of the mental algorithms that determine how words evolve new senses over time. We found that the historical emergence of word senses in English is not arbitrary; Instead, it has exhibited a high degree of predictability over the past millennium. Our findings indicate that the order in which word senses emerge can be best accounted for by a process of nearest-neighbor chaining, which supports the view that the historical development of the lexicon follows a trajectory that tends to minimize cognitive effort. Our current analysis focuses on sense extension within individual word forms, but it would be useful to extend our analysis to examine how different words compete to express novel meanings. Our exploration of the mental algorithms that underlie historical sense extension opens new, interdisciplinary venues for reverse engineering the evolution of the human lexicon.

Acknowledgments

We thank Marc Alexander for helping with data licensing, and the University of Glasgow and the Oxford University Press for making the Historical Lexicon of English available. We thank Charles Kemp and Terry Regier for discussions on chaining and the Computational and Experimental Methods group in Linguistics at UC Berkeley for constructive comments. This project was partly funded by NSF award SBE-1041707 to the Spatial Intelligence and Learning Center (SILC) and NSF award SBE-16302040 to MS.

References

Bréal, M. (1897). *Essai de sémantique: Science des significations*. Paris: Hachette.

The British National Corpus, version 3 (BNC XML Edition). (2007). (Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/)

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733.

Faruqui, M., & Dyer, C. (2015). Non-distributional word vector representations. *arXiv preprint arXiv:1506.05230*.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on world wide web* (pp. 406–414).

Geeraerts, D. (1997). *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford: Oxford University Press.

Gower, J. C., & Ross, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, *18*(1), 54–64.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

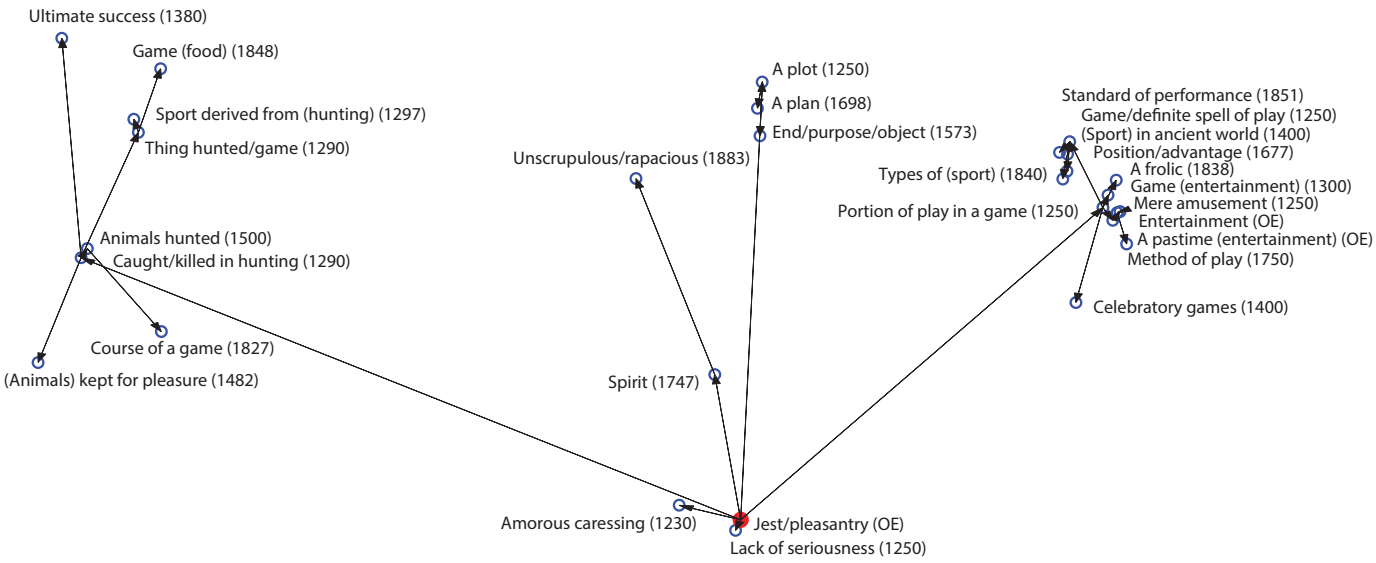

Figure 3: Historical chaining in the English word *game*. The two-dimensional space is generated by multi-dimensional scaling based on sense similarities. The solid red circle marks the earliest meaning. The arrows indicate the predicted path from the chaining algorithm. The annotations include a gloss for the sense and its recorded period of emergence in the HTE.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). New Jersey: Pearson Education.

Kay, C., Roberts, J., Samuels, M., Wotherspoon, I., & Alexander, M. (2015). *The historical thesaurus of english, version 4.2*. Glasgow: University of Glasgow.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Emnlp* (Vol. 14, pp. 1532–1543).

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell Labs Technical Journal*, *36*(6), 1389–1401.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Rosch, E. H. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Basil Blackwell.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on association for computational linguistics* (pp. 133–138).

Xu, Y., Malt, B. C., & Srinivasan, M. (2016). Evolution of polysemous word senses from metaphorical mappings. In *Proceedings of the 38th annual meeting of the cognitive science society.*

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*, 2081–2094.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Boston: Addison-Wesley.