# Word reuse and combination support efficient communication of emerging concepts

Aotao Xu[1,2], Charles Kemp[3], Lea Frermann[2], Yang Xu[1,4]

[1]Department of Computer Science, University of Toronto

[2]School of Computing and Information Systems, University of Melbourne

[3]School of Psychological Sciences, University of Melbourne

[4]Cognitive Science Program, University of Toronto

**Abstract**

A key function of the lexicon is to express novel concepts as they emerge over time through a process known as lexicalization. The most common lexicalization strategies are the reuse and combination of existing words, but they have typically been studied separately in the areas of word meaning extension and word formation. Here we offer an information-theoretic account of how both strategies are constrained by a fundamental tradeoff between competing communicative pressures: word reuse tends to preserve the average length of word forms at the cost of less precision, while word combination tends to produce more informative words at the expense of greater word length. We test our proposal against a large dataset of reuse items and compounds that appeared in English, French and Finnish over the past century. We find that these historically emerging items achieve higher levels of communicative efficiency than hypothetical ways of constructing the lexicon, and both literal reuse items and compounds tend to be more efficient than their non-literal counterparts. These results suggest that reuse and combination are both consistent with a unified account of lexicalization grounded in the theory of efficient communication.

# 1 Introduction

The human lexicon is not static but evolves over time. A central role of the lexicon in this evolutionary process is to lexicalize novel ideas, therefore serving as an adaptive system which supports the symbolic encoding and communication of emerging concepts (Deacon, 1997; Pinker & Bloom, 1990). The most common strategies of lexicalization involve reusing and combining existing words in the lexicon (Marchand, 1969; Algeo, 1980; Brinton & Traugott, 2005; Ramiro, Srinivasan, Malt, & Xu, 2018), although other strategies such as borrowing (e.g., *tofu*) and coinage (e.g., *quark*) also exist. Reuse refers to using the form of an existing word to express something new. For instance, *mouse* was reused to describe a "small device that is moved by hand across a surface to control the movement of the cursor on a computer screen" (Oxford University Press, 2023). Combination refers to concatenating two or more existing words to form a new word, typically known as a compound (e.g., *armchair* combines *arm* and *chair* to express a new type of chair). Word reuse and combination are often viewed and studied separately as two distinct aspects of lexical evolution. Here we present an information-theoretic framework that accounts for both strategies through the lens of efficient communication.

Word reuse has been traditionally discussed in the context of historical semantic change (e.g., Traugott & Dasher, 2001) and word meaning extension (e.g., Williams, 1976). This body of work aims to identify regularities in how words take on new meanings over time. More recent studies using large-scale historical and cross-linguistic data have suggested that words tend to take on new meanings that are semantically related to existing ones (Y. Xu, Regier, & Malt, 2016; Ramiro et al., 2018), and the processes of word meaning extension reflect cognitively economic ways of expanding the referential range of existing words in the lexicon (Srinivasan & Rabagliati, 2015; Y. Xu, Duong, Malt, Jiang, & Srinivasan, 2020). However, this line of research focuses almost exclusively on word reuse (i.e., with no overt changes in word form) and treats it in isolation from word combination.

Word combination has been commonly studied in the literature on word formation and

morphology (e.g., Štekauer & Lieber, 2005). In particular, existing accounts focusing on noun-noun compounds have suggested that compound interpretation involves selecting from a systematic list of predicate relations which in turn constrains possible noun combinations produced by speakers (e.g., Levi, 1978; Lieber, 1983; Levin, Glass, & Jurafsky, 2019). An alternative approach appeals to more functionally motivated principles arguing that compounds should be semantically transparent while shorter forms are preferred if compound constituents are redundant (Downing, 1977; Dressler, 2005; Costello & Keane, 2000). These principles have also been discussed in psycholinguistic work showing that the semantic relatedness between a novel compound and its constituents predicts its acceptability, unless the relatedness is too high (Günther & Marelli, 2016). We believe these functional principles might be equally applicable to explaining word reuse. However, to our knowledge there is no unified account that characterizes both word reuse and combination (c.f. Blank, 2003; A. Xu, Kemp, Frermann, & Xu, 2023).

We propose a unified account of word reuse and combination by building on the view that language is shaped to support efficient communication (e.g., Regier, Kemp, & Kay, 2015; Kemp, Xu, & Regier, 2018; Gibson et al., 2019; Hahn, Jurafsky, & Futrell, 2020). This line of work suggests that linguistic structures are shaped by functional pressures to maximize informativeness and simplicity (or ease of use) in communication. Efficiency-based accounts have been shown to explain word meaning variation across languages (e.g., Kemp & Regier, 2012; Regier et al., 2015; Y. Xu, Liu, & Regier, 2020; Zaslavsky, Kemp, Regier, & Tishby, 2018), the structures of word forms (e.g., Zipf, 1949; Mahowald, Dautriche, Gibson, & Piantadosi, 2018; Bentz & Ferrer Cancho, 2016; Hahn, Mathew, & Degen, 2022), and grammatical form-meaning mappings (Mollica et al., 2021). We extend this growing body of research with the aim to understand the general principles that shape the diverse strategies of lexicalization.

Here we extend existing efficiency-based accounts that assume speakers and listeners use the same static lexicon by considering communicative interactions during the spread of linguistic innovations (e.g., Labov, 2011; Milroy & Milroy, 1985). In particular, we consider

the labels of novel concepts that are yet to be encoded in the lexicon of a large number of language users. We propose that when speakers communicate novel concepts to these language users, word reuse and combination reflect a fundamental tradeoff between speaker effort and information loss (or the inverse of informativeness): on the one hand, speakers can minimize their effort by reusing short words that underspecify intended concepts; on the other hand, information loss is minimized if speakers use relatively long word forms that combine existing words in an informative way. We hypothesize that as speakers repeat these communicative interactions, their encoding of novel concepts is shaped by the pressure to minimize speaker effort and the opposing pressure to minimize information loss, such that the word length and informativeness of both attested reuse items and attested compounds efficiently trade off against each other. We illustrate this idea in Figure 1.

Our proposal is consistent with other functional accounts of word reuse and combination. Previous accounts have separately suggested that combinations should be informative (Lieber, 2004; Clark & Berman, 1984; Downing, 1977) and reuse represents an economical strategy (Štekauer, 2005; Piantadosi, Tily, & Gibson, 2012). Computational studies of reuse items (Ramiro et al., 2018; Brochhagen, Boleda, Gualdoni, & Xu, 2023) and compounds (Günther & Marelli, 2016; Vecchi, Marelli, Zamparelli, & Baroni, 2017; Pugacheva & Günther, 2024) have shown that semantic transparency is preferred across both strategies, which is consistent with a preference for using informative word forms to reduce communicative error. In recent work on language evolution, a pressure for informativeness is often considered necessary for compositional or subword structure to emerge in the lexicon (e.g., Kirby, Tamariz, Cornish, & Smith, 2015; Carr, Smith, Cornish, & Kirby, 2017). Crucially, prior studies have shown that a pressure for informativeness may be sufficient for the emergence of subword structure when speakers have to communicate novel meanings to a large community with limited shared history (Raviv, Meyer, & Lev-Ari, 2019a, 2019b). Here we build on these existing studies to offer a unified functional account of word reuse and combination.

In the following, we first specify our theoretical proposal in formal terms. We then evaluate

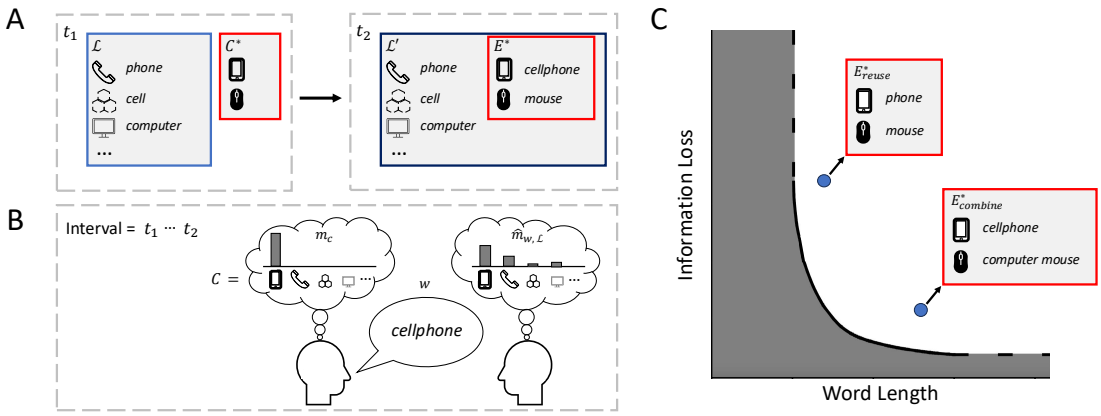


Figure 1: Illustration of our theoretical proposal. Panel (A) illustrates the lexicalization of emerging concepts using examples from English during the historical interval 1980-2000. The existing lexicon $\mathcal{L}$ and the set of emerging concepts $\mathcal{C}^*$ at time $t_1$ are illustrated on the left. At a later time $t_2$, the attested encoding of the novel concepts $E^*$ enters the expanded lexicon $\mathcal{L}'$, which are shown on the right. Panel (B) illustrates the two opposing pressures in a communicative interaction taking place before $t_2$. Here the speaker intends to convey the emerging concept "cellphone" to a listener whose lexicon does not yet have a word for expressing it, and grey bars illustrate probability distributions over a universe of concepts $\mathcal{C}$ that capture uncertainty regarding the intended concept. Our proposal focuses on the pressure for minimizing the length of the utterance, and the pressure for minimizing information loss, or the difference between the speaker and listener distributions over concepts. Panel (C) illustrates possible encodings of the novel concepts in Panel (A). Each point corresponds to the average length and information loss of an encoding of the novel concepts, and the shaded area corresponds to costs that are not attainable. We propose that word reuse and combination reflect a tradeoff between these two costs, and that both attested reuse items and attested compounds achieve tradeoffs that are relatively efficient. Here the example encodings are simplified to contrast reuse and combination, and in reality an encoding can consist of both strategies.

our efficiency-based proposal against a large historical dataset of reuse items and compounds attested in English, French, and Finnish over the past century. Lastly, we discuss the implications of our results and avenues for future work.

## 2 Computational Formulation of Theory

To specify our theoretical proposal, we formulate a scenario in which the attested encoding of emerging concepts spreads within a speech community. Our formulation builds on a standard approach in language change (e.g., Weinreich, Labov, & Herzog, 1968; Milroy & Milroy, 1985; Traugott & Dasher, 2001; Brinton & Traugott, 2005; Labov, 2011), which views the evolution of linguistic structures as a gradual process in which new structures coexist with existing structures as the former spread among speakers.

We illustrate the setting for this scenario in Figure 1A. Here we consider an encoding of concepts as a set of form-concept pairs (or mappings), and we treat the lexicon as an encoding of lexicalized concepts which may incorporate novel pairs over time. Suppose the existing lexicon $\mathcal{L}$ consists of form-concept pairs known by all speakers in the speech community at time $t_1$, and the set $\mathcal{C}^*$ contains novel concepts emerging at $t_1$ but not encoded in the existing lexicon. We assume that the eventual attested encoding of novel concepts, denoted by $E^*$, spreads among speakers until the expanded lexicon $\mathcal{L}' = \mathcal{L} \cup E^*$ has been acquired by all speakers at time $t_2$. In the current study, we assume that forms in $E^*$ always reuse or combine forms that exist in $\mathcal{L}$.

In the following, we will focus on the time interval between $t_1$ and $t_2$ in which both the existing and expanded lexicons coexist within the speech community. We first specify an information-theoretic model of communication. We then build on the model to define two types of communicative cost and specify our theoretical proposal regarding forms in $E^*$.

## 2.1 Model of Communication

To assess the role of communicative efficiency in shaping $E^*$, we first consider the communicative interaction between a speaker who uses the expanded lexicon $\mathcal{L}'$ and a listener who uses the existing lexicon $\mathcal{L}$. We model this interaction by extending previous efficiency-based accounts of the lexicon (Kemp et al., 2018; Zaslavsky et al., 2018) that are grounded in Shannon's original point-to-point model of communication (Shannon, 1948).

We describe our model using an illustration of the interaction in Figure 1B. The speaker's mental representation is a speaker distribution $m_c$ over a universe of concepts $\mathcal{C}$. In general $m_c$ could capture speaker uncertainty, but we assume that $m_c$ corresponds to a single intended concept $c$ and picks it out with certainty. The intended concept $c$ is drawn from a need distribution $p(c|\mathcal{L}')$ which captures the frequency with which different concepts are communicated (e.g., Kemp et al., 2018; Zaslavsky et al., 2018), and here we assume the speaker only communicates concepts encoded in her lexicon. To express $c$, the speaker selects a form $w$ according to her production policy $p(w|c, \mathcal{L}')$ which captures the frequency of using specific forms in her lexicon to communicate an intended concept. In turn, the listener uses $w$ and his lexicon to deterministically construct his mental representation which is a listener distribution $\hat{m}_{w,\mathcal{L}}$ that aims to reconstruct $m_c$.

We define the listener distribution by using a variant of prototype-based categorization models (e.g., Rosch, 1975; Ramiro et al., 2018). In our model, the listener treats each form $w$ as the label of a category of concepts, which is represented by the category prototype $q_w$. The listener uses the category to construct a distribution, so that the probability of concept $c$ is high if it is semantically similar to $q_w$. We specify this distribution via the similarity choice model (Luce, 1963; Nosofsky, 1986):

$$\hat{m}_{w,\mathcal{L}}(c) \propto \exp\{-\gamma d(c, q_w)\} \tag{1}$$

where $d(\cdot, \cdot)$ is semantic distance, and $\gamma \geq 0$ is a sensitivity parameter that controls how fast probability decreases with distance. We require the prototype to be a function of form-concept pairs in $\mathcal{L}$, but its exact definition depends on the specific dataset to which we

apply our framework.

## 2.2 Communicative Costs

To specify our theoretical proposal, we now consider the speaker effort and information loss incurred over repetitions of the above interaction. In reality, these interactions take place across multiple speakers and repeatedly within the same dyads, but these dynamics introduce heterogeneity among listener distributions as listeners adopt new form-concept pairs into their lexicons. For simplicity, we consider the case in which a single speaker interacts once with each of many distinct listeners, such that the listener distributions are independent and identical across interactions. The speaker in this special case may be construed as a leader in spreading linguistic innovations in local communities (e.g., Labov, 2011; Milroy & Milroy, 1985; Del Tredici & Fernández, 2018).

Following efficiency-based accounts of word length (e.g., Zipf, 1949; Mollica et al., 2021), we measure the speaker effort in an interaction via the length of the produced utterance. As the same speaker communicates with many listeners, the average speaker effort over their interactions is given by expected word length:

$$\mathbb{E}[l(W)|\mathcal{L}'] = \sum_{c,w} p(c, w|\mathcal{L}')l(w) \tag{2}$$

where $l(\cdot)$ is the length of a form. Previous studies on coding efficiency have shown that word frequency and length tend to be related in a way that is relatively efficient (e.g., Zipf, 1949; Bentz & Ferrer Cancho, 2016; Mollica et al., 2021). Here we extend these studies by examining the tradeoff between length and information loss in reused and compound forms that express novel concepts.

Following Regier et al. (2015), we define the information loss in a single interaction as the Kullback–Leibler (KL) divergence between the speaker and listener distributions. In our case of a single speaker and many distinct listeners, the average information loss over their

interactions is given by expected KL divergence:

$$\mathbb{E}[D(M||\hat{M})|\mathcal{L}', \mathcal{L}] = \sum_{c,w} p(c, w|\mathcal{L}')h(\hat{m}_{w,\mathcal{L}}(c)) \tag{3}$$

where $h(\cdot) = -\log_2(\cdot)$. In contrast to previous efficiency-based accounts of lexical semantics (e.g., Regier et al., 2015; Zaslavsky et al., 2018), here we consider information loss when the production policy and the listener distribution are conditioned on different lexicons.

### 2.3 Efficiency of Attested Encodings

Our proposal can be specified by considering how much the attested encoding $E^*$ contributes to these communicative costs. This contribution can be summarized by a single objective function obtained from combining and simplifying Equations 2 and 3:

$$L_\beta[E^*|\mathcal{L}] = \mathbb{E}[D(M||\hat{M})|\mathcal{L}', \mathcal{L}] + \beta\mathbb{E}[l(W)|\mathcal{L}'] \tag{4}$$

$$\propto \sum_{(c,w)\in E^*} p(c, w|\mathcal{L}') \cdot (h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w)) \tag{5}$$

where $\beta \geq 0$ is a tradeoff parameter. As in previous efficiency-based approaches (Zaslavsky et al., 2018; Mollica et al., 2021), optimizing Equation 5 for every $\beta$ produces a Pareto frontier that specifies the space of possible encodings of $\mathcal{C}^*$, which we illustrate in Figure 1C. We hypothesize that the attested encoding is shaped by competing pressures for minimizing speaker effort and information loss, which predicts that it will be relatively close to the Pareto frontier in this space of possible encodings.

In *SI Appendix, Section S1*, we provide a more detailed derivation of Equations 2-5. We also provide a discussion on the limitations of previous efficiency-based accounts of the lexicon in terms of accounting for attested combinations of words or morphemes.

## 3 Results

To test our proposal, we instantiated our scenario using Princeton WordNet (Fellbaum, 1998) and its multilingual extensions (Bond & Foster, 2013). These WordNets are conceptually organized dictionaries that map language-specific, orthographic forms to a common

set of word senses or lexicalized concepts. We focused on English, French, and Finnish because they have the largest WordNets among alphabetical languages in terms of the size of their sense inventory (Bond & Foster, 2013). For each language and one of five consecutive intervals over the past century, we instantiated emerging concepts as WordNet senses, and we instantiated their attested encoding and the existing lexicon as sets of form-sense pairs. We identified whether an English form-sense pair is an emerging reuse item or compound by using its first citation in the Historical Thesaurus of English (Kay, Roberts, Samuels, & Wotherspoon, 2017), and we inferred whether a pair is existing based on its estimated frequency in historical text (Davies, 2002; Michel et al., 2011). We implemented the same components for French and Finnish by relabelling emerging and existing senses in the English data with a language-specific form that was attested in historical French or Finnish text (Michel et al., 2011; National Library of Finland, 2014). For tractability, we ignored linking constituents in compounds and we used compounds that have exactly two constituents. More details on data processing are provided in *Materials and Methods*.

To show that our approach applies to both reuse and combination, we controlled for differences in sample size between strategies by instantiating each attested encoding such that it only contains reuse items or compounds. Across our target intervals, we analyzed 518 reuse items and $2,828$ compounds in English, 529 reuse items and 409 compounds in French, and 510 reuse items and 645 compounds in Finnish; sample sizes for specific intervals are provided in *SI Appendix, Section S2*. In Table 1, we show examples of reuse items and compounds that make up these encodings.

Given the existing lexicon, we computed the average length and information loss incurred by a speaker communicating with an encoding of novel concepts as specified in Equation 5. We used the orthographic length of word forms in our subsequent analyses as a proxy for production effort. To estimate information loss, we implemented the listener distribution in Equation 1 in three parts. We first represented each concept or word sense by embedding the text of its WordNet definition (see Table 1 for examples) with a sentence encoder (Reimers & Gurevych, 2019), and we followed approaches that construct the prototype as an average

| Language | Interval | Strategy | Form | Sense Definition |
|---|---|---|---|---|
| English | 1940+ | R | locker | a trunk for storing personal ... |
| | 1940+ | R | printer | an output device that prints the ... |
| | 1940+ | R | dish | directional antenna consisting of a ... |
| | 1900+ | C | birthday card | a card expressing a birthday ... |
| | 1940+ | C | urban renewal | the clearing and rebuilding and ... |
| | 1980+ | C | spreadsheet | a screen-oriented interactive ... |
| French | 1900+ | R | antenne | an electrical device that sends or ... |
| | 1920+ | R | publicité | a commercially sponsored ad on ... |
| | 1960+ | R | émuler | imitate the function of ... |
| | 1900+ | C | turbine à gaz | turbine that converts the chemical ... |
| | 1900+ | C | galaxie spirale | a galaxy having a spiral structure; ... |
| | 1940+ | C | boîte noire | equipment that records information ... |
| Finnish | 1900+ | R | lähetys | message that is transmitted by ... |
| | 1920+ | R | suodatin | an air filter on the end of a ... |
| | 1940+ | R | ajaa | carry out a process or program, as ... |
| | 1900+ | C | sotarikos | a crime committed in wartime; ... |
| | 1900+ | C | taisteluväsymys | a mental disorder caused by stress ... |
| | 1920+ | C | kauppa-apulainen | a salesperson in a store |

Table 1: Examples of reuse items (R) and compounds (C) that emerged in the past century; sense definitions have been truncated for brevity

of the existing senses of a word or its constituents (Reed, 1972; Mitchell & Lapata, 2008). For the sensitivity parameter, we set $\gamma = 10$ based on the informativeness of existing words. These costs were computed for each form-sense pair in the encoding, and then averaged according to their need and production probabilities estimated from historical form-sense frequencies. We specify our implementation in *Materials and Methods*.

In the following, we first directly assess the average-case efficiency of attested reuse-based and combination-based encodings. We then perform fine-grained analyses on the efficiency of individual reuse items and compounds.

| Attested Label | Near-Synonyms |
|---|---|
| locker | deedbox, strongbox, clothespress, storeroom |
| urban renewal | renewal, renovation, urban-renovation |
| publicité | réclame, annonce, pub, emballage |
| turbine à gaz | turbine, générateur, turbine-fluide, moteur-gaz |
| lähetys | lasti, rahti, toimitus, kuorma |
| sotarikos | rikos, laittomuus, sota-laittomuus, hyökätä-rikos |

Table 2: Samples of near-synonyms created for attested labels

### 3.1 Average-case Efficiency

In our first analysis, we compared attested items to optimal encodings on the Pareto frontier and two sets of baseline encodings. The first baseline consists of alternate encodings created from replacing the label of each item in an attested encoding with a near-synonym; examples of near-synonyms are shown in Table 2. To probe the space of all possible alternatives, we created a second baseline by replacing each attested label with a string uniformly sampled from labels in the existing lexicon and their combinations. Details on estimating the Pareto frontier and creation of baseline encodings are specified in *Materials and Methods*.

Figure 2 summarizes the comparisons between attested and alternative encodings. Each Pareto frontier shows the optimal tradeoff that can be achieved by any encoding of the emerging concepts, and intuitively, the closeness of an encoding to the frontier approximates its efficiency. Across strategies, intervals and languages, we observe that both attested encodings (blue) and near-synonym baselines (light blue) tend to be closer to the frontier and more efficient than random baselines (grey), and attested encodings tend to be more efficient than both baselines. By construction, attested and near-synonym encodings tend to be shorter than random baselines because it is more likely to sample a combination of long words than a single short word. The fact that attested labels also dominate near-synonyms indicates the relative efficiency of attested labels does not arise solely due to chance and the prevalence of longer word forms.

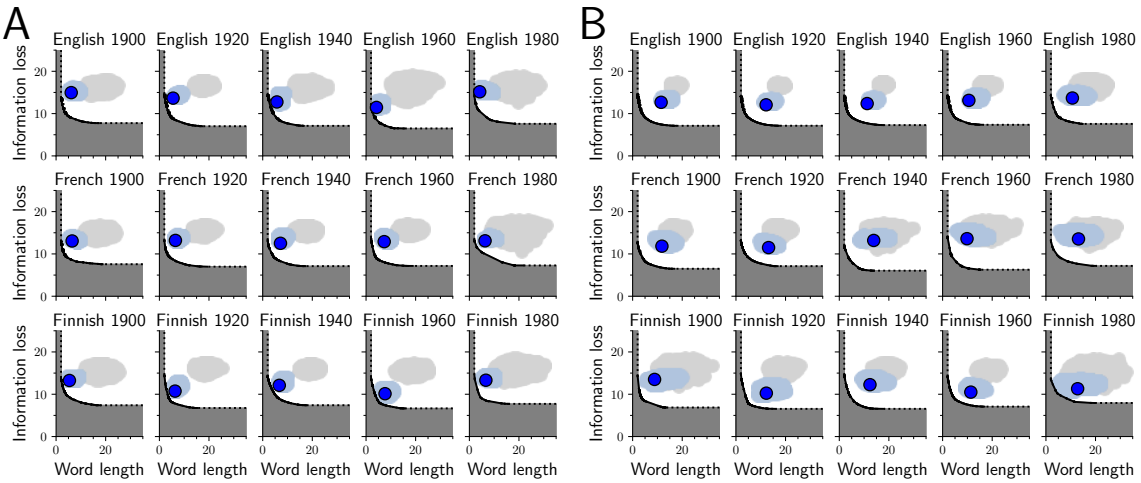


Figure 2: Illustration comparing (A) attested reuse items and (B) attested compounds to the constructed baselines and the Pareto frontier. Every point corresponds to an encoding of emerging concepts for a specific language and interval. Attested cases are marked in blue, near-synonym baselines in light blue, and random baselines in grey. Black solid lines in the bottom left show the estimated Pareto frontier, and the shaded areas show costs that are not attainable.

Figure 3 compares attested encodings against baselines using a quantitative measure of efficiency loss (see *Materials and Methods*), which overall confirms our qualitative observations. In *SI Appendix, Section S4*, we show that attested reuse items and compounds remain more efficient than the constructed baselines under different implementations of our scenario of lexical evolution. First, we show our results are robust if we use a uniform distribution over attested items and different values for the sensitivity parameter. Second, we show the results hold up across different communication channels via an implementation that represents word forms using phonemes instead of letters. Third, we describe an analysis based on historical embeddings (Mikolov, Chen, Corrado, & Dean, 2013; Hamilton, Leskovec, & Jurafsky, 2016) to address the concern that our approach may be biased by using contemporary embeddings to study historical change. Lastly, we show these findings are robust to alternative datasets of lexicalized concepts by considering another English dictionary and

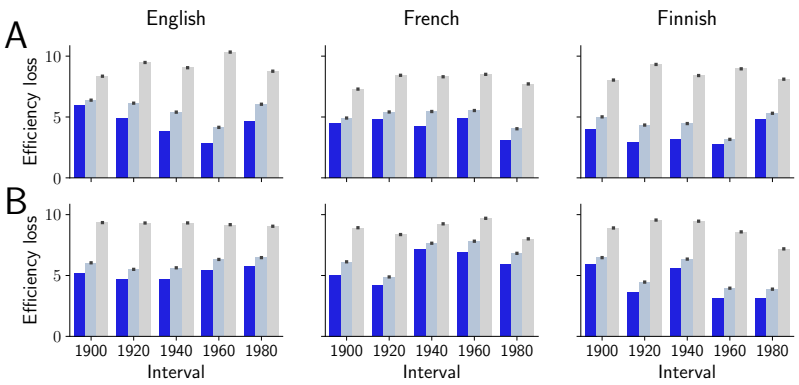


Figure 3: Efficiency loss of attested encodings for (A) reuse items and (B) compounds relative to the average loss of baselines. Attested loss is marked in blue, and the average loss of near-synonym and random baselines is marked in light blue and grey, respectively. Error bars show bootstrapped 95% confidence intervals.

by assuming one-to-one correspondence between concept and form.

### 3.2 Item-level Variation

In Figure 2, we observe non-trivial gaps between Pareto frontiers and attested encodings. This suggests that the communicative efficiency of attested encodings could be improved by replacing some attested reuse items and compounds with more efficient forms. We thus compared individual items to optimized forms by using the same implementation of our scenario, except we replaced full encodings with singletons that contain individual items.

Figure 4 shows efficiency losses for individual items, which measure their deviation from optimized forms, and each distribution is aggregated over all time intervals. As in the previous analysis, the item-level loss is approximated by the distance between attested items and Pareto frontiers, which is illustrated in Figure 5. We observe that attested items tend to be much closer to optimized forms (loss = 0) than most randomly sampled labels (grey), but nonetheless the right tails of attested items overlap with random distributions. In *SI Appendix, Section S5.A*, we show that item-level loss based on orthographic forms

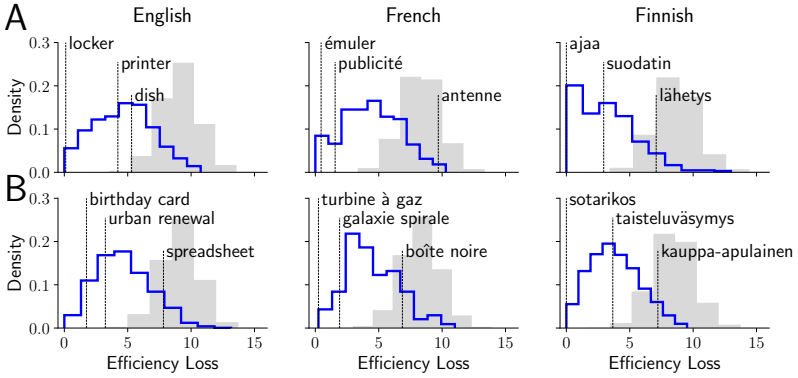


Figure 4: Efficiency loss of individual attested items for (A) reuse and (B) compounding and randomly sampled labels. The distributions for attested and random are marked in blue and grey, respectively. Examples in Table 1 are annotated.

strongly correlates with item-level loss based on phonemic forms. The variation from near-optimal to near-random among attested reuse items and compounds reveals that some items are more strongly shaped by our proposed tradeoff than others.

Here we characterize this variation using two well-studied subclasses of lexical items. Endocentric compounds are the most well-studied subclass of compounds that are defined by the relation between intended and existing constituent concepts (e.g., Downing, 1977; Jackendoff, 2010). The head word of an endocentric compound encodes a superordinate category of the intended concept and is a literal expression of this concept (e.g., *birthday card* is a *card*), in contrast to non-literal (or exocentric) compounds (e.g., *blue-collar*). Similarly, a subclass of reuse items often expresses an intended concept that is more narrow than an existing sense of the reused word (e.g., Bloomfield, 1933), for instance the modern use of *car* as a motorized vehicle refers to a narrower set of concepts compared to its original meaning of *wheeled cart*. Across strategies, these literal expressions may be more efficient because they are more transparent to the listener than non-literal ones. For example, in Figure 5, we observe that literal reuse items and endocentric compounds tend to be more efficient than their non-literal counterparts (e.g., *birthday card* vs *dish* or *antenne*).

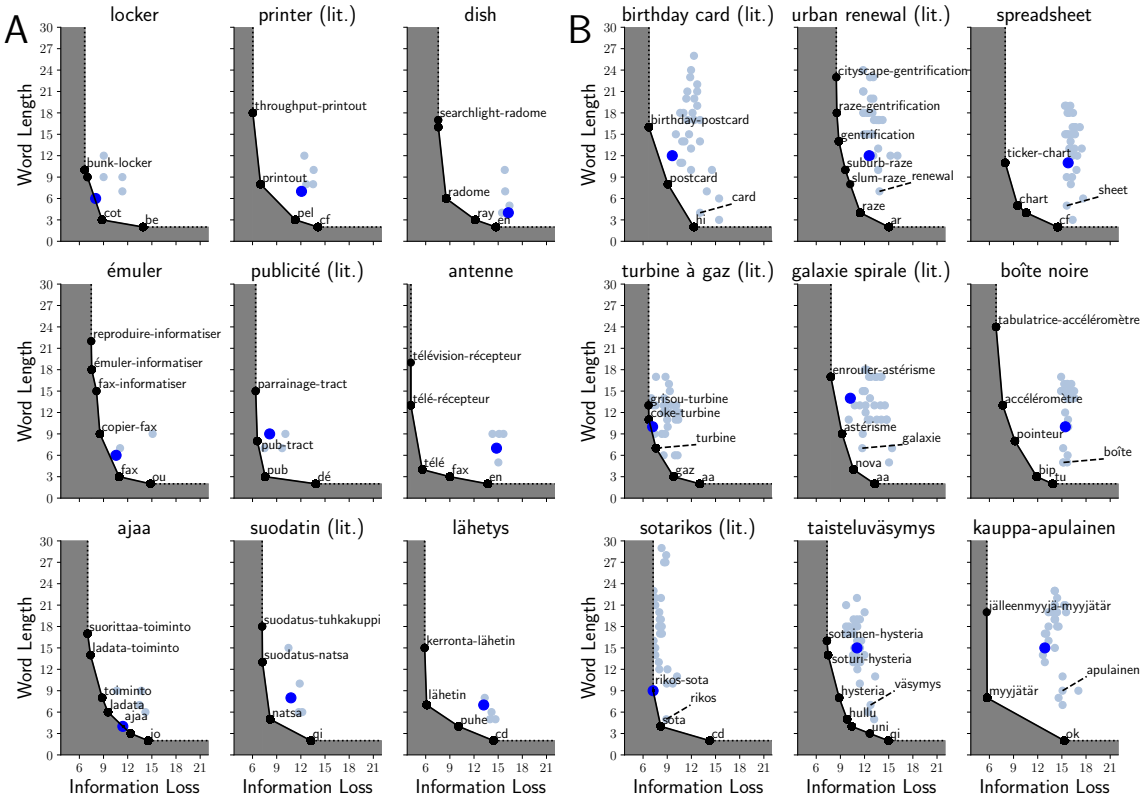


Figure 5: Item-level illustration for (A) attested reuse items and (B) attested compounds. Headers correspond to the examples in Table 1, with additional marking for literal items (lit.). Each dark blue dot corresponds to an attested form. Black dots correspond to the item-level Pareto frontier, and light blue dots correspond to the near-synonym set generated for this item; the size of markers for attested items is larger than the size of other markers for improved visibility. A sample of optimal labels and compound head words are shown as text. Note that the axes are swapped relative to Figure 2 and the x-axis is truncated so there is more space to display optimal labels.

To test this hypothesis, we leveraged the WordNet taxonomic hierarchy to classify reuse items and compounds into literal and non-literal cases. We performed a quantitative analysis that compares the efficiency loss of literal and non-literal items; we supplemented attested reuse items with additional data by using the head words of attested compounds, since

the original English WordNet does not explicitly encode literal items (Miller, 1998) (see *Materials and Methods* for details). We find that in French and Finnish, literal reuse items are significantly more efficient than non-literal items ($t(527) = 4.70$, $p < .001$; $t(508) = 6.37$, $p < .001$), and the same trend holds between literal and non-literal head words in English ($t(2793) = 23.60$, $p < .001$), French ($t(396) = 7.32$, $p < .001$), and Finnish ($t(631) = 9.40$, $p < .001$); endocentric compounds also tend to be more efficient on average in English ($t(2826) = 17.26$, $p < .001$), French ($t(407) = 3.68$, $p < .001$), and Finnish ($t(643) = 7.58$, $p < .001$). We illustrate these comparisons in *SI Appendix, Section S5.D*. These results suggest our efficient tradeoff proposal applies more strongly to labels that encode novel concepts in a literal way across both strategies.

In *SI Appendix, Section S5.E and Section S5.F*, we explore the variation in efficiency among attested reuse items and compounds in two further analyses. In the first analysis, we used taxonomic distance measures (Wu & Palmer, 1994; Leacock, Chodorow, & Miller, 1998) as a continuous version of the literal and non-literal distinction. In line with our findings above, we found that taxonomic distance measures are positively correlated with efficiency loss across all languages and across both strategies. In the second analysis, we investigated whether frequent items are closer to Pareto frontiers since this implies less total efficiency loss. We did not find that frequency differentiates variation in item-level loss. This may be due to frequency effects in lexicalization beyond the scope of our account. We return to other factors that underlie lexical evolution in *Discussion*.

We demonstrate our findings with examples in Figure 5. Along each Pareto frontier, optimal labels gradually increase in length from the shortest but uninformative words (e.g., *be*) to the most informative compounds. We observe that near-optimal items are qualitatively similar to these optimal labels (e.g., *locker* and *bunk locker*; *birthday card* and *birthday postcard*). On the other hand, suboptimal items like *dish* and *spreadsheet* tend to relate to the intended concept in a less literal way when compared to optimal labels. These examples showcase the finding that both attested reuse items and combinations are in part explained by an efficient tradeoff between word length and informativeness.

### 3.3 Strategy Comparison

In Figure 2, we also observe that reuse-based encodings and combination-based encodings tend to occupy different neighbourhoods in the space of possible encodings. To compare differences between the two strategies, we compared informativeness and word length between all attested reuse items and compounds, aggregated across intervals for each language. We show the statistics in *SI Appendix, Section S5.D*, finding that on average, attested reuse items tend to be shorter than attested compounds across all three languages, and attested compounds tend to be more informative than attested reuse items in English and French. This mirrors existing proposals that cast reuse as an economical lexicalization strategy (Štekauer, 2005; Piantadosi et al., 2012) and compounding as an informative strategy (Downing, 1977; Clark & Berman, 1984).

## 4 Discussion

We have presented evidence that word reuse and combination are shaped by competing pressures of informativeness and length minimization that affect the lexicalization of emerging concepts. We formulated this view in information-theoretic terms, and we tested our proposal using large-scale resources over history and across languages. We found that both attested reuse items and attested compounds that emerged over the past century in English, French, and Finnish are more efficient than random and near-synonym baselines, and that literal items are generally more efficient than non-literal items across both strategies.

Our work makes several contributions to efficiency-based accounts of language. First, our work establishes a new connection between efficiency-based accounts and word formation. Previous accounts have focused on form length and meanings (e.g., Mollica et al., 2021) and morpheme ordering (Hahn et al., 2022), but not the specific combination of certain morphemes as opposed to others in compositional words. In particular, one account suggests that the presence of subword structures in word forms hinders efficient length minimization (Pimentel, Nikkarinen, Mahowald, Cotterell, & Blasi, 2021). We show that in

the case of compounding, these structures may nonetheless support efficient communication in settings where the speaker and listener do not fully share the same lexicon. Second, another line of investigation uses ideas from information theory to show that existing form-meaning mappings support accurate reconstruction of intended meanings under cognitive constraints (e.g., Regier et al., 2015; Zaslavsky et al., 2018). Our work shows that information-theoretic formulations of efficiency can also account for generalizations to novel concepts not yet encoded in the lexicon. Lastly, previous work has also examined the tradeoff between simplicity and informativeness in the lexicon, but only within restricted domains (Kemp & Regier, 2012; Y. Xu et al., 2016; Zaslavsky et al., 2018; Y. Xu, Liu, & Regier, 2020; Denić, Steinert-Threlkeld, & Szymanik, 2020; Steinert-Threlkeld, 2021; Zaslavsky, Maldonado, & Culbertson, 2021; Chen, Futrell, & Mahowald, 2023). Our computational framework offers a promising venue for extending simplicity-informativeness analyses toward the broader lexicon beyond individual semantic domains.

Although we focused on the lexicalization strategies of reuse and compounding, the functional principles invoked by our theory are general and our framework can be applied to other types of word formation. For instance, derived words are also highly productive (Algeo, 1980) and previous work has suggested that they are subject to a preference for informativeness and brevity (Lieber, 2004; Marelli & Baroni, 2015). Since derived words are combinations of existing words and morphemes, one way to extend our approach might be to incorporate a representation of affix meanings and model their combination with stem words (Marelli & Baroni, 2015; Westbury & Hollis, 2019). Derivational morphemes tend to be shorter than free morphemes and might yield more efficient strategies for expressing emerging or novel concepts; for example, the derived word *physicist* is more compact than the compound *physics scientist*. As a result, derivation may sit at a midpoint between economy of production and informativeness, flanked by the strategies examined in the current study along the Pareto frontier of efficiency.

Our account suggests that speakers select for more efficient reuse items and compounds as they repeatedly communicate novel concepts to listeners, but it does not capture all

aspects of the historical evolution of the lexicon. First, while both strategies can create efficient lexical labels, our account does not explain why certain concepts are encoded via compounding but not reuse or vice versa. Consistent with Zipf's law of abbreviation (e.g., Zipf, 1949; Bentz & Ferrer Cancho, 2016; Pimentel et al., 2021; Kanwal, Smith, Culbertson, & Kirby, 2017), one possibility is that concepts with high communicative need are less likely to be expressed via compounds since they are relatively long and their morphological structure does not offer additional informativeness once they enter the listener's lexicon. Second, our account emphasizes the tradeoff between length and informativeness but there are complementary and potentially overriding factors. Recent work on language learning suggests new meanings are more easily learned if they are encoded by semantically transparent existing words (Floyd & Goldberg, 2021) or novel complex words (Brusnighan & Folk, 2012). Since informative reuse items and compounds also tend to be transparent, our results may be alternatively construed to reflect a cognitive pressure for ease of learning (c.f. Brochhagen & Boleda, 2022). On the other hand, frequent lexical innovations are more likely to be picked up by speakers (Bryden, Wright, & Jansen, 2018) and subsequently these lexical items are more likely to persist (e.g., Bybee, 1998; Pagel, Atkinson, & Meade, 2007; Lieberman, Michel, Jackson, Tang, & Nowak, 2007; Hamilton et al., 2016). Frequent lexical innovations may outcompete alternative, more efficient innovations that express the same concepts but emerged later and were used less frequently by speakers.

Our work is limited in that meaning representations and concept emergence are derived from historical data based on English. Recent work suggests that word meanings across languages show considerable variation (e.g., Thompson, Roberts, & Lupyan, 2020; Lewis, Cahill, Madnani, & Evans, 2023) and our analysis of French and Finnish items may be revisited using representations entirely based on French and Finnish resources. However, adapting these representations to a historical setting is non-trivial. For example, unlike existing crosslinguistic studies on reuse that analyze word meanings on a per-word basis (e.g., Fugikawa et al., 2023), our approach requires a large sample of existing form-sense pairs in a historical period. To apply our methodology for English to other languages, a large dataset

of historical documents and high-quality historical dictionaries (e.g., COHA and the OED) are required, which is challenging in many languages.

For simplicity, we measured the informativeness of reuse items or compounds by using representations that are derived from word sense definitions. This may have contributed to the result that literal reuse items and compounds are more efficient than non-literal cases, since the latter may reflect similarity and contiguity relations (Bloomfield, 1933; Jackendoff, 2010) that are not always encoded in sense definitions. For example, *computer memory* is a metaphorical extension of *human memory*, and the function of storage is encoded in both definitions; in contrast, visual similarities in the metaphor *computer mouse* and animal-related *mouse* are not directly encoded in their definitions. The situation for compounds is further complicated by the fact that the constituents may relate in a non-literal way (e.g., *ghost town*) or the constituents may relate in a literal way but their product may reflect non-literal extension (e.g., *white-collar*). Future work may build on recent work on polysemy and multi-modality (Brochhagen et al., 2023) and integrate it with computational models of compound interpretation (e.g., Mitchell & Lapata, 2008; Tratz & Hovy, 2010; Nakov, 2013) to further differentiate genuinely inefficient labels (e.g., homonyms and opaque compounds) from other non-literal cases.

In prior literature, word reuse and combination are typically treated as distinct areas of research. Our work provides a unified account of both lexicalization strategies by appealing to the general idea that language supports efficient communication. Previous efficiency-based approaches have focused on syntactic and semantic structures, but our work shows that the same general approach can capture the tradeoff between communicative pressures in different strategies for lexicalization. Our work therefore suggests that the view that language is shaped to support efficient communication has the potential to explain a wide spectrum of lexicalization strategies in the evolution of the lexicon.

# 5 Materials and Methods

## 5.1 Treatment of Data

We focused on five idealized historical intervals in the past century, setting $t_1 = 1900, 1920, ..., 1980$ and $t_2 = t_1 + 19$. For each interval, we instantiated the components $\mathcal{L}$ and $E^*$ using form-sense pairs in language-specific WordNets (Fellbaum, 1998; Bond & Foster, 2013), and we set $\mathcal{C}^*$ to be the concepts encoded in $E^*$. For tractability, we set the universe $\mathcal{C}$ to be the union of $\mathcal{C}^*$ and concepts encoded in $\mathcal{L}$.

We first instantiated $\mathcal{L}$ and $E^*$ using the English WordNet (Fellbaum, 1998) for each interval. Before setting up the components, we standardized word forms in the dataset to facilitate estimation of frequency and word length. We assigned form-sense pairs to $\mathcal{L}$ if their frequencies during $[t_1 - 20, t_2]$ exceeded certain thresholds, based on token frequencies from the Google Ngrams corpus (English 2020 version; Michel et al., 2011) and sense frequencies estimated using the Corpus of Historical American English (COHA; Davies, 2002) and a state-of-the-art word sense disambiguation algorithm, EWISER (Bevilacqua & Navigli, 2020). We obtained $E^*$ by first collecting reuse items and compounds with first citations in $[t_1, t_2]$ according to the Historical Thesaurus of English (Kay et al., 2017), and we then took a coarse-grained approach that assumed these items emerged at $t_1$ and have entered the lexicon by $t_2$. Lastly, we processed both $\mathcal{L}$ and $E^*$ to ensure they are disjoint. We provide a full description of this data processing pipeline in *SI Appendix, Section S2.A*.

We instantiated the same components for each interval using French and Finnish WordNets (Sagot & Fišer, 2008; Lindén & Carlson, 2010). Here, we assumed concepts encoded in the English lexicon $\mathcal{L}$ are also encoded in the French or Finnish lexicon $\mathcal{L}$ for the same interval, and we implemented $\mathcal{L}$ by labeling these concepts with forms that are attested in historical sections of the Google Ngrams corpus (French 2020 version; Michel et al., 2011) and the Newspaper and Periodical Corpus of the National Library of Finland (FNC; National Library of Finland, 2014). We also assumed the set of emerging concepts $\mathcal{C}^*$ is the same as the English set for the same interval, and we implemented $E^*$ by pairing each

$c \in \mathcal{C}^*$ with one of its French or Finnish forms if the form is in $\mathcal{L}$ or combines forms in $\mathcal{L}$. We provide a full description of this data processing pipeline in *SI Appendix, Section S2.B.*

### 5.2 Need and Production Distributions

To implement the need distribution and the production policy, we rewrote their product as $p(c, w|\mathcal{L}') = p(w|\mathcal{L}')p(c|w, \mathcal{L}')$ and separately estimated the first and second terms on the right-hand side in each language.

In the case of English, we estimated the first term using token frequencies in historical texts that appeared during $[t_1, t_2]$ in the English Google Ngrams corpus (Michel et al., 2011). We defined the first term as $p(w|\mathcal{L}') \propto f_w$, where $f_w$ is the frequency of the form $w$. We estimated the second term in two steps. If $w$ is a unigram, we reused sense frequencies based on text that appeared during $[t_1, t_2]$ in COHA (Davies, 2002), so that given sense $c$ with frequency $f_{c,w}$, we have $p(c|w, \mathcal{L}') \propto f_{c,w}$. Otherwise, we used a uniform distribution over concepts in $\mathcal{L}'$ that are labeled by $w$ since our sense disambiguation method did not apply to open compounds. Lastly, we applied add-one smoothing to form-concept pairs in $\mathcal{L}'$.

In the case of French and Finnish, we used the same method to estimate the first term, except we used historical text in the French Google Ngrams corpus (Michel et al., 2011) and the FNC (National Library of Finland, 2014). We estimated the second term based on how an English speaker would infer the distribution via Bayes rule. Specifically, we first assumed that the speaker's prior $p_e(c)$ is the need probability of $c$ estimated from English text during $[t_1, t_2]$ and their likelihood is uniform over all labels of $c$ in the language-specific lexicon $\mathcal{L}'$. We then defined the second term as the posterior, which is given by $p(c|w, \mathcal{L}') \propto p_e(c)$ if $(c, w) \in \mathcal{L}'$ and zero otherwise. We also applied add-one smoothing to form-concept pairs in $\mathcal{L}'$.

### 5.3 Prototype Model

Here we specify our variant of prototype-based categorization models in Equation 1. We first represented each $c \in \mathcal{C}$ by embedding its definition using a state-of-the-art sentence encoder, Sentence-BERT (Reimers & Gurevych, 2019), yielding a semantic vector for each concept $c$. Although WordNet definitions were compiled during the past century (Fellbaum, 1998), the encoder was trained on contemporary natural language data, and we view these vectors and the semantic space as an approximation of listener representation of concepts in our target intervals. Throughout this study, we used cosine distance for $d(\cdot, \cdot)$ following Reimers and Gurevych (2019).

Since the word $w$ is either an existing word in the listener lexicon $\mathcal{L}$ or a combination of existing words, we defined the prototype $q_{w,\mathcal{L}}$ in two parts. If $w$ is in $\mathcal{L}$, we extended the model in Reed (1972) and defined the prototype as a weighted average of category exemplars, i.e., the concepts encoded by $w$ in $\mathcal{L}$; alternatively, if $w$ is a combination of N constituents, we used the additive composition function (Mitchell & Lapata, 2008) to recursively define a composite prototype that combines the prototypes of its constituents (e.g., Smith, Osherson, Rips, & Keane, 1988):

$$q_{w,\mathcal{L}} = \begin{cases} \sum_c p(c|w,\mathcal{L})c & \text{if } w \in \mathcal{L} \\ \sum_i q_{w_i,\mathcal{L}} & \text{else if } w = w_1 ... w_{\mathrm{N}} \in \mathcal{L}^{\mathrm{N}} \end{cases} \tag{6}$$

Here the expression $w \in \mathcal{L}$ implies $(c, w) \in \mathcal{L}$ for at least one $c \in \mathcal{C}$, and $p(c|w,\mathcal{L})$ was estimated from the relative frequencies of items in $\mathcal{L}$. In *SI Appendix, Section S3*, we validated our embedding space and construction of prototypes using datasets of English word similarity and compound meaning predictability.

Intuitively, the parameter $\gamma$ simulates how much the listener prefers to infer the most transparent interpretation of a word. In *SI Appendix, Section S3.C*, we show that the average information loss incurred over communicating existing concepts (i.e., part of the omitted term in Equation 5) is minimized when $\gamma \in [15, 20]$. This suggests a reasonable range for $\gamma$ is $(0, 15]$ since if $\gamma$ is too low then the listener does not distinguish among

concepts, and if $\gamma$ is too high then the listener will incur high information loss whenever the word form expresses an extended sense. For this reason, in the main text we set $\gamma = 10$. We note that our argument is based on the information loss of existing words, and we leave more fine-grained modeling of the listener distribution for compounds for future work.

### 5.4 Estimating the Pareto Frontier

For each tradeoff parameter $\beta = 0, 0.01, ..., 10$, we computed the optimal encoding $E^*_\beta$ that encodes concepts in $\mathcal{C}^*$ and minimizes Equation 5. We assume that need probabilities are constant with respect to how concepts are encoded. In this case, each novel concept independently contributes to the overall cost in Equation 5, and this optimization is equivalent to finding a form $w$ for each $c \in \mathcal{C}^*$ such that $w$ jointly minimizes word length and surprisal for a certain $\beta$. That is, we want to optimize the following item-level objective over existing words and possible combinations:

$$L_\beta[w|c, \mathcal{L}] = h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w) \tag{7}$$

For tractability, we greedily selected a first string $u \in \mathcal{L}$ that optimizes Equation 7, and we greedily selected a second string $u' \in \mathcal{L}$ or the empty string such that the concatenation of the selected strings minimizes Equation 7. The final concatenation is the approximately optimal form for $c$ and this form-concept pair is added to $E^*_\beta$.

### 5.5 Baseline Encodings

We constructed near-synonym and random encodings as baselines for every attested encoding $E^*$. To construct a near-synonym encoding, we started by constructing a near-synonym set for every form-sense pair in $E^*$. Suppose that the form contains modifier $w$ and syntactic head $u$; we assumed all English and Finnish compounds are right-headed and all French compounds are left-headed due to their relative prevalence (Lieber, 2011; Hyvärinen, 2019; Van Goethem & Amiot, 2019). For the constituent $w$, we selected the top $k = 5$ forms, among all existing forms $x \in \mathcal{L}$ that are closest to $w$ in terms of the cosine distance between

$q_{w,\mathcal{L}}$ and $q_{x,\mathcal{L}}$ but are not antonyms of $w$ in WordNet. We repeated this procedure for $u$, except we also made sure the possible word classes of the generated constituents overlap with the possible word classes of $u$. The near-synonym set is defined as $\{xy : x \in S_u, y \in S_w\} \cup S_u$, where $S_w$ and $S_u$ are the forms generated for $w$ and $u$, respectively. We then created a near-synonym encoding by replacing the form in each attested form-sense pair with a random sample from its near-synonym set. We constructed random encodings for each $E^*$ similarly by replacing every attested label with a form uniformly sampled from forms in $\mathcal{L}$ and their combinations. For every $E^*$, we created $100,000$ near-synonym and random encodings, respectively.

### 5.6 Efficiency Loss

We compared each attested encoding $E^*$ against the generated baselines by computing their efficiency loss relative to the Pareto frontier following Zaslavsky et al. (2018). Given $E^*_\beta$ for $\beta = 0, 0.01, ..., 10$, the efficiency loss of the encoding $E^*$ or its corresponding baselines is defined as follows:

$$\epsilon = \min_{\beta} \Big( L_\beta[E^*|\mathcal{L}] - L_\beta[E^*_\beta|\mathcal{L}] \Big) \tag{8}$$

This measures the deviation of the encoding $E^*$ from optimality, or its deviation from the lowest possible amount of information loss given a specific value of average length.

### 5.7 Literal and Non-Literal Items

We classified a reuse item $(c, w)$ as a literal item if the novel concept $c$ is a hyponym of an existing sense of $w$, and otherwise we classified the pair as a non-literal item. We classified a compound item $(c, w)$ as an endocentric compound if $c$ and the head word constitute a literal item, and otherwise we classified the item as an exocentric compound; we made the same assumptions on head positions as in our construction of near-synonyms. Since Princeton WordNet was made to avoid linking a sense and its hyponyms to the same word (Miller, 1998), there are few literal reuse items for English (N = 3) and we supplemented attested reuse items for all languages with additional data by replacing the

compound in each attested compound-sense pair with its head word. A small number of endocentric compounds with head positions different from our assumption were not used in data augmentation. We provide hypernyms of literal items from Table 1 in *SI Appendix, Section S5.B.*

### 5.8 Data Availability

All data and code used in analyses are available at `https://osf.io/dmgh6`

## 6 Acknowledgments

We thank Frank Mollica, Gemma Boleda, Guillaume Thomas, and Michael Hahn for their feedback on the manuscript. We also thank the editor and the reviewers for their constructive feedback. This research is funded partly by U of T-UoM International Research Training Group program. A.X. and Y.X. are supported by Natural Sciences and Engineering Research Council of Canada Discovery Grant RGPIN-2018-05872 and Ontario Early Researcher Award #ER19-15-050. C.K. is supported by Australian Research Council Grant FT190100200.

## References


Algeo, J. (1980). Where do all the new words come from? *American Speech*, *55*(4), 264–277.

Bentz, C., & Ferrer Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In C. Bentz, G. Jäger, & I. Yanovich (Eds.), *Proceedings of the leiden workshop on capturing phylogenetic algorithms for linguistics* (pp. 1–4).

Bevilacqua, M., & Navigli, R. (2020, July). Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2854–2864). Online: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/2020.acl-main.255`

Blank, A. (2003). Words and concepts in time: Towards diachronic cognitive onomasiology. *Trends In Linguistics Studies And Monographs*, *143*, 37–66.

Bloomfield, L. (1933). *Language*. Holt.

Bond, F., & Foster, R. (2013). Linking and extending an open multilingual wordnet. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1352–1362). Association for Computational Linguistics.

Brinton, L. J., & Traugott, E. C. (2005). *Lexicalization and language change*. Cambridge University Press.

Brochhagen, T., & Boleda, G. (2022). When do languages use the same word for different meanings? the goldilocks principle in colexification. *Cognition*, *226*, 105179.

Brochhagen, T., Boleda, G., Gualdoni, E., & Xu, Y. (2023). From language development to language evolution: A unified view of human lexical creativity. *Science*, *381*(6656), 431–436.

Brusnighan, S. M., & Folk, J. R. (2012). Combining contextual and morphemic cues is beneficial during incidental vocabulary acquisition: Semantic transparency in novel

compound word processing. *Reading Research Quarterly*, *47*(2), 172–190.

Bryden, J., Wright, S. P., & Jansen, V. A. (2018). How humans transmit language: horizontal transmission matches word frequencies among peers on twitter. *Journal of The Royal Society Interface*, *15*(139), 20170738.

Bybee, J. (1998). The emergent lexicon. In M. C. Gruber, D. Higgins, & K. S. Olson (Eds.), *Chicago linguistic society* (Vol. 34, pp. 421–435). University of Chicago.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive science*, *41*(4), 892–923.

Chen, S., Futrell, R., & Mahowald, K. (2023). An information-theoretic approach to the typology of spatial demonstratives. *Cognition*, *240*, 105505.

Clark, E. V., & Berman, R. A. (1984). Structure and use in the acquisition of word formation. *Language*, 542–590.

Costello, F. J., & Keane, M. T. (2000). Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science*, *24*(2), 299–349.

Davies, M. (2002). *The corpus of historical American English (COHA): 400 million words, 1810-2009*. Brigham Young University.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. W.W. Norton & Company.

Del Tredici, M., & Fernández, R. (2018). The road to success: Assessing the fate of linguistic innovations in online communities. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 1591–1603). Association for Computational Linguistics.

Denić, M., Steinert-Threlkeld, S., & Szymanik, J. (2020). Complexity/informativeness trade-off in the domain of indefinite pronouns. In J. Rhyne, K. Lamp, N. Dreier, & C. Kwon (Eds.), *Semantics and linguistic theory* (pp. 166–184). Linguistic Society of America.

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, 810–842.

Dressler, W. U. (2005). Word-formation in natural morphology. In P. Štekauer & R. Lieber (Eds.), *Handbook of word-formation* (pp. 267–284). Springer.

Fellbaum, C. (1998). *Wordnet: An electronic lexical database*. MIT press.

Floyd, S., & Goldberg, A. E. (2021). Children make use of relationships across meanings in word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(1), 29.

Fugikawa, O., Hayman, O., Liu, R., Yu, L., Brochhagen, T., & Xu, Y. (2023). A computational analysis of crosslinguistic regularity in semantic change. *Frontiers in Communication*, *8*, 1136338.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*, *23*(5), 389–407.

Günther, F., & Marelli, M. (2016). Understanding karma police: The perceived plausibility of noun compounds as predicted by distributional models of semantic representation. *PloS one*, *11*(10), e0163200.

Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, *117*(5), 2347–2353.

Hahn, M., Mathew, R., & Degen, J. (2022, June). Morpheme ordering across languages reflects optimization for processing efficiency. *Open Mind: Discoveries in Cognitive Science*, *5*, 208-232. Retrieved from `https://direct.mit.edu/opmi/article/doi/10.1162/opmi_a_00051/109033/Morpheme-Ordering-Across-Languages-Reflects` doi: 10.1162/opmi_a_00051

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016, August). Diachronic word embeddings reveal statistical laws of semantic change. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1489–1501). Berlin, Germany: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P16-1141` doi: 10.18653/

v1/P16-1141

Hyvärinen, I. (2019). Compounds and multi-word expressions in Finnish. In B. Schlücker (Ed.), *Complex lexical units* (pp. 307–336). De Gruyter.

Jackendoff, R. (2010). *Meaning and the lexicon*. Oxford University Press Oxford.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Kay, C., Roberts, J., Samuels, M., & Wotherspoon, I. (2017). *The Historical Thesaurus of English, version 4.21*. Glasgow, UK: University of Glasgow. Retrieved from `http://historicalthesaurus.arts.gla.ac.uk/`

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

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

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

Labov, W. (2011). *Principles of linguistic change, volume 3: Cognitive and cultural factors* (Vol. 3). John Wiley & Sons.

Leacock, C., Chodorow, M., & Miller, G. A. (1998). Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, *24*(1), 147–165.

Levi, J. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.

Levin, B., Glass, L., & Jurafsky, D. (2019). Systematicity in the semantics of noun compounds: The role of artifacts vs. natural kinds. *Linguistics*, *57*(3), 429–471.

Lewis, M., Cahill, A., Madnani, N., & Evans, J. (2023). Local similarity and global variability characterize the semantic space of human languages. *Proceedings of the National Academy of Sciences*, *120*(51), e2300986120.

Lieber, R. (1983). Argument linking and compounds in english. *Linguistic inquiry*, *14*(2),

251–285.

Lieber, R. (2004). *Morphology and lexical semantics*. Cambridge University Press.

Lieber, R. (2011). IE, Germanic: English. In R. Lieber & P. Štekauer (Eds.), *The oxford handbook of compounding* (pp. 357–369). Oxford University Press.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713–716.

Lindén, K., & Carlson, L. (2010). Finnwordnet-wordnet på finska via översättning. *LexicoNordica–Nordic Journal of Lexicography*, *17*, 119–140.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). Wiley.

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive science*, *42*(8), 3116–3134.

Marchand, H. (1969). *The categories and types of present-day English word formation: a synchronic diachronic approach*. München, Germany: C.H. Beck'sche Verlagsbuchhandlung.

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, *122*(3), 485.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, G. B., ... Norvig, P. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

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

Miller, G. A. (1998). Nouns in wordnet. In C. Fellbaum (Ed.), *Wordnet: An electronic lexical database* (pp. 23–46). MIT press.

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of linguistics*, *21*(2), 339–384.

Mitchell, J., & Lapata, M. (2008, June). Vector-based models of semantic composition. In

J. D. Moore, S. Teufel, J. Allan, & S. Furui (Eds.), *Proceedings of acl-08: Hlt* (pp. 236–244). Columbus, Ohio: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P08-1028

Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., & Kemp, C. (2021). The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, *118*(49).

Nakov, P. (2013). On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, *19*(3), 291–330.

National Library of Finland. (2014). *The Finnish N-grams 1820-2000 of the Newspaper and Periodical Corpus of the National Library of Finland* [data set]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2014073038 (Avaiable at www.kielipankki.fi/download/FNC1/. Accessed December 21, 2023.)

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

Oxford University Press. (2023). *mouse (n.), sense I.13.* Retrieved from https://www.oed.com/dictionary/mouse_n?tab=meaning_and_use (Accessed on 2024-03-07)

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, *449*(7163), 717–720.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., & Blasi, D. (2021, June). How (non-)optimal is the lexicon? In K. Toutanova et al. (Eds.), *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4426–4438). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.350 doi: 10.18653/v1/2021.naacl-main.350

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, *13*(4), 707–727.

Pugacheva, V., & Günther, F. (2024). Lexical choice and word formation in a taboo game paradigm. *Journal of Memory and Language*, *135*, 104477.

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, *115*(10), 2323–2328.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019a). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019b). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), 20191262.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O. Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Wiley Online Library.

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D19-1410` doi: 10.18653/v1/D19-1410

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *104*(3), 192.

Sagot, B., & Fišer, D. (2008). Building a free French wordnet from multilingual resources. In A. Oltramari, L. Prévot, C.-R. Huang, P. Buitelaar, & P. Vossen (Eds.), *Ontolex.* European Language Resources Association.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A

selective modification model. *Cognitive science*, *12*(4), 485–527.

Srinivasan, M., & Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, *157*, 124–152.

Steinert-Threlkeld, S. (2021). Quantifiers in natural language: Efficient communication and degrees of semantic universals. *Entropy*, *23*(10), 1335.

Štekauer, P. (2005). Onomasiological approach to word-formation. In P. Štekauer & R. Lieber (Eds.), *Handbook of word-formation* (pp. 207–232). Springer.

Štekauer, P., & Lieber, R. (2005). *Handbook of word-formation* (Vol. 64). Springer Science & Business Media.

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, *4*(10), 1029–1038.

Tratz, S., & Hovy, E. (2010). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In J. Hajič, S. Carberry, S. Clark, & J. Nivre (Eds.), *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 678–687). Association for Computational Linguistics.

Traugott, E. C., & Dasher, R. B. (2001). *Regularity in semantic change*. Cambridge University Press. doi: 10.1017/CBO9780511486500

Van Goethem, K., & Amiot, D. (2019). Compounds and multi-word expressions in French. In B. Schlücker (Ed.), *Complex lexical units* (pp. 127–152). De Gruyter.

Vecchi, E. M., Marelli, M., Zamparelli, R., & Baroni, M. (2017). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive science*, *41*(1), 102–136.

Weinreich, U., Labov, W., & Herzog, M. (1968). Empirical foundations for a theory of language change. In W. P. Lehmann & Y. Malkiel (Eds.), *Directions for historical linguistics*. University of Texas Press.

Westbury, C., & Hollis, G. (2019). Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector

averaging. *Behavior Research Methods*, *51*(3), 1371–1398.

Williams, J. M. (1976). Synaesthetic adjectives: A possible law of semantic change. *Language*, 461–478.

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

Xu, A., Kemp, C., Frermann, L., & Xu, Y. (2023). Predicting strategy choice in word formation: A case study of reuse and compounding. In M. Goldwater, F. Anggoro, B. Hayes, & D. Ong (Eds.), *Proceedings of the 45th annual meeting of the cognitive science society.* Cognitive Science Society.

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition*, *201*, 104280.

Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind: Discoveries in Cognitive Science*, *4*, 57–70.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive science*, *40*(8), 2081–2094.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.

Zaslavsky, N., Maldonado, M., & Culbertson, J. (2021). Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the annual meeting of the cognitive science society* (Vol. 43). Cognitive Science Society.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Ravenio Books.