Supporting Information for

Word reuse and combination support efficient communication of emerging concepts

Aotao Xu, Charles Kemp, Lea Frermann, and Yang Xu

Aotao (John) Xu
E-mail: a26xu@cs.toronto.edu

This PDF file includes:

Supporting Information Text

1. Additional Information on Computational Formulation

We provide additional information on the computational formulation of our theoretical proposal. We first provide a formal derivation of the objective function we used to specify the Pareto frontier. We then provide additional discussion on previous efficiency-based accounts of the lexicon.

A. Derivation of the Objective Function. Here we provide a formal derivation of Equations 2-5 in the main text. Recall that the expanded lexicon $\mathcal{L}'$ is the union of the existing lexicon $\mathcal{L}$ and an encoding of emerging concepts E^*. Now, consider the case in which the same speaker interacts once with each of n distinct listeners, for a large number n. In the i-th interaction, the speaker samples an intended concept from the need distribution $C_i \sim p(c|\mathcal{L}')$ and a word form from the production policy $W_i \sim p(w|c, \mathcal{L}')$; if $W_i = w$, then the i-th listener approximates the speaker's mental representation using the distribution $\hat{m}_w^{(i)}$. We assume the listeners are distinct but have the same listener distribution, i.e., for every w, $\hat{m}_w^{(1)}, \hat{m}_w^{(2)}, ..., \hat{m}_w^{(n)} \overset{\text{iid}}{\sim} \hat{m}_{w,\mathcal{L}}$. We will consider a large sample of these interactions, represented by the sequence $(C_1, W_1), (C_2, W_2), ..., (C_n, W_n)$.

A.1. Average Word Length. First, we relate the average length over the above interactions to the expectation of length. Let L_{length} be the average length over these interactions. We define this random variable by first multiplying the length of w with the fraction of times the form-concept pair (w, c) shows up in the sample of interactions, and then summing over all possible pairs:

$$L_{\text{length}} = \sum_{c,w} l(w) \cdot \left[\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[W_i = w, C_i = c] \right] = \sum_{c,w} l(w) \cdot p(c, w|\mathcal{L}') \quad [1]$$

The last step shows the average length over many interactions converges to the expected length with probability 1, which follows from the strong law of large numbers and a property of the indicator function. We thus define average length as the expectation of length.

A.2. Average Information Loss. Next, we relate the average information loss over the above interactions to the expectation of KL divergence. Let L_{error} be the average information loss over these interactions. Similar to ref. 1, because we assumed speaker distributions have perfect accuracy, the information loss incurred in the i-th interaction reduces to the surprisal $h(\hat{m}_w^{(i)}(c))$ if c and w were selected by the speaker. Since listener distributions are independent and identically distributed, the information loss incurred in any interaction is $h(\hat{m}_{w,\mathcal{L}}(c))$ if c and w were selected by the speaker. Thus similar to above, we define the random variable L_{error} by first multiplying the surprisal of c given w with the fraction of times the form-concept pair (w, c) shows up in the sample of interactions, and then summing over all possible pairs:

$$L_{\text{error}} = \sum_{c,w} h(\hat{m}_{w,\mathcal{L}}(c)) \cdot \left[\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[W_i = w, C_i = c] \right] = \sum_{c,w} h(\hat{m}_{w,\mathcal{L}}(c)) \cdot p(c, w|\mathcal{L}') \quad [2]$$

The last step shows the average information loss over many interactions converges to the expected surprisal with probability 1, which follows from the same steps used in Equation 1. We thus define average information loss as the expectation of KL divergence (which reduces to expected surprisal).

A.3. Objective Function. To assess the role of communicative efficiency in shaping the encoding E^*, we combine and simplify the average length and information loss as defined above. First, we make two simplifying assumptions: 1) the frequency of concepts encoded in the existing lexicon during the target interval between t_1 and t_2 remains proportional to their frequency at time t_1, the point at which all speakers use the existing lexicon by assumption; and 2) the production policy for these concepts during the target interval remains the same as the policy at time t_1. These assumptions imply that for every concept-form pair $(c, w) \in \mathcal{L}$, we have $p(c|\mathcal{L}') \propto p(c|\mathcal{L})$ and $p(w|c, \mathcal{L}') = p(w|c, \mathcal{L})$. The objective function for the encoding E^* can now be obtained as follows:

$$\begin{aligned}
L_\beta[E^*|\mathcal{L}] &= \mathbb{E}[D(M||\hat{M})|\mathcal{L}', \mathcal{L}] + \beta\mathbb{E}[l(W)|\mathcal{L}'] && [3]\\
&= \sum_{(c,w)\in\mathcal{L}'} p(c, w|\mathcal{L}') \cdot (h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w)) && [4]\\
&= \sum_{(c,w)\in\mathcal{L}} p(c, w|\mathcal{L}') \cdot (h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w)) + \sum_{(c,w)\in E^*} p(c, w|\mathcal{L}') \cdot (h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w)) && [5]\\
&\propto \sum_{(c,w)\in E^*} p(c, w|\mathcal{L}') \cdot (h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w)) && [6]
\end{aligned}$$

where the second last line follows from $\mathcal{L}' = \mathcal{L} \cup E^*$ and the fact that $\mathcal{L}$ and E^* are disjoint, and the last line follows from the fact that the first sum in the previous line is constant in terms of E^* after applying our assumptions on the need distribution and the speaker's production policy.

B. Additional Discussion. We have hypothesized that word reuse and combination reflect a tradeoff between speaker effort and information loss, and that both attested reuse items and combinations are constrained by competing pressures to minimize these communicative costs. We formalized this idea by extending previous models of communication (e.g., refs. 1, 2) to the setting in which new lexical items spread among language users. One consideration is that existing frameworks or simpler variants are sufficient for explaining both attested reuse items and compounds, such that it is less parsimonious to use an extended framework. In particular, under an existing framework, ref. 3 has shown that a pressure for informativeness constrains the products of word meaning extension in the domain of container names, and it may be likely that this finding generalizes across domains. Here we briefly discuss why existing frameworks and their underlying information-theoretic principles cannot also account for compound words or structures at the subword level in general. Our discussion will focus on how subwords relate to expected word length and informativeness under existing frameworks.

B.1. Word length. We consider the two-stage approach in ref. 4 which decomposes the speaker's production policy into a meaning encoder and a form encoder and mirrors earlier work in lossy data compression (e.g., ref. 5). Under this approach, the meaning encoder maps each intended message to a distribution over a finite list of indexes, and the form encoder maps each index to a unique string. Given a fixed meaning encoder, the optimal form encoder that minimizes expected word length can be obtained from classic results in lossless data compression (6), and the set of strings assigned by this encoder is likely to be shorter on average and less structured than an alternate set of strings that contains a high amount of subword structure (7). However, under this existing approach, an optimal form encoder does not impose systematicity between subwords and meaning even at the item level. Crucially, the optimal string assignment to each index only depends on the length of the string and the probability of the index. Observe that this condition does not impose systematicity because exchanging any two forms of the same length does not affect expected word length, even if these forms are substrings of longer forms whose index assignment remained unchanged. For example, the strings *fly* and *man* have the same length and are frequent compound head words; suppose that now *fly* expresses the meaning of *man* and vice versa, but the meanings of all other words remain the same. This change violates attested systematicity in endocentric compounds headed by these words, e.g., *fireman* becomes a *fly* and *firefly* becomes a *man*, but the expected word length of the lexicon remains the same as before.

An important feature of existing frameworks based on variable-length lossless data compression is the use of entropy as the lower bound on expected length (4, 7). This implicitly assumes that word forms are uniquely decodable, which does not hold over the whole lexicon in general (8). This is apparent in the case of compound words: for example, there are two ways to decode the string *black sea*, since it can be parsed as a single named entity but also as two separate words *black* and *sea*. In speech, distinctions in word stress are generally helpful for disambiguating these cases in certain languages (e.g., English; ref. 9) but not others (e.g., French; ref. 10).

B.2. Informativeness. In efficiency-based accounts of the lexicon that invoke the point-to-point model of communication, the expected KL divergence between speaker and listener distributions over world states has become the standard way to measure information loss (1–4, 11–13). The exact implementation of speaker and listener distributions differs across studies, but a common assumption is that the listener has full access to the speaker's production policy (e.g., in ref. 14, the listener distribution is a function of the shared lexicon as well as the need distribution used by the speaker). Given this assumption and KL divergence-based information loss, optimal reconstruction can always be obtained by only combining the meanings of the speaker's utterance that exist in the shared lexicon (2). This implies under this general existing approach, subword information in word forms is irrelevant for accurately reconstructing messages intended by speakers.

Differing from these existing accounts, in our approach the speaker may use word forms that do not exist in the lexicon of the listener. The listener distribution for new word forms depends on data-driven modelling decisions and is less general than listener distributions that only depend on the mathematical formulation of the communication model itself (e.g., ref. 2). Nonetheless, experimental studies on language evolution (e.g., refs. 15, 16) and language production (e.g., refs. 17, 18) show that listeners are able to accurately infer intended messages using labels that combine subwords in an informative way, suggesting that plausible listener distributions for new word forms will be dependent on subword level information.

2. Historical Data

Here we provide a full description of our data processing pipeline for instantiating our scenario of lexical evolution. We first describe the pipeline for English analyses, which builds towards and is followed by a description of the pipeline for French and Finnish analyses. Lastly, we provide an evaluation of the historical sense frequencies used in this study.

A. Processing of English Data. Here we describe the pipeline for processing the English data. It involves the following steps: 1) standardizing word forms in WordNet (19); 2) estimating the existing lexicon $\mathcal{L}$; 3) identifying emerging reuse items; 4) identifying emerging compounds; and 5) compiling $\mathcal{L}$ and the attested encoding E^*.

A.1. Standardizing Word Forms. We standardized the space of English word forms to facilitate measuring word length and frequency. We defined a standard word form as a lemma (since WordNet only provides lemma forms) that contains only alphabetical, lowercase characters or a delimiter in the form of dash or hyphen and does not have alternate spellings from which it is vastly dissimilar. To satisfy the second criterion, we removed any form-sense pair such that 1) the form is a synonym of a proper noun or an initialism, 2) the form is a number in orthographic form, or 3) the sense definition includes the words *slang* or *informal term*. We also removed forms that can be further lemmatized using the WordNet lemmatizer in NLTK (20).

A.2. Existing Lexicon. To estimate the lexicon $\mathcal{L}$ existing at time t_1, we first estimated the subset of form-sense pairs that existed during the interval $[t_1, t_1 + 19]$ and another subset that existed during the interval $[t_1 - 20, t_1 - 1]$, and then we set $\mathcal{L}$ to be the intersection between the subsets. We estimated if a form-sense pair existed during interval $[t, t']$ by checking if their frequencies in historical corpora exceed certain thresholds. We first obtained sense frequencies from the Corpus of Historical American English (COHA; ref. 21). From documents published between t and t', we extracted contiguous sentences that contain at least one ambiguous word and identified the intended sense of each ambiguous word using the word sense disambiguation algorithm EWISER (22). In *SI Appendix, Section 2.C*, we verify that the resulting sense frequencies reflect historical changes in word meaning. Second, to more accurately estimate the frequencies of historically rare words, we used unigram frequencies provided by the Google Ngrams corpus (English 2020 version; ref. 23). Since infrequent word forms and senses are more sensitive to noise, we included a form-sense pair in $\mathcal{L}$ only if 1) the unigram frequency of the form exceeds $\tau_{\text{token}} = 10$ and 2) the ratio between sense and unigram frequencies, i.e., the proportion of times the sense is expressed by the word, exceeds $\tau_{\text{sense}} = 0.1$ during the interval.

A.3. Attested Encoding via Combination. As mentioned in the main text, we determined the set of emerging concepts and its attested encoding for an interval $[t_1, t_2]$ by collecting attested reuse items and compounds that contain a first citation in the interval, and here we describe how these items and their first citations were collected. The first step was to identify a list of WordNet lemmas that are compound words. We identified a lemma as a closed compound if it is an entry that is singular and correctly parsed in the Large Database of English Compounds (LADEC; ref. 24) or if it is etymologically formed via compounding in an extraction of Wiktionary (25); both sources also provided us with their correct constituent segmentation. To identify open and hyphenated compounds, we took an inclusive approach by extracting all WordNet lemmas that can be parsed into two lemmas by underscore or hyphen. We only extracted compounds that have exactly two constituents for tractability, and we removed lemmas in which the second constituent is a preposition so that the compound head is usually on the right. Given this list of compounds, the second step is to determine the first citation of form-sense pairs in which the form is a listed compound. We used the Historical Thesaurus of English (HTE; ref. 26) to determine the first citation of the compounds we collected, and if a compound form w has first citation at year t, we assumed all form-sense pairs containing w also has first citation at t.

A.4. Attested Encoding via Reuse. As in the description of compound-based encodings, here we focus on describing how we collected reuse items and their first citations. We used two methods to collect this dataset. In the first method, we manually annotated the first citations of $1,331$ form-sense pairs in WordNet by checking the HTE and the associated sense definitions in the Oxford English Dictionary (OED). In the second method, we used first citations of compounds to automatically estimate the first citations of compound constituents that were reused to express the senses of compounds (e.g., *cell* and *cellphone*). In the following, we first describe how we obtained the manually annotated form-sense pairs, and then we describe the second automatic method.

To select the subset of form-sense pairs to annotate, we used two automatic methods to shortlist pairs that are likely to have emerged in the past century and are expressed by polysemous words. First, we used the HTE to find words such that all of their HTE senses first appeared in the 20th century, and we shortlisted these words along with all of their WordNet senses. Second, we used the HTE to find words that have at least one HTE sense with a first citation before 1900 and at least one with a first citation after 1900. Since the second list is large, we pruned the words in the list by checking if the definitions of their senses overlap with culturally salient keywords useful for novel word sense detection (27). Each sense definition is processed by removing non-alphabetical characters, converting to lowercase, and lemmatizing its tokens using the WordNet lemmatizer from NLTK (20). We kept a form-sense pair in the list only if the processed sense definition contains one of the keywords, which yields the second shortlist.

Since ref. 27 focused on senses emerging after year 2000, we selected a new set of keywords that are culturally salient during the past century. We selected these keywords semi-automatically by first selecting keyword candidates based on frequency information in COHA. Specifically, we first selected candidates using two kinds of frequency information from five subcorpora of COHA (i.e., the full corpus plus the four genres consisting of fiction, magazine, news, and non-fiction). For every subcorpus

of COHA, we first ranked all lemmas that appear in both the pre-1900 and post-1900 parts of the subcorpus by changes in their frequency. That is, let $f_h(w)$ and $f_m(w)$ be the frequencies of lemma w in these two parts respectively; we followed ref. 28 and ranked the lemmas by the ratio of their modern and historical frequencies, $f_m(w)/f_h(w)$. We then ranked all lemmas that only appear in the post-1900 part of the subcorpus by their raw frequency. For both ranked lists, we selected the top 100 lemmas. We applied this method to all of COHA and its four genres, yielding in total ten 100-word lists. Given these word lists, we then manually selected words that are 1) related to technological or cultural innovations in the past century and 2) likely to exist in the definitions of a large number of word senses. We also manually augmented the selected words with additional ones that are etymologically related (e.g., *electronic* and *electric*). In total, we identified 32 keywords, which are shown in Table S1.

aircraft, airport, basketball, broadcast, broadcasting, car, chocolate, cigarette, cinema, computer, data, electric, electrical, electronic, film, hockey, internet, jazz, motor, motorcycle, online, petro, petroleum, phone, pilot, radar, radio, rocket, software, spaceship, television, video

Table S1. Set of cultural keywords

Finally, we used the first citations of English compounds to help supplement our dataset of English reuse items. For each form-sense pair in this set of compounds, we checked if the compound has a synonym in WordNet that is also a constituent of the compound (e.g., *line* and *assembly line*). We added these synonymous constituents to our set of reuse items and assumed their first citation is the same as the compound. Note that even though a word can be first formed via compounding and later reused in the same interval, our setting does not allow the same word form to exist in both $\mathcal{L}$ and $\mathcal{L}'$, and for simplicity we did not include emerging reuse items that are also emerging compounds in the same interval.

A.5. Additional Processing. After obtaining the existing lexicon $\mathcal{L}$ using frequency statistics from historical corpora and obtaining the attested encoding of novel concepts E^* by compiling reuse items or compounds with a first citation in $[t_1, t_2]$, we synced the encoding E^* with the lexicon $\mathcal{L}$ so that each novel item communicates a novel concept via reuse or combination. Specifically, we first updated the lexicon $\mathcal{L}$ by removing every form-concept pair that is a reuse item or compound in the encoding E^*. We then updated the encoding E^* by removing form-concept pairs if 1) the concept is encoded in the updated $\mathcal{L}$, or 2) the form itself or one of its constituents does not exist in the updated $\mathcal{L}$. The descriptive statistics for each final existing lexicon and attested encoding are summarized in Table S2. Note that the smaller sample sizes of the final interval are caused by the sparsity of post-1980 entries in the HTE.

Interval	# existing forms	# existing senses	# reuse	# combination
1900+	47,572	43,649	136	766
1920+	49,876	45,658	127	914
1940+	51,353	47,194	142	627
1960+	52,317	48,070	92	474
1980+	52,957	48,555	21	47

Table S2. Number of existing forms and senses, and novel items across intervals for English

B. Processing of French and Finnish Data. Here we describe the data processing pipeline for French and Finnish, involving two steps: 1) estimating the existing lexicon; and 2) estimating attested encodings consisting of reuse items or compounds.

B.1. Existing Lexicon. For each interval, we approximated the existing lexicon $\mathcal{L}$ by reusing historical English data and language-specific historical corpora. Specifically, we first assumed the senses that existed in the English lexicon also existed in French or Finnish. We then obtained a candidate set of existing words by including the lemma forms for every concept in the English lexicon and standardized them in the same way we standardized the space of English word forms, except we did not remove word forms that can be further lemmatized. The candidate set was further filtered to ensure each word form exists at least $\tau_{\text{token}} = 10$ times in the relevant historical interval of the Google Ngrams corpus (French 2020 version; ref. 23) for French and the Newspaper and Periodical Corpus of the National Library of Finland (FNC; ref. 29) for Finnish. The descriptive statistics for each existing lexicon are summarized in Tables S3 and S4. Note that fluctuations in the number of existing words in Table S4 are due to sparsity of the Finnish corpus for certain historical intervals.

B.2. Attested Encodings of Novel Concepts. To obtain novel reuse items for an interval, we started with all language-specific form-sense pairs in which the sense emerges in the same interval in the English data and the form exists in the language-specific existing lexicon $\mathcal{L}$. We removed pairs in which the form is a stopword, and we removed duplicates by keeping only the shortest form in every set of lemmas that express an identical set of senses. The remaining pairs were added to the reuse-based encoding E^*. To obtain novel compounds, we started with all pairs in which the sense emerges in the same interval in the English data but the form is not in the language-specific existing lexicon $\mathcal{L}$. We then selected pairs in which the form is 1) a closed compound in Wiktionary (25), 2) an open or hyphenated compound according to our inclusive approach for English compounds, or 3) a prepositional compound in the case of French. These compounds were added to the combination-based encoding E^* if their constituents do not contain stopwords. The descriptive statistics for each attested encoding are summarized in Tables S3 and S4. Note that the smaller number of novel items in the last interval is due to sparsity in the English data.

Interval	# existing forms	# existing senses	# reuse	# combination
1900+	19,429	22,840	147	124
1920+	19,948	23,541	143	127
1940+	20,470	24,235	141	78
1960+	21,121	24,730	82	67
1980+	21,682	25,160	16	13

Table S3. Number of existing forms and senses, and novel items across intervals for French

Interval	# existing forms	# existing senses	# reuse	# combination
1900+	17,033	24,552	149	209
1920+	20,221	27,404	155	201
1940+	19,634	27,487	118	128
1960+	14,340	23,370	77	95
1980+	13,848	22,968	11	12

Table S4. Number of existing forms and senses, and novel items across intervals for Finnish

We note that internally inflected compounds were excluded from our WordNet-based data because only lemma forms are listed in the existing lexicons. This may have a greater impact for the French and Finnish analyses since these languages are more morphologically complex than English. We leave the investigation of the role of internal inflection in our efficiency-based account for future work.

C. Evaluation of Sense Frequencies. Our main analyses depended on the frequency of English form-sense pairs in a historical interval. Since historical text only provides the frequency of word forms, we used a state-of-the-art word sense disambiguation algorithm, EWISER (22), to determine the proportion of the usages of a word that corresponds to one of its word senses. To validate whether the resulting sense proportions reflect variation across different historical intervals, we used an existing evaluation framework for semantic change detection (30). The framework consists of a dataset of 100 words that were rated by 5 annotators on a four-point scale, indicating the degree of word meaning change between the 1960s and the 1990s. We followed previous computational work (27, 31) and compared these human ratings of meaning change against scores computed from sense proportions.

Specifically, we computed a score for each word in the dataset by using novelty scores for its word senses relative to the 1960s and the 1990s. Let w be a word with m senses $s_1, s_2, ..., s_m$ in WordNet. Further, let $p_h(s_i)$ be the proportion of s_i in the historical usages of w from the 1960s subset of COHA, and similarly let $p_m(s_i)$ be the proportion from the modern 1990s subset of COHA. We computed a novelty score for s_i using the following formula:

$$N(s_i) = \frac{p_m(s_i) + \alpha}{p_h(s_i) + \alpha} \quad [7]$$

Here α is a constant that prevents division by zero, which we set as 0.01 following ref. 31. Finally, the change score of w is defined as the maximum over sense-level novelty scores; we applied a log transformation to the computed scores to reduce skewness. Since WordNet consists of mappings between synsets and lemmas, we manually lemmatized the entries in the dataset from ref. 30, and we computed change scores for the lemmatized entries.

In Table S5, we compare the correlation between our computed scores and averaged human ratings against correlations obtained in previous studies that were also based on the frequency distribution of word senses: 1) ref. 32 used a collection of historical corpora (21, 33, 34) and measured change following ref. 27 with a state-of-the-art dynamic Bayesian model; 2) ref. 31 used the same data and methods described above, except word senses were obtained from Oxford Dictionary* and disambiguation was based on a nearest-neighbour approach; and 3) refs. 35 and 36 used COHA and word senses inferred from clustering BERT-based contextualized word embeddings (37). Our method yielded Pearson ($p < .001$, $n = 100$) and spearman ($p < .001$, $n = 100$) correlations significantly above zero, and we observe that the performance of our method is comparable to previous results. This suggests that the sense frequencies we computed reflect attested changes in frequency distributions of word senses across historical intervals.

Method	Corpus	Pearson	Spearman
Frermann (2016)	COHA, DTE, CLMET3.0	-	0.377
Hu (2019)	COHA	0.520	0.428
Montariol (2021)	COHA	-	0.510
Our method	COHA	0.472	0.435

Table S5. Correlations between semantic change scores based on sense frequencies and human ratings of change

*https://www.lexico.com/

3. Listener Distribution

We present further analyses on design choices involved in our implementation of the listener distribution. First, we evaluated our semantic (or embedding) space and prototype representations in two parts: 1) we evaluated prototypes of existing words by using them to reconstruct human ratings of pairwise word similarity; and 2) we evaluated composite prototypes by using them to reconstruct human ratings of compound meaning predictability. Then, we examined the sensitivity parameter γ when the speaker communicates existing concepts to listeners.

A. Prototype and Word Similarity. Here we evaluate our WordNet-based semantic space and prototypes constructed by weighted averaging. In the following, we first describe datasets of word similarity ratings, an additional English dictionary, and baselines that involve other choices of semantic space and methods for prototype construction. We then use pairwise word similarity to compare different combinations of embedding method and prototype construction.

A.1. Datasets. We used three standard datasets of pairwise word similarity ratings: WordSim-353 (38), MEN (39), and SimLex-999 (40). WordSim-353 contains 153 English word pairs rated by 13 subjects and 200 English words pairs rated by another 16 subjects; in both cases, the subjects are near-native English speakers and the rating scale is between 0 (totally unrelated) and 10 (very much related or identical). WordSim-353 was further divided into a set of 203 pairs focused on similarity and a set of 252 pairs focused on relatedness (41). MEN contains 3,000 English word pairs, and each pair has a normalized relatedness score between 0 and 1; these scores were obtained from MTurk crowdworkers who were presented with target and control pairs and asked to judge whether the target pair is more related. SimLex-999 contains 999 English word pairs, and each pair has a score between 0 and 10 that was transformed from a 0-6 rating (less to more similar) given by approximately 50 MTurk crowdworkers. Compared to the earlier datasets, SimLex-999 more explicitly focuses on pairwise similarity instead of relatedness or association.

In the main text, our primary source of word senses and sense definitions is WordNet (42). For comparison, we used Oxford Dictionary as an alternate source of word senses. We retrieved primary senses and corresponding definitions for words from two word lists. First, we obtained a subset of the dictionary by recording $3,353$ compounds that appear in both LADEC (24) and the online version of Oxford Dictionary[†]. Second, we obtained $43,482$ COHA lemmas and their definitions that appear in archived webpages of Oxford Dictionary[‡]. In total, this provided us with a set of $77,359$ senses for $30,760$ words.

A.2. Methods. In the main text, we mapped each word sense c to a vector by embedding the definition of c with Sentence-BERT (43). For comparison, we considered a baseline approach where we embedded each word sense by using the bag of words in its definition. For each word sense c, we first obtained its definition in WordNet or Oxford Dictionary, and we recorded the words that are not stopwords and only contain alphabetical characters. We then embedded c by averaging the word embeddings of these words. We used a set of pre-trained word2vec embeddings that were trained on a 600B-token corpus from Common Crawl using the continuous bag-of-words architecture (44, 45).

We considered the following approaches to construct the prototype $q_{w,\mathcal{L}}$ for an existing word w. For WordNet-based embeddings, as in the main text, we constructed the prototype by taking the weighted average of the embeddings of its word senses; we set the weights as sense frequencies obtained from the post-2000 subset of COHA. We also considered two baseline approaches where the weight is uniform or all weight is concentrated on the most common sense (mcs) in WordNet. For Oxford Dictionary-based embeddings, we only considered unweighted averages since we did not have easily accessible frequency information that fully covers our dataset. Finally, since word similarity datasets are usually used to evaluate word embeddings, we directly used the same pre-trained word2vec embeddings (44, 45) as prototype vectors. We always set the existing lexicon $\mathcal{L}$ to be the full set of form-sense pairs in a dictionary since we used contemporary datasets of human similarity ratings.

To evaluate our semantic spaces and prototypes, we first computed the cosine distance between $q_{w,\mathcal{L}}$ and $q_{u,\mathcal{L}}$ for every word pair w and u in one of the word similarity datasets. We then computed the correlation between cosine distances and human ratings of pairwise similarity. To directly compare different approaches, we ensured only word pairs that exist in the vocabulary of every embedding method are used for evaluation. This yielded a 181-pair subset of WordSim-353 focused on similarity, a 219-pair subset of WordSim-353 focused on relatedness, a 2647-pair subset of MEN, and 957-pair subset of SimLex-999.

A.3. Results. We summarize our results for the similarity subset of WordSim-353 in Table S6, for the relatedness subset of WordSim-353 in Table S7, for MEN in Table S8, and for SimLex-999 in Table S9. For all datasets, we observe that directly using word2vec embeddings usually yields the highest correlations, and weighted averaging of WordNet senses encoded by Sentence-BERT is the best sense-level approach. We can also observe that simple averaging of WordNet senses performs better than simple averaging of Oxford Dictionary senses, and Sentence-BERT encoded definitions outperform the bag-of-words baseline. These results provide validation for our model of the listener distribution by showing that the WordNet-based semantic space and weighted averaging approach reflect human judgements of word similarity.

[†] https://www.lexico.com/

[‡] https://archive.org

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	$\mathbf{-0.834}$	< 0.001	$\mathbf{-0.836}$	< 0.001	181
OD W2V (average)	-0.456	< 0.001	-0.452	< 0.001	–
OD SBERT (average)	-0.588	< 0.001	-0.576	< 0.001	–
WN W2V (mcs)	-0.549	< 0.001	-0.497	< 0.001	–
WN W2V (average)	-0.521	< 0.001	-0.506	< 0.001	–
WN W2V (weighted)	-0.659	< 0.001	-0.628	< 0.001	–
WN SBERT (mcs)	-0.617	< 0.001	-0.571	< 0.001	–
WN SBERT (average)	-0.697	< 0.001	-0.695	< 0.001	–
WN SBERT (weighted)	$\mathbf{-0.764}$	< 0.001	$\mathbf{-0.741}$	< 0.001	–

Table S6. Evaluating semantic spaces and word-level prototypes using word similarity ratings (WordSim-353 sim); W2V = word2vec, SBERT = sentence-BERT, OD = Oxford Dictionary, and WN = WordNet

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	$\mathbf{-0.699}$	< 0.001	$\mathbf{-0.728}$	< 0.001	219
OD W2V (average)	-0.173	0.010	-0.190	0.005	–
OD SBERT (average)	-0.283	< 0.001	-0.278	< 0.001	–
WN W2V (mcs)	-0.344	< 0.001	-0.353	< 0.001	–
WN W2V (average)	-0.308	< 0.001	-0.274	< 0.001	–
WN W2V (weighted)	-0.439	< 0.001	-0.416	< 0.001	–
WN SBERT (mcs)	-0.339	< 0.001	-0.303	< 0.001	–
WN SBERT (average)	-0.424	< 0.001	-0.392	< 0.001	–
WN SBERT (weighted)	$\mathbf{-0.509}$	< 0.001	$\mathbf{-0.493}$	< 0.001	–

Table S7. Evaluating semantic spaces and word-level prototypes using word similarity ratings (WordSim-353 rel); abbreviations follow Table S6

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	$\mathbf{-0.819}$	< 0.001	$\mathbf{-0.838}$	< 0.001	2647
OD W2V (average)	-0.388	< 0.001	-0.391	< 0.001	–
OD SBERT (average)	-0.487	< 0.001	-0.493	< 0.001	–
WN W2V (mcs)	-0.515	< 0.001	-0.515	< 0.001	–
WN W2V (average)	-0.448	< 0.001	-0.445	< 0.001	–
WN W2V (weighted)	-0.574	< 0.001	-0.580	< 0.001	–
WN SBERT (mcs)	-0.568	< 0.001	-0.574	< 0.001	–
WN SBERT (average)	-0.595	< 0.001	-0.600	< 0.001	–
WN SBERT (weighted)	$\mathbf{-0.666}$	< 0.001	$\mathbf{-0.683}$	< 0.001	–

Table S8. Evaluating semantic spaces and word-level prototypes using word similarity ratings (MEN); abbreviations follow Table S6

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	$\mathbf{-0.457}$	< 0.001	-0.404	< 0.001	957
OD W2V (average)	-0.074	0.022	-0.082	0.012	–
OD SBERT (average)	-0.275	< 0.001	-0.288	< 0.001	–
WN W2V (mcs)	-0.170	< 0.001	-0.181	< 0.001	–
WN W2V (average)	-0.208	< 0.001	-0.194	< 0.001	–
WN W2V (weighted)	-0.259	< 0.001	-0.246	< 0.001	–
WN SBERT (mcs)	-0.293	< 0.001	-0.313	< 0.001	–
WN SBERT (average)	-0.401	< 0.001	-0.378	< 0.001	–
WN SBERT (weighted)	$\mathbf{-0.438}$	< 0.001	$\mathbf{-0.426}$	< 0.001	–

Table S9. Evaluating semantic spaces and word-level prototypes using word similarity ratings (SimLex-999); abbreviations follow Table S6

B. Composite Prototype and Meaning Predictability. Here we evaluate our WordNet-based semantic space and composite prototypes. We first describe a compound dataset of meaning predictability and additional methods for constructing prototypes for compounds. We then use predictability ratings to compare different combinations of embedding method and construction of composite prototypes.

B.1. Compound Dataset. The Large Database of English Compounds (LADEC; ref. 24) contains 8,957 English closed compounds compiled from the Brown, CELEX and COCA corpora and a dataset of phrases from ref. 46. Every entry in the database is an adjective-noun or noun-noun compound and contains information on the segmentation of the compound into head and modifier,

whether the segmentation is correct, and the corresponding human meaning predictability judgement. Meaning predictability was obtained by asking 1,772 native English speakers how predictable a compound meaning is from its parts on a scale of 0 (not very predictable) to 100 (very predictable). As before, we removed entries that are plural or incorrectly segmented.

B.2. Methods. As in the main text, here we construct composite prototypes by applying composition functions to simple prototypes (i.e., the prototypes of lexicalized words instead of novel compounds). Since we needed to compute information loss for and thus apply composition to a large number of potential compounds, we prioritized tractability and focused on linear composition functions (47). Let $w = xy$ be a compound with constituents x and y; the linear composition of the prototype vectors of its constituents is defined as follows:

$$q_{w,\mathcal{L}} = Aq_{x,\mathcal{L}} + Bq_{y,\mathcal{L}} \quad [8]$$

Here $q_{w,\mathcal{L}}$ is the composite prototype vector of w, and A and B are real-valued matrices; as before we set $\mathcal{L}$ to be the full set of form-sense pairs. We considered a weighted additive version of Equation 8 where the matrices are replaced by scalar weights (47). Here we replaced (A, B) with $(1-\lambda, \lambda)$ for $\lambda = 0, 0.25, 0.5, 0.75, 1$; note that $\lambda = 0.5$ is equivalent to the simple additive function used in the main text since we used cosine distance throughout. We also considered a full version of Equation 8 by using linear projection (ref. 48; also see refs. 49, 50). Specifically, we obtained matrices (A, B) by minimizing the mean squared error between the simple and composite prototypes of every compound in LADEC. This was implemented using batch gradient descent for 10,000 iterations and the Adam optimizer (51) with learning rate $\alpha = 0.001$ and decay rates $(\beta_1, \beta_2) = (0.9, 0.999)$.

We first evaluated these composition functions by using the best-performing simple prototypes in the previous section. After intersecting LADEC with WordNet and our set of pre-trained word2vec embeddings, we obtained a set of 3,336 compounds with meaning predictability ratings. We then further evaluated our semantic space and simple prototypes from the previous section. After intersecting LADEC additionally with Oxford Dictionary, we obtained a set of 2,121 compounds with meaning predictability ratings. In both cases, we computed the cosine distance between the composite and simple prototypes of every compound in LADEC, and we correlated these distances with human ratings.

B.3. Results. The correlations based on the best simple prototypes are shown in Table S10. We observe that both types of embeddings yielded similar results across composition functions, and that the correlations tend to be relatively low when the weight on one of the constituents is zero. For simplicity, we focused on the simple additive function ($\lambda = 0.5$) since it tends to achieve the best results. In Table S11, we further evaluate semantic spaces and simple prototypes based on word2vec, Oxford Dictionary, and WordNet by using the simple additive function, and we observe the same general trends as in the previous section. These results provide further validation for our model of the listener distribution by showing that our composite prototypes reflect human judgements of compound meaning predictability.

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
W2V Additive $\lambda = 0$	-0.353	< 0.001	-0.362	< 0.001	3336
W2V Additive $\lambda = 0.25$	$\mathbf{-0.388}$	< 0.001	$\mathbf{-0.394}$	< 0.001	–
W2V Additive $\lambda = 0.5$	-0.376	< 0.001	-0.384	< 0.001	–
W2V Additive $\lambda = 0.75$	-0.291	< 0.001	-0.306	< 0.001	–
W2V Additive $\lambda = 1$	-0.203	< 0.001	-0.218	< 0.001	–
W2V Linear	-0.310	< 0.001	-0.315	< 0.001	–
WN Additive $\lambda = 0$	-0.330	< 0.001	-0.347	< 0.001	–
WN Additive $\lambda = 0.25$	-0.377	< 0.001	-0.390	< 0.001	–
WN Additive $\lambda = 0.5$	$\mathbf{-0.385}$	< 0.001	$\mathbf{-0.392}$	< 0.001	–
WN Additive $\lambda = 0.75$	-0.329	< 0.001	-0.341	< 0.001	–
WN Additive $\lambda = 1$	-0.258	< 0.001	-0.273	< 0.001	–
WN Linear	-0.363	< 0.001	-0.370	< 0.001	–

Table S10. Evaluating composition functions based on WordNet (WN) sense-level representations and word-level word2vec (W2V)

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	-0.385	< 0.001	-0.391	< 0.001	2121
OD W2V (average)	-0.128	< 0.001	-0.109	< 0.001	–
OD SBERT (average)	-0.256	< 0.001	-0.252	< 0.001	–
WN W2V (mcs)	-0.154	< 0.001	-0.157	< 0.001	–
WN W2V (average)	-0.172	< 0.001	-0.171	< 0.001	–
WN W2V (weighted)	-0.267	< 0.001	-0.258	< 0.001	–
WN SBERT (mcs)	-0.300	< 0.001	-0.317	< 0.001	–
WN SBERT (average)	-0.346	< 0.001	-0.348	< 0.001	–
WN SBERT (weighted)	$\mathbf{-0.408}$	< 0.001	$\mathbf{-0.417}$	< 0.001	–

Table S11. Evaluating semantic spaces and simple prototypes using meaning predictability ratings; abbreviations follow Table S6

C. Sensitivity Parameter. In the main text, we combined sense and word-level representations with the similarity choice model (52, 53) to instantiate the listener distribution. The model contains a single sensitivity parameter γ that determines how likely the listener prefers to infer the most transparent interpretation of the word that was uttered by the speaker. We focused on the communication of novel concepts, but in our scenario the speaker also communicates concepts in the existing lexicon $\mathcal{L}$ to listeners that use identical listener distributions for existing words. Here we examine how the sensitivity parameter influences the average information loss incurred in the latter case.

To do so, we reused our WordNet-based datasets for English, French, and Finnish, except for each historical interval, we focused on the existing lexicon $\mathcal{L}$ instead of the encoding E^*. In other words, we computed the average cost of communicating existing concepts by using the omitted sum in Equation 6:

$$L_\beta[\mathcal{L}] = \sum_{(c,w)\in\mathcal{L}} p(c,w|\mathcal{L}) \cdot (h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w)) \quad [9]$$

We used two implementations of the weight $p(c,w|\mathcal{L})$ and the listener distribution $\hat{m}_{w,\mathcal{L}}$. First, following the main text, we implementing each weight using sense frequencies from COHA (21) and token frequencies from the Google Ngrams corpus (23) and the FNC (29), and we used the same definition of the listener distribution as in the main text. Second, we repeated the first implementation by changing the weight to a uniform distribution over existing form-sense pairs while keeping the listener distribution constant.

Since word length is independent of γ in Equation 9, we only computed the average information loss across five intervals for English, French and Finnish. We set the sensitivity parameter γ as $0, 1, 2, ..., 29$. The results are summarized in Figure S1. As expected, information loss is high when γ is small since that implies the listener does not distinguish among senses. However, loss decreases with γ until it reaches $[15, 20]$. This is because an overly large γ puts most probability mass on a single sense that is the nearest neighbour of the prototype $q_{w,\mathcal{L}}$, which increases information loss when the same word is used to express multiple senses. We can also observe that French and Finnish tend to have higher information loss when the weight $p(c,w|\mathcal{L})$ is uniform. This is because large γ penalizes words with many senses more severely when the weight is uniformly distributed as opposed to being concentrated on prototypical or frequent senses. Similarly, the difference between English (uniform) and the other languages (uniform) is because the greater coverage of lexical items in the original Princeton WordNet caused the English lexicons to have more monosemous words (average percentage of monosemous words = 82.1%) than French lexicons (average percentage of monosemous words = 61.1%) and Finnish lexicons (average percentage of monosemous words = 50.8%).

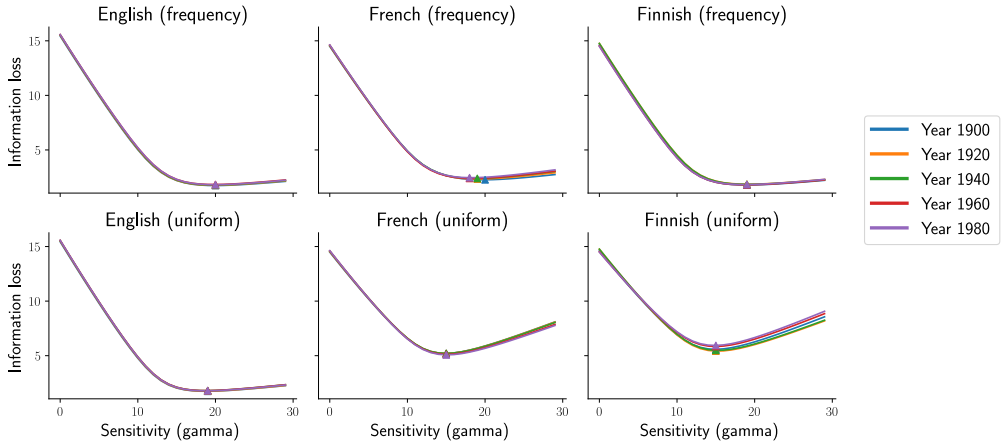

Fig. S1. Information loss incurred over communicating existing concepts. Triangles indicate the lowest information loss for a specific language and interval. Legend on the right indicates the starting year of the interval.

4. Additional Average-Case Analyses

We present additional average-case analyses of attested reuse items and compounds. First, we present analyses using variants of the main-text implementation with different sensitivity parameters and a uniform frequency distribution of attested items. Second, we present an average-case analysis that uses phonological representations of word forms. We then present an analysis of the English data by representing concepts with historical word embeddings. Finally, we present two analyses on alternative datasets of lexicalized concepts.

A. Parameter Variation. We tested the robustness of our main-text results by computing average-case efficiency loss under different settings of the sensitivity parameter γ and the joint distribution for need and production. We used $\gamma = 5, 6, ..., 15$ and both frequency-based and uniform distributions. For tractability, unlike the main text we used a small number of tradeoff parameters, setting $\beta = 0, 0.15, 0.5, 1, 10$, and we reduced the number of baselines created per attested encoding to $1,000$. All other procedures follow the methods in the main text. Figure S2 summarizes the efficiency loss of attested items and baselines. Each bar corresponds to the average over all languages, intervals, and strategies. We observe the same trend as in the main text: attested encodings are more efficient relative to both baselines, and near-synonym baselines are more efficient than random baselines. We can also observe that efficiency loss tends to increase with γ. This is likely because larger γ increases the gap in information loss between any non-optimal encoding and the Pareto frontier.

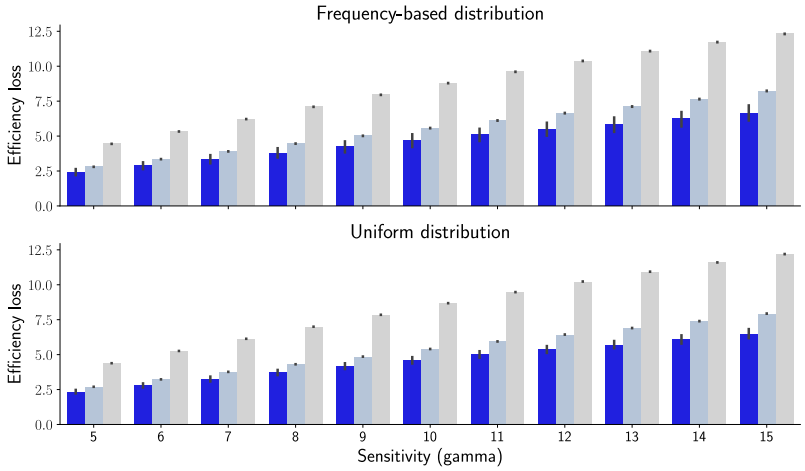

Fig. S2. Efficiency loss of attested reuse items and compounds relative to the average loss of baselines. As in the main text, blue bars correspond to attested items, light blue bars correspond to near-synonym baselines, and grey bars correspond to random baselines. Error bars show bootstrapped 95% confidence intervals.

B. Phonological Representation. In the main text, we used orthographic representations of word forms to approximate speaker production effort. Here we present the same analysis by using a better approximation that represents word forms using phonemes. We focused on English and French since the Finnish grapheme-phoneme mapping is essentially one-to-one (54).

B.1. Sound Dictionaries. To test our proposal using phonological representation of word forms, we used the following sound dictionaries. For English, we used the unstressed version of the CMU Pronouncing Dictionary (55) which is commonly used to evaluate models for grapheme to phoneme conversion. The dictionary contains $133,854$ unique entries, and each entry is an English word in orthographic form paired with a phonemic form. The phoneme inventory of the dictionary contains 39 unique phonemes. For French, we used the Lexique database (version 3.83; ref. 56). This database contains $142,694$ unique entries. Each entry contains a series of variables that describe a lexical item, but we focused on each item's orthographic form, phonemic form, and part of speech. The phoneme inventory of the dictionary contains 38 unique phonemes. For both dictionaries, we removed entries in which the orthographic form contains non-alphabetical characters. For French, if entries with the same orthographic form and part of speech map to multiple phonemic forms, we used the first phonemic form in the dictionary's default ordering. We did the same for English except we did not consider part of speech as it was not provided by the dictionary. This provided us with $116,507$ unique entries for English and $137,819$ unique entries for French.

We note that these sound dictionaries were compiled relatively recently and thus do not account for sound changes that took place over the target periods of our study. While chain shifts preserve form-concept mappings, other types of sound change (e.g., deletions or mergers) may affect the length of a word form or the distinction between phonemes and thus the informativeness of a word form (57). We leave more fine-grained phonological representations and their application to an efficiency-based analysis of reuse and combination for future work.

B.2. Out of Vocabulary Words. Although these sound dictionaries are large, they only cover a limited subset of lemmas in the English and French WordNets because many entries correspond to inflected forms. To ensure a comprehensive analysis of our historical data, we obtained the phonemic form of a word w not in the above sound dictionaries in one of two ways. If w is a

compound in some attested encoding E^*, we approximated its phonemic form using the concatenation of the phonemic forms of its constituents. Otherwise, if w is a word in the existing lexicon $\mathcal{L}$, we approximated its phonemic form using a state-of-the-art grapheme-to-phoneme model that is based on the transformer (58, 59). We first evaluated the model by applying it to the sound dictionaries described above and using 10-fold cross validation. We obtained an accuracy of 0.694 and a phoneme error rate of 0.0831 for English, and an accuracy of 0.960 and a phoneme error rate of 0.0831 for French; the English accuracy is slightly worse than the accuracy reported in ref. 59 for the same dataset and the high French accuracy is likely due to a large number of inflicted forms in the data, but nonetheless we take this as evidence that the model is able to generalize well. We thus trained an English model and a French model on the full dictionaries respectively, and used them to map every form in existing lexicons to a phonological representation.

B.3. Results. We used these phonological representations of English and French words to test our proposal. We reused the implementations of the scenario described in the main text, except for changes in form-concept mappings. For both languages, we replaced the orthographic form in every original form-concept pair with its unique corresponding phonemic form; for French, some orthographic forms have multiple phonemic forms corresponding to different parts of speech, and in these cases we assigned the correct phonemic form based on the part of speech of the WordNet sysnet for the concept. For both the existing lexicon $\mathcal{L}$ and the attested encoding E^*, we removed duplicate form-concept pairs after the conversion to phonemic forms. We removed compounds in E^* if their phonemic forms were contained in the phoneme-based existing lexicon, but this only removed a very small number of attested compounds. We summarize descriptive statistics for our final datasets in Tables S12 and S13.

Interval	# existing forms	# existing senses	# reuse	# combination
1900+	45,919	43,649	136	766
1920+	48,101	45,658	127	914
1940+	49,534	47,194	140	627
1960+	50,452	48,070	92	473
1980+	51,061	48,555	21	47

Table S12. Number of existing forms and senses, and novel items for English after mapping orthographic forms to phonemic forms

Interval	# existing forms	# existing senses	# reuse	# combination
1900+	18,531	22,840	147	124
1920+	19,031	23,541	143	127
1940+	19,530	24,235	140	77
1960+	20,121	24,730	82	67
1980+	20,643	25,160	16	13

Table S13. Number of existing forms and senses, and novel items for French after mapping orthographic forms to phonemic forms

To assess the efficiency of a phoneme-based encoding E^*, we measured the word length of each phonemic form using the number of phonemes in the string. We measured the information loss of a form-concept pair by reusing our implementation of the listener distribution in the main text, after finding that γ behaves similarly when existing concepts are communicated via phonemic forms. These costs were then averaged by reusing frequencies from the main text and combining the frequencies of homophones. As in the main text, we compare attested encodings of reuse items or compounds to Pareto frontiers and baseline encodings. These comparisons are summarized in Figure S3. We observe that similar to the main text, attested encodings tend to be closer to the frontier than both near-synonym and random encodings. We show quantitative measures of efficiency loss in Figure S4 which confirms our qualitative observations. These results show that our tradeoff proposal is robust to different channels of communication.

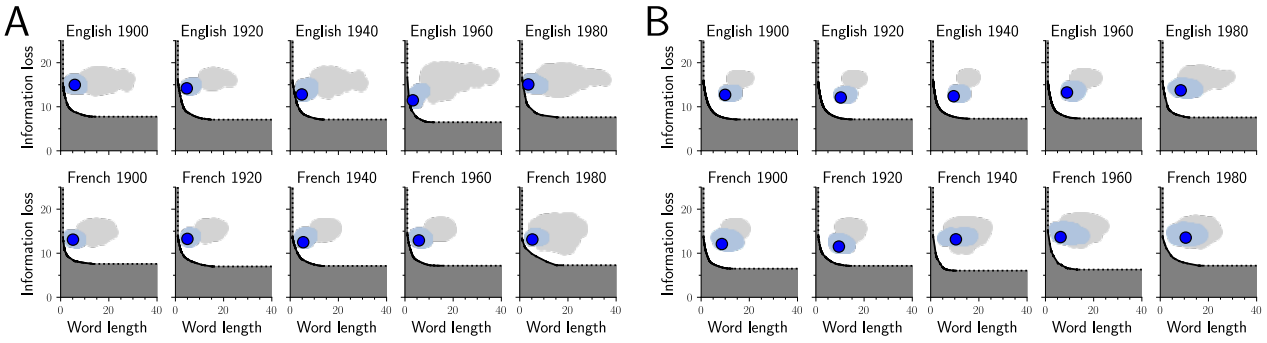

Fig. S3. Illustration comparing (A) attested reuse items and (B) attested compounds to the constructed baselines and the Pareto frontier using an implementation based on phonological representations of word forms.

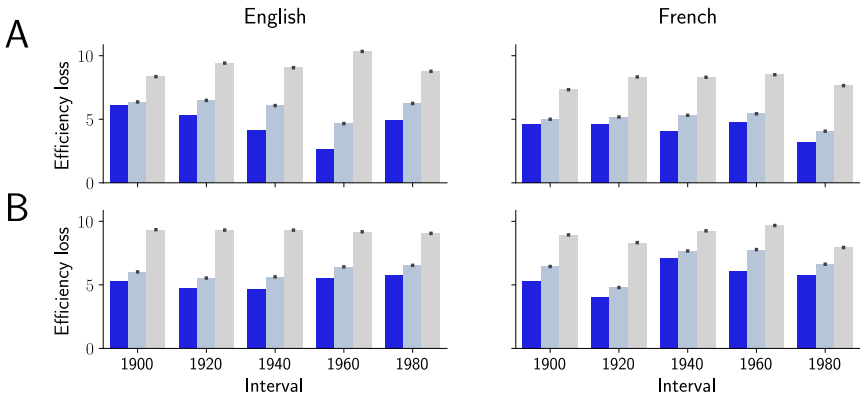

Fig. S4. Efficiency loss of (A) attested reuse items and (B) attested compounds based on phonological representations of word forms relative to the average loss of baselines. Error bars show bootstrapped 95% confidence intervals.

C. Historical Embedding-based Implementation. In the main text, we implemented the listener distribution by assuming the similarity between any pair of concepts is constant across time, and we used a sentence encoder trained on contemporary text and sentence similarity ratings (43). Here we relax this assumption by measuring conceptual similarity with historical word embeddings (44, 60), and we analyze a subset of the English data used in the main text under this alternative implementation.

C.1. Methods. We used pre-trained English historical word embeddings provided by HistWords (60). Specifically, we used the version that was based on the Google Ngrams corpus (All English version 2; ref. 23), which covers the most frequent 100,000 words in the 1800-2000 period of the corpus. For each decade in that period, a set of word embeddings is trained solely on the corresponding subset of the corpus using the skip-gram architecture (44). After training, these sets of embeddings were aligned across time periods while preserving the cosine similarities among embeddings from the same decade. Importantly, the embedding space for each decade is not influenced by novel word sense extensions or novel words that emerged in future decades.

To apply these embeddings to our efficiency analysis, we reused our method in *SI Appendix, Section S3* to represent WordNet senses by taking the bag of words in their definitions and averaging the corresponding word embeddings. We evaluated these representations by reusing the same method and datasets of pairwise word similarity and compound meaning predictability ratings in *SI Appendix, Section S3*. More specifically, for each word in the evaluation datasets, we represented the word by embedding its most common sense (mcs) in WordNet, by using the simple average of the embeddings of all definitions, or by using the weighted average of the embeddings of all definitions. In the case of compound meaning predictability, we focused on using simple additive composition. In all evaluations, we used HistWords embeddings from the decade of 1990 since this is the most recent available decade in HistWords and the evaluation datasets are contemporary; we compared the HistWords-based sense representations to two alternate approaches: 1) sense representations based on sentence-BERT (37), and 2) a simple approach in which we computed the cosine similarity between static word embeddings from HistWords.

We summarize our evaluation results for the similarity subset of WordSim-353 in Table S14, for the relatedness subset of WordSim-353 in Table S15, for MEN in Table S16, for SimLex-999 in Table S17, and for LADEC compound meaning predictability ratings in Table S18. Here we observe that the representation we used in the main text (WN SBERT) is the best sense-level representation across all datasets, but we can also observe that the HistWords-based representation that is weighted by sense frequency tends to outperform several baselines. These results provide evidence that our HistWords-based representation of word senses reflect human judgements of word similarity and compound meaning predictability to a reasonable extent.

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	-0.715	< 0.001	-0.716	< 0.001	194
WN W2V (mcs)	-0.540	< 0.001	-0.508	< 0.001	–
WN W2V (average)	-0.577	< 0.001	-0.536	< 0.001	–
WN W2V (weighted)	-0.664	< 0.001	-0.644	< 0.001	–
WN SBERT (mcs)	-0.616	< 0.001	-0.576	< 0.001	–
WN SBERT (average)	-0.694	< 0.001	-0.688	< 0.001	–
WN SBERT (weighted)	$\mathbf{-0.761}$	< 0.001	$\mathbf{-0.738}$	< 0.001	–

Table S14. Evaluating semantic spaces using word similarity ratings (WordSim-353 sim); W2V = word2vec using HistWords, SBERT = sentence-BERT, and WN = WordNet

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	$\mathbf{-0.604}$	< 0.001	$\mathbf{-0.614}$	< 0.001	235
WN W2V (mcs)	-0.327	< 0.001	-0.323	< 0.001	–
WN W2V (average)	-0.362	< 0.001	-0.334	< 0.001	–
WN W2V (weighted)	-0.411	< 0.001	-0.399	< 0.001	–
WN SBERT (mcs)	-0.340	< 0.001	-0.308	< 0.001	–
WN SBERT (average)	-0.407	< 0.001	-0.377	< 0.001	–
WN SBERT (weighted)	$\mathbf{-0.489}$	< 0.001	$\mathbf{-0.474}$	< 0.001	–

Table S15. Evaluating semantic spaces using word similarity ratings (WordSim-353 rel); abbreviations follow Table S14

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	$\mathbf{-0.685}$	< 0.001	$\mathbf{-0.698}$	< 0.001	2911
WN W2V (mcs)	-0.482	< 0.001	-0.483	< 0.001	–
WN W2V (average)	-0.528	< 0.001	-0.520	< 0.001	–
WN W2V (weighted)	-0.583	< 0.001	-0.590	< 0.001	–
WN SBERT (mcs)	-0.570	< 0.001	-0.574	< 0.001	–
WN SBERT (average)	-0.598	< 0.001	-0.601	< 0.001	–
WN SBERT (weighted)	$\mathbf{-0.671}$	< 0.001	$\mathbf{-0.686}$	< 0.001	–

Table S16. Evaluating semantic spaces using word similarity ratings (MEN); abbreviations follow Table S14

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	-0.234	< 0.001	-0.231	< 0.001	989
WN W2V (mcs)	-0.182	< 0.001	-0.180	< 0.001	–
WN W2V (average)	-0.245	< 0.001	-0.228	< 0.001	–
WN W2V (weighted)	-0.272	< 0.001	-0.246	< 0.001	–
WN SBERT (mcs)	-0.303	< 0.001	-0.316	< 0.001	–
WN SBERT (average)	-0.399	< 0.001	-0.373	< 0.001	–
WN SBERT (weighted)	$\mathbf{-0.441}$	< 0.001	$\mathbf{-0.424}$	< 0.001	–

Table S17. Evaluating semantic spaces using word similarity ratings (SimLex-999); abbreviations follow Table S14

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	-0.190	< 0.001	-0.200	< 0.001	1208
WN W2V (mcs)	-0.185	< 0.001	-0.187	< 0.001	–
WN W2V (average)	-0.283	< 0.001	-0.273	< 0.001	–
WN W2V (weighted)	-0.327	< 0.001	-0.318	< 0.001	–
WN SBERT (mcs)	-0.293	< 0.001	-0.300	< 0.001	–
WN SBERT (average)	-0.341	< 0.001	-0.331	< 0.001	–
WN SBERT (weighted)	$\mathbf{-0.417}$	< 0.001	$\mathbf{-0.415}$	< 0.001	–

Table S18. Evaluating semantic spaces using compound meaning predictability ratings; abbreviations follow Table S14

C.2. Results. We used English historical word embeddings to further test our proposal. We reused the implementations in the main text except we represented each word sense via the averaging method described above and word embeddings from the decade of $t_0 = t_1 - 10$. To make sure the sense representations do not misrepresent the corresponding word senses, we removed word senses in the existing lexicon $\mathcal{L}$ and attested encoding E^* if one of the non-stopword words in their definitions does not exist in the vocabulary of the embeddings from t_0. The descriptive statistics of our final datasets are summarized in Table S19.

Interval	# existing forms	# existing senses	# reuse	# combination
1900+	34,886	31,538	60	366
1920+	37,249	33,655	54	474
1940+	37,691	34,197	57	341
1960+	40,680	37,042	50	287
1980+	43,938	40,137	15	30

Table S19. Number of existing forms and senses, and novel items across intervals after intersecting with HistWords

As in the main text, we assessed the efficiency of an attested encoding by measuring orthographic length and measuring information loss with $\gamma = 10$. These costs were averaged using the same frequencies from the main text plus normalization. We compared the attested reuse-based and compound-based encodings to Pareto frontiers and the same types of baseline encodings, which are summarized in Figure S5. Similar to the main text, we observe that attested encodings tend to be closer to the

Pareto frontier than both the near-synonym and random baselines. Figure S6 shows comparisons based on the quantitative measure of efficiency loss, and we observe a trend that is consistent with our qualitative observations. These results suggest that our findings do not arise because of potential systematic biases in using contemporary embeddings.

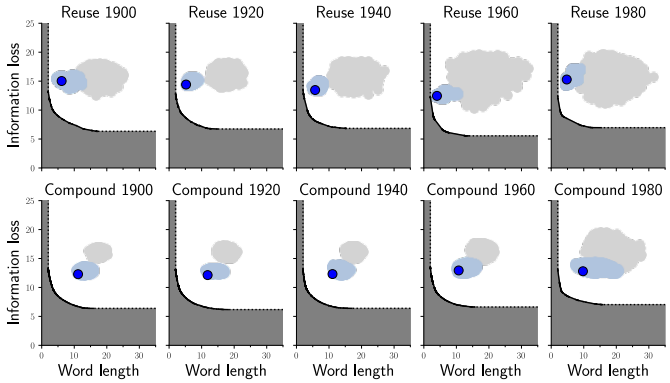

Fig. S5. Illustration comparing attested reuse items and compounds to constructed baselines and Pareto frontiers using an implementation based on HistWords.

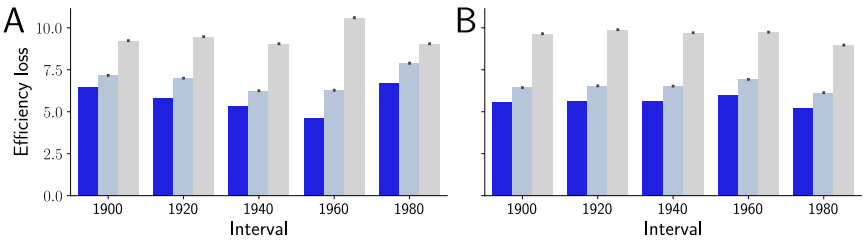

Fig. S6. Efficiency loss of (A) attested reuse items and (B) attested compounds relative to the average loss of baselines, under an implementation of our scenario using HistWords. Error bars show bootstrapped 95% confidence intervals.

D. Alternative Datasets of Lexicalized Concepts. Here we present two further analyses to show our findings are robust to datasets of lexicalized concepts that are alternative to WordNet. We first describe an implementation of our scenario based on Oxford Dictionary. We then describe an implementation that assumes one-to-one correspondence between concept and form.

D.1. Oxford Dictionary. In *SI Appendix, Section S3*, we showed that WordNet-based semantic representations tend to be better at reconstructing human ratings of word similarity than Oxford Dictionary-based representations, but the latter nonetheless significantly correlated with human ratings. We thus tested the robustness of our findings and performed an analysis of English reuse items and compounds using Oxford Dictionary (OD) definitions. We first describe how we processed OD-based form-sense pairs, and then we describe how we adapted our methods and assessed the efficiency of attested items.

We first obtained candidates of novel reuse items and compounds by using our OD dataset from the previous sections and intersecting it with the HTE (26). To obtain novel reused items, we used the same keyword matching method (see *SI Appendix, Section S2.A*) and obtained 325 candidate form-sense pairs, which were then manually checked against the OED and yielded 149 pairs that emerged during the past century. To obtain novel compounds, we focused on closed compounds since we only retrieved unigrams from OD and we identified these compounds using LADEC (24) and Wiktionary (25). As in the WordNet implementation, we assumed each compound-sense pair emerged at the first citation year of the compound form, which provided us with 399 compound-sense pairs from the past century.

To obtain emerging and existing concepts and their encodings for each interval, we first filtered form-sense pairs that contained *unknown*, *abbreviation*, or *proper noun* in their part of speech information, and we then reused the procedures for processing WordNet-based English data. We used COHA-based frequency data for OD senses provided by ref. 31 to estimate the sets of existing form-sense pairs; polysemous words not in their dataset were excluded to simplify the calculation of form-sense frequencies. The total number of existing forms, existing senses, and novel items are summarized in Table S20.

To assess the efficiency of these attested items, we reused the main-text implementation with the exception of the need and production distributions and the semantic space. The need and production distributions were estimated using the Google Ngrams corpus (23) and add-one smoothing as in the main text, but sense frequencies were estimated using the dataset from ref. 31. The listener distribution was also implemented in the same way, but here we used sentence embeddings of OD sense definitions and used the frequencies of these senses to construct prototypes. Note that reusing the same implementation implies we set $\gamma = 10$. We also created the near-synonym and random baselines in the same way. We approximated every Pareto frontier via our greedy method for $\beta = 0, 0.01, ..., 10$.

Interval	# existing forms	# existing senses	# reuse	# combination
1900+	27,533	31,121	30	46
1920+	28,471	32,214	28	62
1940+	29,167	33,047	37	48
1960+	29,753	33,696	31	53
1980+	30,287	34,197	23	14

Table S20. Number of existing forms and senses, and novel items across intervals for the Oxford Dictionary-based analysis

We compare these English reuse items and closed compounds to baselines in Figure S7. As in the main text, each point corresponds to an encoding of novel concepts within a certain interval, and the closer a point is to the Pareto frontier the more efficient it is. For this smaller set of English reuse items and compounds, we observe that attested items tend to be more efficient relative to both baselines as in the main results, except for reuse items in the 1960+ interval. Figure S8 shows the average-case efficiency loss of these items which confirms our qualitative observations. Taken together, these results suggest our findings are generally robust to different dictionaries of word senses.

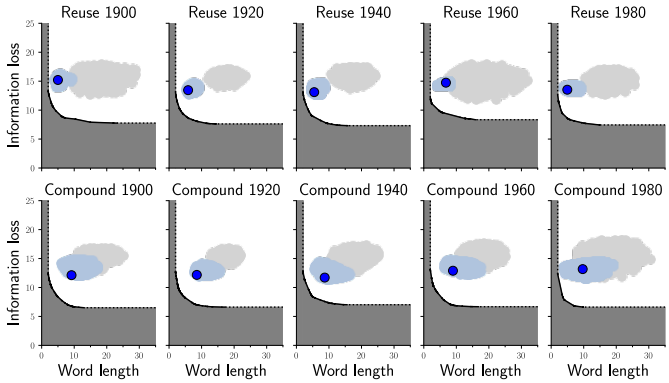

Fig. S7. Illustration comparing attested reuse items and compounds to constructed baselines and Pareto frontiers under an Oxford Dictionary-based implementation.

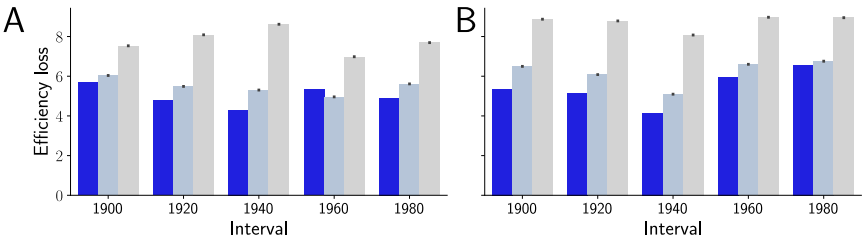

Fig. S8. Efficiency loss of (A) attested reuse items and (B) attested compounds relative to the average loss of baselines, under an implementation of our scenario using the Oxford Dictionary. Error bars show bootstrapped 95% confidence intervals.

D.2. One-to-one Correspondence. A more coarse-grained way to implement our framework is to assume one-to-one correspondence between concept and form instead of using word senses. An intuitive extension of this assumption is to operationalize novel concepts as new forms, which would be based on more accessible and accurate first citation and frequency information as it does not require sense-level information. In the following, we describe how we adapted this alternate implementation to our methods for the main text and used it to assess the efficiency of novel items in English.

First, for each historical interval used in the main text, we listed the sets of existing words, novel compounds, and reuse items. For an interval starting at year t_1, we set the list of existing words as all word forms in the HTE (26) that existed in the year $t_1 - 1$ and do not contain non-alphabetical or uppercase characters; the size of these lists range from 91,816 to 112,094. We obtained novel English compounds by using closed compounds in LADEC (24) and Wiktionary (25) and checking their first citation in the HTE (26); we removed compounds in which the rightmost constituent is a preposition, and this provided us with 565 compounds that emerged in the past century. One drawback of the one-to-one assumption is that we no longer have access to word senses, but we can still examine the subset of reuse items that involve existing words gaining the meaning of a new word. We thus used compound head words to approximate reuse items that express the meanings of attested novel compounds; we assumed the rightmost constituent of each novel compound is the head word.

Instead of using word sense definitions and their embeddings to implement the listener distribution, a natural approach in this word-level analysis is to use word embeddings (e.g., ref. 44). Here, we used the same pre-trained word2vec embeddings (44, 45)

that we validated against human ratings of English word similarity and compound meaning predictability in a previous section. We intersected these embeddings with the lists of existing words and novel compounds, and for each interval we removed a form from the set of novel compounds if one of the constituents is not an existing word. The total number of existing forms and novel items are summarized in Table S21.

Interval	# existing forms	# compounds
1900+	60,055	98
1920+	62,601	114
1940+	65,936	117
1960+	68,895	114
1980+	71,182	27

Table S21. Number of existing forms and novel compounds across intervals for the word-level analysis

To assess the efficiency of these attested compounds and approximate reuse items, we used a simplified version of the implementation in the main text that is based on word-level statistics. Since each form corresponds to a single concept, we estimated the need and production distributions using unigram frequency from the Google Ngrams corpus (23) and add-one smoothing. We used the same settings for the listener distribution except we used a single pre-trained word2vec embedding (45) to represent the prototype of an existing word. Lastly, for simplicity we used $\gamma = 10$, but we note that due to the one-to-one assumption, the information loss incurred over communicating existing concepts is minimized when $\gamma \to \infty$ instead of the lower values observed in *SI Appendix, Section S3.C*. The baselines were created in the same way as in the main text. We estimated Pareto frontiers using the same method for $\beta = 0, 0.01, ..., 10$.

Figure S9 shows the average information loss and word length of attested closed compounds and approximate reuse items. As usual, the closer an encoding is to the Pareto frontier the more efficient it is, but here the frontier is the same for both reuse and compound in the same historical interval because the set of novel concepts is the same. Figure S10 shows the quantitative efficiency loss that intuitively corresponds to how far each encoding is away from the Pareto frontier. Overall, we observe that attested items are relatively more efficient than the corresponding baselines. We can also observe that approximate reuse items tend to have lower values of efficiency loss than attested compounds. This may be due to the fact that our word embeddings were trained on contemporary text in which the historically dominant meanings of reused words were replaced by historically emerging concepts (e.g., *phone* gaining the sense of *cellphone* and losing the sense of rotary dial phone).

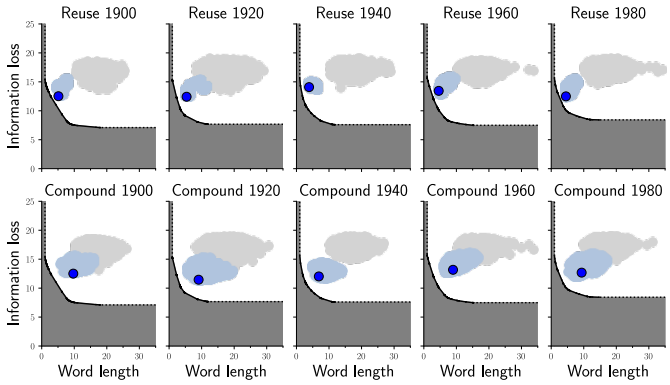

Fig. S9. Illustration comparing attested reuse items and compounds to constructed baselines and Pareto frontiers under the one-to-one assumption.

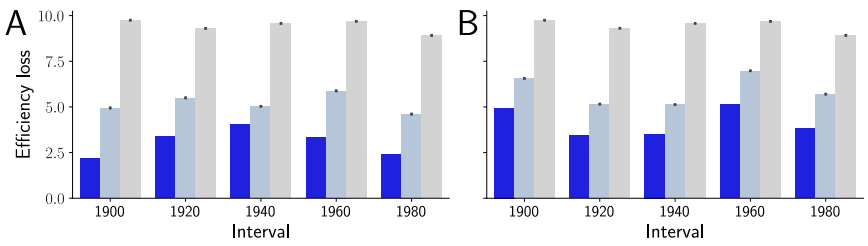

Fig. S10. Efficiency loss of (A) approximate reuse items based on head words and (B) attested compounds relative to the average loss of baselines, under an implementation of our scenario using word embeddings. Error bars show bootstrapped 95% confidence intervals.

5. Additional Item-Level Analyses

We present additional analyses in which we analyze individual reuse items or compounds as opposed to encodings (or sets of form-concept pairs). We first compare item-level efficiency loss between orthographic and phonemic representations of the same attested items. The remainder of the section focuses on our implementation in the main text. We provide examples of novel items that we classified as literal items, and we illustrate the main-text literal vs non-literal comparisons. We next examine distributions of item-level efficiency loss alongside distributions of information loss and word length. Lastly, we explore two additional factors as predictors of variation in item-level efficiency.

A. Comparing Orthographic and Phonemic Forms. Here we examine the item-level variation in efficiency based on phonemic forms. We reused our implementation in *SI Appendix, Section S4.B* and replaced each attested encoding E^* with a singleton containing a single attested item as in the main text. We performed two analyses: 1) we tested whether item-level efficiency differs between representations on average, and 2) we tested whether the relative rank of an attested item in terms of efficiency differs across representations. In the first analysis, we did not find there are significant differences in efficiency for English reuse items ($t(1034) = -0.660, p = 0.509$), French reuse items ($t(1056) = 1.063, p = 0.288$), English compounds ($t(5654) = -0.565, p = 0.572$), or French compounds ($t(812) = 0.273, p = 0.785$). In the second analysis, we found that item-level efficiency is highly correlated across representations. The correlation statistics are summarized in Table S22. Overall, we did not find significant differences in item-level variation between the representations, and thus we focused on analyzing item-level variation using orthographic forms.

Group	Spearman ρ	p-value	N
English reuse	0.925	< 0.001	518
English compound	0.967	< 0.001	2828
French reuse	0.950	< 0.001	529
French compound	0.958	< 0.001	407

Table S22. Correlations between orthography-based item-level efficiency and phoneme-based item-level efficiency

B. Examples of Literal Items. Table S23 shows examples of literal items from the main text.

Label	Head	Head hypernym definition
printer	printer	a machine that prints
birthday card	card	a rectangular piece of stiff paper used to send messages
publicité	publicité	a message issued in behalf of some product or cause or idea or person or institution
turbine à gaz	turbine	rotary engine in which the kinetic energy of a moving fluid is converted into mechanical energy by causing a bladed rotor to rotate
suodatin	suodatin	device that removes something from whatever passes through it
sotarikos	rikos	activity that transgresses moral or civil law

Table S23. Examples of literal items and their hypernyms in existing lexicons

C. Efficiency in Literal and Non-Literal Items. Figure S11 illustrates the comparisons between literal and non-literal items presented in the main text.

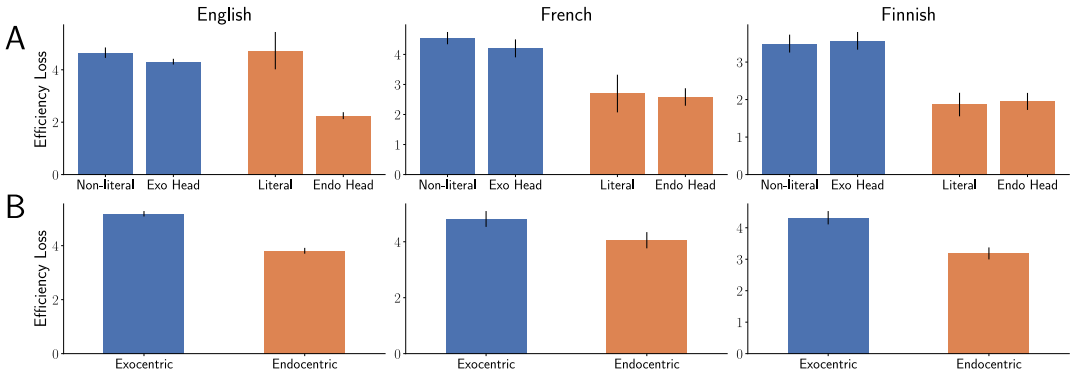

Fig. S11. Averages of item-level efficiency loss for literal and non-literal (A) reuse items and (B) compounds. For each plot in row (A), the left shows attested non-literal reuse items and additional non-literal reuse items based on exocentric compounds; the right shows attested literal reuse items and additional literal reuse items based on endocentric compounds. Error bars indicate bootstrapped 95% confidence intervals.

D. Strategy Comparison. We compare distributions of item-level costs between reuse and compounding, separately for the languages examined in the main text. Figure S12 summarizes the distributions for English: on average, there is no evidence that item-level efficiency loss is significantly different between reuse and compounding ($t(3346) = -0.130$, $p = 0.897$), but reuse items tend to have higher information loss than compounds ($t(3346) = 13.046$, $p < .001$) and lower word length ($t(3346) = -34.908$, $p < .001$). Distributions of item-level costs for French and Finnish are summarized in Figures S13 and S14. A similar trend is observed for French: reuse items do not differ significantly from compounds in efficiency loss ($t(936) = -0.498$, $p = 0.618$), but they have higher information loss ($t(936) = 5.784$, $p < 0.001$) and lower length ($t(1154) = -31.907$, $p < 0.001$). However, Finnish reuse items are both lower in efficiency loss ($t(1154) = -5.225$, $p < 0.001$) and length ($t(1154) = -26.909$, $p < 0.001$), while information loss is not significantly different ($t(1249) = 0.0202$, $p = 0.984$).

In sum, reuse items have lower word length than compounds on average in all languages, and information loss is lower in compounds relative to reuse items in English and French. Nonetheless, differences between strategies are smaller in terms of information loss compared to differences in length across all languages, and Finnish compounds are not more informative on average. Since our listener distributions for compounds were implemented using very simple models, we hypothesize that it could have systematically underestimated the informativeness of compounds and their relative advantage over reuse items.

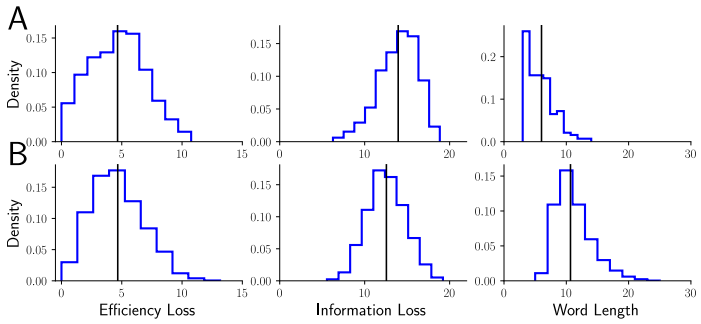

Fig. S12. Distributions of item-level efficiency loss, information loss, and word length for English (A) reuse items and (B) compounds. In each plot, the vertical line indicates the mean of the distribution.

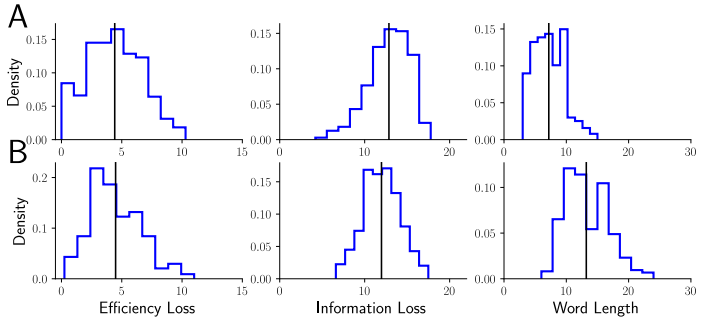

Fig. S13. Distributions of item-level efficiency loss, information loss, and word length for French (A) reuse items and (B) compounds. In each plot, the vertical line indicates the mean of the distribution.

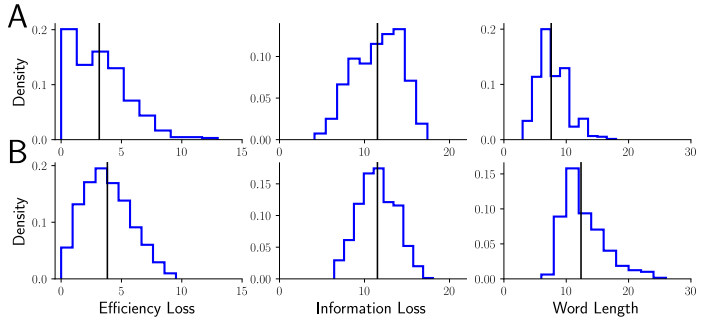

Fig. S14. Distributions of item-level efficiency loss, information loss, and word length for Finnish (A) reuse items and (B) compounds. In each plot, the vertical line indicates the mean of the distribution.

E. Taxonomic Distance Measures. In the main text, we classified both reuse items and compounds into literal and non-literal items by using hyponymic relations in the WordNet taxonomy. However, there are two limitations with this approach: 1) literal reuse items may be diminished because the original Princeton WordNet avoided linking a sense and its hyponyms to the same word (61), and 2) hyponymic relations can be taxonomic but also functional (62). Here we attempt to address the second limitation by using conceptual similarity measures to approximate hyponymic relations not captured in WordNet. To this end, we used Leacock-Chordorow similarity (63) and Wu-Palmer similarity (64). These taxonomic similarities are based on the position of the lowest common hypernym of any pair of senses, and may help to approximate hyponymic relations not encoded in WordNet while preserving encoded relations. For example, *poker* is a type of metal rod used as fire iron, and even though *poker* is not a hyponym of *rod* in WordNet, they are closely associated by their common hypernym *implement* (a piece of equipment or tool).

To assess how these similarity measures are related to item-level efficiency, we extended our binary classification in the main text. Let c be a novel sense, w be an existing word, and $c_1, ..., c_n$ be the existing senses of w. If a reuse item consists of novel sense c and existing word w, then we measured the taxonomic similarity of the item as the maximum between novel and existing senses:

$$\text{sim}(c, w) = \max\{\text{wn-sim}(c, c_i) : i = 1, ..., n\} \quad [10]$$

Here wn-sim$(\cdot, \cdot)$ is either Leacock-Chordorow or Wu-Palmer similarity. If a compound item consists of a novel sense c and head word w, we plugged c and w into the similarity measure in Equation 10 to obtain an extension of endocentricity. We excluded reuse and compound items where there is no path between c and the existing senses from our analysis.

We summarize the spearman correlations between item-level efficiency loss and Leacock-Chordorow similarity across languages and strategies in Tables S24, and summarize analogous results based on Wu-Palmer similarity in Table S25. In all cases, we observe that a novel item tends to be more efficient when it is closer to the existing senses of a reused word or head word. This is consistent with our results in the main text which showed literal items tend to be more efficient.

Group	Spearman ρ	p-value	N
English reuse	-0.202	< 0.001	414
English compound	-0.368	< 0.001	2505
French reuse	-0.254	< 0.001	485
French compound	-0.277	< 0.001	395
Finnish reuse	-0.413	< 0.001	465
Finnish compound	-0.323	< 0.001	637

Table S24. Correlations between Leacock-Chordorow similarity and item-level efficiency loss

Group	Spearman ρ	p-value	N
English reuse	-0.193	< 0.001	414
English compound	-0.383	< 0.001	2505
French reuse	-0.249	< 0.001	485
French compound	-0.292	< 0.001	395
Finnish reuse	-0.413	< 0.001	465
Finnish compound	-0.367	< 0.001	637

Table S25. Correlations between Wu-Palmer similarity and item-level efficiency loss

F. Item Frequency. Usage frequency is often correlated with the economy of use of a word form (e.g., refs. 65, 66). Here we explore whether frequency also predicts item-level efficiency in attested reuse items and compounds. Intuitively, we expect frequent items to be more optimized than infrequent items, so that the total efficiency loss aggregated over many communicative interactions tends to be lower than otherwise.

To investigate the relation between frequency and efficiency loss, we reused the historical frequencies of form-concept pairs that were used to implement the need and production distributions. We removed items that had no frequency before applying add-one smoothing, and we normalized the frequency of each item over the total frequency of both emerging and existing items in the same interval. The results for reuse items are summarized in Table S26 and the results for compounds are summarized in Table S27. Contrary to our prediction, we do not observe any significant correlation between item-level efficiency and frequency. One possible reason is that our frequency estimates are not sufficiently accurate; for example, a small amount of noise in word sense disambiguation may inflate the frequency of a peripheral sense drastically if the corresponding word is highly frequent. Another possible reason is that there are frequency-related factors beyond the scope of our account, and future work should investigate how these factors and our account are connected.

Language	Pearson ρ	p-value	Spearman ρ	p-value	N
English	−0.0606	0.242	−0.0488	0.346	375
French	−0.0549	0.209	−0.0646	0.140	524
Finnish	0.0261	0.558	0.0564	0.206	505

Table S26. Correlations between the relative frequency of reuse items and item-level efficiency

Language	Pearson ρ	p-value	Spearman ρ	p-value	N
English	−0.00599	0.775	−0.00647	0.757	2284
French	0.0770	0.177	0.0357	0.531	310
Finnish	0.0257	0.641	0.00123	0.982	332

Table S27. Correlations between the relative frequency of compounds and item-level efficiency

References

1. T Regier, C Kemp, P Kay, Word meanings across languages support efficient communication. *The Handb. Lang. Emergence* pp. 237–263 (2015).
2. N Zaslavsky, C Kemp, T Regier, N Tishby, Efficient compression in color naming and its evolution. *Proc. Natl. Acad. Sci.* **115**, 7937–7942 (2018).
3. Y Xu, T Regier, BC Malt, Historical semantic chaining and efficient communication: The case of container names. *Cogn. science* **40**, 2081–2094 (2016).
4. F Mollica, et al., The forms and meanings of grammatical markers support efficient communication. *Proc. Natl. Acad. Sci.* **118** (2021).
5. PA Chou, T Lookabaugh, RM Gray, Entropy-constrained vector quantization. *IEEE Transactions on acoustics, speech, signal processing* **37**, 31–42 (1989).
6. TM Cover, JA Thomas, *Elements of information theory*. (Wiley-Interscience), (2006).
7. T Pimentel, I Nikkarinen, K Mahowald, R Cotterell, D Blasi, How (non-)optimal is the lexicon? in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. (Association for Computational Linguistics, Online), pp. 4426–4438 (2021).
8. R Ferrer-i Cancho, C Bentz, C Seguin, Optimal coding and the origins of Zipfian laws. *J. Quant. Linguist.* **29**, 165–194 (2022).
9. H Marchand, *The categories and types of present-day English word formation: a synchronic diachronic approach*. (C.H. Beck'sche Verlagsbuchhandlung, München, Germany), (1969).
10. K Van Goethem, D Amiot, Compounds and multi-word expressions in French in *Complex Lexical Units*. (De Gruyter), pp. 307–336 (2019).
11. Y Xu, E Liu, T Regier, Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind* **4**, 57–70 (2020).
12. N Zaslavsky, M Maldonado, J Culbertson, Let's talk (efficiently) about us: Person systems achieve near-optimal compression in *Proceedings of the annual meeting of the cognitive science society*. Vol. 43, (2021).
13. S Chen, R Futrell, K Mahowald, An information-theoretic approach to the typology of spatial demonstratives. *Cognition* **240**, 105505 (2023).
14. C Kemp, T Regier, Kinship categories across languages reflect general communicative principles. *Science* **336**, 1049–1054 (2012).
15. L Raviv, A Meyer, S Lev-Ari, Larger communities create more systematic languages. *Proc. Royal Soc. B* **286**, 20191262 (2019).
16. L Raviv, A Meyer, S Lev-Ari, Compositional structure can emerge without generational transmission. *Cognition* **182**, 151–164 (2019).
17. P Downing, On the creation and use of English compound nouns. *Language* pp. 810–842 (1977).
18. V Pugacheva, F Günther, Lexical choice and word formation in a taboo game paradigm. *J. Mem. Lang.* **135**, 104477 (2024).
19. C Fellbaum, *WordNet: An electronic lexical database*. (MIT press), (1998).
20. S Bird, E Klein, E Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. (" O'Reilly Media, Inc."), (2009).
21. M Davies, *The corpus of historical American English (COHA): 400 million words, 1810-2009*. (Brigham Young University), (2002).
22. M Bevilacqua, R Navigli, Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. (Association for Computational Linguistics, Online), pp. 2854–2864 (2020).
23. JB Michel, et al., Quantitative analysis of culture using millions of digitized books. *science* **331**, 176–182 (2011).

24. CL Gagné, TL Spalding, D Schmidtke, Ladec: the large database of English compounds. *Behav. research methods* **51**, 2152–2179 (2019).
25. W Wu, D Yarowsky, Computational etymology and word emergence in *Proceedings of The 12th Language Resources and Evaluation Conference*. (European Language Resources Association, Marseille, France), (2020).
26. C Kay, J Roberts, M Samuels, I Wotherspoon, The Historical Thesaurus of English, version 4.21 (2017).
27. P Cook, JH Lau, D McCarthy, T Baldwin, Novel word-sense identification in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. pp. 1624–1635 (2014).
28. M Ryskina, E Rabinovich, T Berg-Kirkpatrick, DR Mortensen, Y Tsvetkov, Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. *Proc. Soc. for Comput. Linguist.* **3**, 43–52 (2020).
29. National Library of Finland, The Finnish N-grams 1820-2000 of the Newspaper and Periodical Corpus of the National Library of Finland (2014).
30. K Gulordava, M Baroni, A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. in *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, eds. S Pado, Y Peirsman. (Association for Computational Linguistics, Edinburgh, UK), pp. 67–71 (2011).
31. R Hu, S Li, S Liang, Diachronic sense modeling with deep contextualized word embeddings: An ecological view in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pp. 3899–3908 (2019).
32. L Frermann, M Lapata, A bayesian model of diachronic meaning change. *Transactions Assoc. for Comput. Linguist.* **4**, 31–45 (2016).
33. HJ Diller, H De Smet, J Tyrkkö, A european database of descriptors of English electronic texts. *The Eur. Engl. Messenger* **19**, 21–35 (2011).
34. O Popescu, C Strapparava, Semeval 2015, task 7: Diachronic text evaluation in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. pp. 870–878 (2015).
35. M Giulianelli, M Del Tredici, R Fernández, Analysing lexical semantic change with contextualised word representations in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds. D Jurafsky, J Chai, N Schluter, J Tetreault. (Association for Computational Linguistics, Online), pp. 3960–3973 (2020).
36. S Montariol, M Martinc, L Pivovarova, Scalable and interpretable semantic change detection in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, eds. K Toutanova, et al. (Association for Computational Linguistics, Online), pp. 4642–4652 (2021).
37. J Devlin, MW Chang, K Lee, K Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. (Association for Computational Linguistics, Minneapolis, Minnesota), pp. 4171–4186 (2019).
38. L Finkelstein, et al., Placing search in context: The concept revisited in *Proceedings of the 10th international conference on World Wide Web*. pp. 406–414 (2001).
39. E Bruni, NK Tran, M Baroni, Multimodal distributional semantics. *J. artificial intelligence research* **49**, 1–47 (2014).
40. F Hill, R Reichart, A Korhonen, Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Comput. Linguist.* **41**, 665–695 (2015).
41. E Agirre, et al., A study on similarity and relatedness using distributional and wordnet-based approaches in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 19–27 (2009).
42. GA Miller, WordNet: A lexical database for English in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. (1992).
43. N Reimers, I Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks in *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 (2019).
44. T Mikolov, K Chen, G Corrado, J Dean, Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
45. T Mikolov, E Grave, P Bojanowski, C Puhrsch, A Joulin, Advances in pre-training distributed word representations in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. (European Language Resources Association (ELRA), Miyazaki, Japan), (2018).
46. FJ Costello, T Veale, S Dunne, Using WordNet to automatically deduce relations between words in noun-noun compounds in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. (Association for Computational Linguistics, Sydney, Australia), pp. 160–167 (2006).
47. J Mitchell, M Lapata, Composition in distributional models of semantics. *Cogn. science* **34**, 1388–1429 (2010).
48. M Yazdani, M Farahmand, J Henderson, Learning semantic composition to detect non-compositionality of multiword expressions in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. (Association for Computational Linguistics, Lisbon, Portugal), pp. 1733–1742 (2015).
49. E Guevara, A regression model of adjective-noun compositionality in distributional semantics in *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, eds. R Basili, M Pennacchiotti. (Association for Computational Linguistics, Uppsala, Sweden), pp. 33–37 (2010).
50. M Marelli, CL Gagné, TL Spalding, Compounding as abstract operation in semantic space: Investigating relational effects

through a large-scale, data-driven computational model. *Cognition* **166**, 207–224 (2017).
51. DP Kingma, J Ba, Adam: A method for stochastic optimization in *ICLR*. (2015).
52. RD Luce, Detection and recognition. *Handb. mathematical psychology* pp. 103–189 (1963).
53. RM Nosofsky, Attention, similarity, and the identification–categorization relationship. *J. experimental psychology: Gen.* **115**, 39 (1986).
54. M Aro, Learning to read finnish. *Learn. to read across languages writing systems* p. 416 (2017).
55. K Lenzo, The cmu pronouncing dictionary (version 0.7 b) (2014).
56. B New, C Pallier, M Brysbaert, L Ferrand, Lexique 2: A new french lexical database. *Behav. Res. Methods, Instruments, & Comput.* **36**, 516–524 (2004).
57. W Labov, *Principles of linguistic change, volume 3: Cognitive and cultural factors*. (John Wiley & Sons) Vol. 3, (2011).
58. A Vaswani, et al., Attention is all you need. *Adv. neural information processing systems* **30** (2017).
59. S Wu, R Cotterell, M Hulden, Applying the transformer to character-level transduction in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, eds. P Merlo, J Tiedemann, R Tsarfaty. (Association for Computational Linguistics, Online), pp. 1901–1907 (2021).
60. WL Hamilton, J Leskovec, D Jurafsky, Diachronic word embeddings reveal statistical laws of semantic change in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. (Association for Computational Linguistics, Berlin, Germany), pp. 1489–1501 (2016).
61. GA Miller, Nouns in wordnet in *WordNet: An electronic lexical database*. (MIT press), pp. 23–46 (1998).
62. A Wierzbicka, Apples are not a "kind of fruit": The semantics of human categorization. *Am. Ethnol.* **11**, 313–328 (1984).
63. C Leacock, M Chodorow, GA Miller, Using corpus statistics and wordnet relations for sense identification. *Comput. Linguist.* **24**, 147–165 (1998).
64. Z Wu, M Palmer, Verbs semantics and lexical selection in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. pp. 133–138 (1994).
65. GK Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*. (Ravenio Books), (1949).
66. K Mahowald, I Dautriche, E Gibson, ST Piantadosi, Word forms are structured for efficient use. *Cogn. science* **42**, 3116–3134 (2018).