Predicting strategy choice in word formation: A case study of reuse and compounding

Aotao Xu1,2 (a26xu@cs.toronto.edu)
Charles Kemp3 (c.kemp@unimelb.edu.au)
Lea Frermann2 (lea.frermann@unimelb.edu.au)
Yang Xu1,4 (yangxu@cs.toronto.edu)

1Department of Computer Science, University of Toronto
2School of Computing and Information Systems, University of Melbourne
3School of Psychological Sciences, University of Melbourne
4Cognitive Science Program, University of Toronto

Abstract

Natural language expresses new concepts by reusing existing words or coining new ones. Previous studies have examined these word formation strategies separately through a functional lens, but it is unclear why one strategy might be preferred over another. In this study, we hypothesize that communicative and cognitive efficiency might predict the choice between lexical reuse and compounding for expressing an emerging concept. We test our hypothesis by developing a computational analysis of English word meanings that emerged over the past century. Our results suggest that strategy choice may be explained partly by a pressure for least effort. Our work contributes a novel connection between strategy choice in word formation and functional theories of language.

Keywords: the lexicon; word formation; word meaning extension; compounding; efficiency

Introduction

Natural language adapts to an evolving culture by assigning word forms to emerging concepts. For example, the English lexicon used the form car to express a version of the motor vehicle as this concept emerged during the early twentieth century. This example reflects two common strategies through which the lexicon expresses emerging meanings: reusing an existing form, such as the horse-drawn car, or combining existing forms into a compound such as motor vehicle. What factors predict the word formation strategy for expressing an emerging concept? Here we address this question with a computational analysis of historical cases of word reuse and compounding in English.

We begin our investigation from a functional perspective suggesting that language reflects communicative and cognitive efficiency (e.g., Rosch, 1978; Zipf, 1949). Existing work in linguistics has extended this perspective to account for word formation (Dressler, 2005; Stekauer, 2005), the study of patterns with which the lexicon forms new words (Marchand, 1960). This functional theory of word formation is supported by empirical work on historically attested cases of word meaning extension (Y. Xu, Malt, & Srinivasan, 2017; Ramiro, Srinivasan, Malt, & Xu, 2018), conventionalized complex words (A. Xu, Kemp, Frermann, & Xu, 2022), and loan words (Monaghan & Roberts, 2019), but these different word formation strategies have typically been examined in isolation. We extend these previous studies by suggesting that communicative and cognitive efficiency might also predict the specific strategy for expressing an emerging meaning, and in this initial work we focus on the strategy choice between reuse and compounding.

One established account of communicative efficiency originates from the observation that word frequency and word length tend to be anti-correlated (Zipf, 1949). This anti-correlation reflects an optimization of the average word length in the lexicon, hence allowing meanings frequently talked about to be expressed in shorter forms and those rarely talked about to be expressed in longer forms. Recent work has formalized this idea in information-theoretic terms and rigorously examined this tendency in attested lexicons (Ferrer-i Cancho, Bentz, & Seguin, 2022; Mollica et al., 2021; Pimentel, Nikkarinen, Mahowald, Cotterell, & Blasi, 2021). This least effort account has also been supported by experiments on artificial language learning (Kanwal, Smith, Culbertson, & Kirby, 2017) and repeated reference games (Krauss & Weinheimer, 1964; Hawkins et al., 2022). Here we hypothesize that the lexicon adapts to the communicative need of emerging meanings, i.e., the frequency with which the concept is encountered: as meanings enter the lexicon, their need should constrain the length of their corresponding word forms, so that the lexicon remains relatively compact in a Zipfian sense. In particular, since the plausible compounds for expressing a given meaning tend to be longer than plausible existing word forms (e.g., motor car is longer than car), the view of least effort predicts that the lexicon should prefer expressing high-need emerging meanings with word reuse instead of compounding.

We also consider an alternative account of efficiency rooted in earlier theories from cognitive linguistics (Lakoff, 1987; Geeraerts, 1997). Using formal categorization models and large-scale historical data, this line of work found that the lexicon tends to express novel meanings with existing words that are semantically similar (Ramiro et al., 2018; Greewal & Xu, 2021; Yu & Xu, 2021), which may reflect ease in learning (Srinivasan, Al-Mughairy, Foushee, & Barner, 2017; Floyd & Goldberg, 2021). However, there might not always be an existing word available that is sufficiently similar to a novel meaning. Certain concepts (e.g., radioactivity or nuclear winter) may be highly novel relative to the concepts ex-
pressed by the existing lexicon, thereby impeding the learning or communication of those emerging concepts. We therefore hypothesize that the lexicon should avoid reusing existing words to express highly novel meanings. Intuitively, compounding might be a more efficient choice in the case of high novelty, since a large number of conceptual combinations covers novel meanings more compactly in similarity space than existing meanings.

We illustrate our hypotheses in Figure 1. In the left panel, the least effort hypothesis postulates that high-need meanings are more likely to be expressed by reuse than by compounding. In the right panel, the novelty hypothesis postulates that high-novelty meanings should prompt compounding over reuse. In the following, we introduce our dataset of (historical) word meanings and our operationalization of need and novelty. We then evaluate the least effort and novelty hypotheses separately, and conduct a predictive analysis to test whether the two constraints can jointly predict the attested word formation strategy of meanings that emerged over the historical development of English.

Data
To investigate strategy choice in word formation, we used three sources of data. First, we used a large historical text corpus to measure communicative need and recreate historically existing lexicons. Second, we used an English dictionary that records words and their meanings to measure relative novelty. To analyze historically emerging meanings, we timestamped the first occurrence for a subset of the compounds and polysemous words.

Historical text data. We collected historical word usages from the Corpus of Historical American English (COHA; Davies, 2002). The corpus contains English text published between 1810 to 2009, spanning across four genres. To more accurately estimate the frequencies of historically rare words, we supplemented COHA with unigram frequencies provided by Google Books Ngrams (English 2012 version; Michel et al., 2011). In total, we obtained 605K lemma types from COHA and 465B word tokens between 1810 and 2012 from Google Books.

Dictionary data. We primarily used the record of established words and their meanings provided by Hu, Li, and Liang (2019), which was originally based on the English version of the Oxford Dictionary (OD). This dataset consists of 3,220 frequent polysemous words, containing sense definitions and historical sense frequencies for every decade between 1810 and 2009, which are estimated using COHA and supervised word sense disambiguation. Since their dataset only contains polysemous words, we supplemented this dataset with monosemous words from two sources: 1) we obtained 3,353 compounds that appear in the Large Database of English Compounds (LADEC; Gagné, Spalding, & Schmidtke, 2019) and the online version of OD, and 2) we obtained 43,482 COHA lemmas and their definitions that appear in archived webpages of OD. In total, this provided us with a set of 77,359 senses for 30,760 words. Example sense definition are shown in Table 4.

Word sense emergence. Since OD was created for contemporary English, we manually timestamped the first occurrence of a subset of OD word senses that emerged in the 20th century. We identified this subset by using a method of shortlisting sense definitions that contain cultural key-
words (Cook, Lau, McCarthy, & Baldwin, 2014). To identify these keywords, the first author manually selected culturally salient words that have changed the most in frequency since 1900 in COHA. Our keywords are shown in Table 1. We used these keywords to shortlist senses of polysemous words from Hu et al. (2019) and compounds from LADEC. Since most compound words are nouns, we focused on noun senses in the shortlist. After timestamping the shortlist with the OED, we obtained 67 cases of reuse and 46 compounds, as well as their exact year of first occurrence in the OED. To guarantee positive communicative need⁶, we filtered out senses that have zero frequency for two decades after their emergence. This left us with our final dataset of 63 meanings with an attested reuse form, and 46 meanings expressed with compounds.

Figure 2 shows descriptive statistics for our dataset. The senses we collected emerged during the whole span of the 20th century. The orthographic (word) lengths of senses expressed via reuse tend to be shorter than the lengths of compound words in the dataset.

### Computational methods

To evaluate our hypotheses on word formation strategies for emerging meanings, we first define the space of meanings obtained from our datasets. We then define the measures we use to quantify the properties of emerging meanings.

**Meaning space.** Let \( M = \{ m_1, m_2, \ldots, m_k \} \) be the set of meanings we will use in our analyses. We defined each meaning \( m_i \) as a word sense we obtained from OD. We represented each meaning as a 768-d vector by embedding its sense definition with Sentence-BERT (Reimers & Gurevych, 2019)³.

We ensured that the embedding space reflects genuine semantic relatedness using three datasets of pairwise word similarity ratings: WordSim-353 (Finkelstein et al., 2001), SimLex-999 (Hill, Reichart, & Korhonen, 2015), and MEN (Bruni, Tran, & Baroni, 2014). We compared our sense embeddings based on sentence-BERT to a version where we represent senses by averaging the Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018) vectors of non-stopwords in their definitions. We also include a word (rather than sense) level method using pre-trained Word2Vec embeddings. This method solely serves an upper bound in the context of our word similarity validation task, but cannot be applied in our main analyses which rely on sense-level representations.

To evaluate these embeddings, we computed the spearman correlation of human ratings and the cosine distances between word embeddings. For sense-level embeddings, we represented each word by averaging the embeddings of its senses. The evaluation results are summarized in Table 2. We see that Sentence-BERT embeddings are significantly better than the OD (Word2Vec) baseline, possibly because the former is able to better capture the dependencies among words within each definition. While OD senses do not perform as well as the word-level Word2Vec upper bound, both correlate consistently and significantly with human similarity ratings. We proceeded with the Sentence-BERT representation, and leave its improvement for future work.

**Communicative need of an emerging meaning.** The need probability of a newly emerged meaning is difficult to measure, and in our historical framework we approximate it through the observed frequency of a meaning after emergence in a diachronic corpus. Suppose a meaning \( m_i \) emerged in year \( t \), and suppose \( w \) is its attested word form. Let \( p_x(w) \) be the relative frequency of \( w \) in year \( x \) according to Google Books; let \( p_x(m|w) \) be the proportion of tokens of \( w \) expressing meaning \( m \) in year \( x \), obtained from Hu et al. (2019). Since the frequency of new words tends to be sparse when they just emerged, we estimated the communicative need of \( m_i \) in year \( t \), denoted \( f(m_i) \), by averaging its frequency over a specified time window \( X \):

\[
f(m) = \frac{1}{|X|} \sum_{x \in X} p_x(m|w)p_x(w)
\]

For robustness, we used two time windows: a historical window, where \( X = \{t, t + 1, \ldots, t + 19\} \) for each \( m_i \) and a contemporary window, where \( X = \{2000, 2001, \ldots, 2012\} \). In our analyses, we multiplied the estimated need by a constant \( 10^6 \) to avoid numerical issues.

**Relative novelty to the existing lexicon.** To quantify the novelty of an emerging meaning \( m \), relative to the existing lexicon \( L_t \), we first defined \( L_t \). We started by dividing up the time period between 1880 and 2000 into consecutive intervals of 20 years, denoted \( I_1, \ldots, I_7 \). We then identified the list of senses that existed in \( I_{t-1} \) if year \( t \in I_t \). Specifically, we automatically selected every sense \( m \) such that 1) if its attested
Table 2: Evaluation of embedding space using word similarity ratings. Each cell shows the spearman correlation between human ratings and embedding cosine distances, and embeddings with more negative correlations are better at capturing human ratings. The labels "+" and "++" denote significance at $p < 0.01$ and $p < 0.001$, respectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>OD (Sentence-BERT)</th>
<th>OD (Word2Vec)</th>
<th>Word2Vec</th>
<th>Sample size (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordSim-353 (sim)</td>
<td>-0.576 ***</td>
<td>-0.077</td>
<td>-0.836 ***</td>
<td>181</td>
</tr>
<tr>
<td>WordSim-353 (rel)</td>
<td>-0.276 ***</td>
<td>-0.072</td>
<td>-0.731 ***</td>
<td>220</td>
</tr>
<tr>
<td>SimLex-999</td>
<td>-0.289 ***</td>
<td>0.111 **</td>
<td>-0.404 ***</td>
<td>964</td>
</tr>
<tr>
<td>MEN</td>
<td>-0.494 ***</td>
<td>0.007</td>
<td>-0.837 ***</td>
<td>2659</td>
</tr>
</tbody>
</table>


distances from $m_i$ to $L_d$:

$$d(m_i, L_d) = \min_{m \in L_d} \cos-dist(m_i, m)$$

Intuitively, this measure characterizes the upper bound on the semantic similarity between $m_i$ and the existing lexicon. A low novelty score (i.e., small cosine distance to the nearest existing word) implies the opportunity to achieve more cognitive efficiency via reuse, and vice versa.\(^8\)

Results

In this section, we first show our results for each of the two hypotheses we outlined previously. We then use the predictors motivated by these hypotheses jointly to predict the historically attested choices of word formation strategy, and qualitatively analyze our predictions.

Evaluating the least effort hypothesis

Since the least effort hypothesis implies the lexicon prefers to express high-need meanings with reuse, we expected this preference to distinguish the average communicative need of reuse meanings and compound meanings. The comparison is illustrated in Figure 3. In both settings, the communicative need of meanings expressed by reuse tends to be higher than meanings expressed as compounds. To assess the statistical significance of this tendency, we compared the mean need of each group using independent t-tests. For needs estimated from historical data, we obtained $t(107) = 7.957$, $p < 0.001$. For needs estimated from contemporary data, we obtained $t(107) = 4.920$, $p < 0.001$. This provides evidence for our least effort hypothesis.

Evaluating the novelty hypothesis

Similar to the previous hypothesis, we compared the average novelty of reuse and compound meanings to assess the novelty hypothesis. The results are illustrated in Figure 4. We observe that on average, the distances for meanings expressed by compound tend to be higher than meanings expressed by reuse. Similar to the previous hypothesis, we compared the mean novelty of each group using t-tests. For distances computed at the sense level, we obtained $t(107) = -0.801$, $p = 0.425$. For distances computed at the word level, we ob-

\(^8\)This is graphically illustrated in Figure 1b.)
ings tend to have a wide range of novelty and communicative
meanings. In both plots of Figure 5, we observe that mean-
ings with respect to sense-level novelty and both of the need
satisfaction 0
pressures for communicative and cognitive efficiency ac-
tween reusing an existing word and coining a novel form.

Building on previous functional accounts of the lexicon (e.g.,
Ramiro et al., 2018; Mollica et al., 2021), we hypothesized
that communicative need and novelty can jointly predict
the attested strategy of an emerging meaning.

Since we focused on reuse and compounding, we formulate
this task as a binary prediction problem. We proceeded by
using a logistic regression model of the following form:

\[
y(m) \sim \left(1 + \exp\left(-\beta_0 + \beta_1 f_{\text{hist}}(m) + \beta_2 f_{\text{now}}(m) + \beta_3 d(m, L^\text{sense}) + \beta_4 d(m, L^\text{word})\right)\right)^{-1}
\]

Here \(y(m)\) refers to whether \(m\) is expressed by reuse or
compound, and \(f_{\text{hist}}, f_{\text{now}}\) refer to the communicative need of
\(m\) estimated using historical and contemporary frequencies.
We implemented the model using statsmodel (Seabold &
Perktold, 2010).

To evaluate the model, we trained it on word meanings that
emerged before 1950 (\(n = 66\)), and we tested it using word
meanings that emerged after 1950 (\(n = 43\)). The accuracy of
the model is 0.767, which is higher than the percentage of
the most frequent class, 0.558. We also tested the statistical
significance of each individual predictor using a Wald test.
The test results are summarized in Table 3. We observe that
\(f_{\text{hist}}(m)\) has the strongest effect size, while other predictors
have no statistical significance.

To better understand how the predictors relate to reuse and
compounding, we plotted individual meanings in Figure 5;
meanings for labelled cases are detailed in Table 4. Since
the novelty measures are highly colinear (Pearson correlation
0.978, \(p < 0.001\), \(n = 109\)), we focus on plotting mean-
ings with respect to sense-level novelty and both of the need
measures. In both plots of Figure 5, we observe that mean-
ings tend to have a wide range of novelty and communicative
need. Aligned with our predictions, most meanings with a
need higher than one standard deviation away from the mean
tend to be expressed as reuse, whereas in the other direction
most meanings tend to be expressed as compounds. In con-
trast, both low novelty and high novelty meanings are found
across both groups.

We also observe in Figure 5a) that \(f_{\text{hist}}(m)\) more strongly
separates the two attested strategies than \(f_{\text{now}}(m)\) in Figure
5b). This can be seen from the meanings with the most
extreme communicative needs across the two strategies: for
\(f_{\text{hist}}(m)\), both the most needed and least needed reuse mean-
ings (mask and edition) had higher need than their compound
counterparts (spreadsheet and kidvid); however, for \(f_{\text{now}}(m)\),
the most needed in the compound group (software) actually
has higher need than its counterpart in the reuse group (ap-
lication). This suggests communicative need and strategy
choice are sensitive to cultural changes over time.

**Discussion**

In this paper, we sought to explain the strategy choice be-
tween reusing an existing word and coining a novel form.
Building on previous functional accounts of the lexicon (e.g.,
Ramiro et al., 2018; Mollica et al., 2021), we hypothesized
that pressures for communicative and cognitive efficiency ac-
count for the choice between reuse and compounding. Specif-
ically, our least effort hypothesis predicts that new mean-
ings with high communicative need should be expressed by

![Figure 5: Scatter plots showing the communicative need and novelty of each emerging meaning. Each dot corresponds to an emerging meaning, and the extreme cases for each strategy are annotated. Dashed lines correspond to one standard deviation away from the mean over all meanings.](image)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>(\beta)</th>
<th>Z statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.5353</td>
<td>-0.977</td>
<td>0.328</td>
</tr>
<tr>
<td>(f_{\text{hist}}(m))</td>
<td>-2.8210</td>
<td>-3.488</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(f_{\text{now}}(m))</td>
<td>0.3194</td>
<td>0.618</td>
<td>0.537</td>
</tr>
<tr>
<td>(d_s(m, L^\text{sense}))</td>
<td>3.7112</td>
<td>0.144</td>
<td>0.886</td>
</tr>
<tr>
<td>(d_w(m, L^\text{word}))</td>
<td>-7.6745</td>
<td>-0.309</td>
<td>0.757</td>
</tr>
</tbody>
</table>

Table 3: Summary of logistic regression results.
Table 4: Examples of OD sense definitions and their year of emergence according to the OED.

<table>
<thead>
<tr>
<th>Word</th>
<th>Time</th>
<th>Sense Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>rally</td>
<td>1911</td>
<td>A long-distance race for motor vehicles over public roads or rough terrain, typically in several stages.</td>
</tr>
<tr>
<td>skirt</td>
<td>1912</td>
<td>A surface that conceals or protects the wheels or underside of a vehicle or aircraft.</td>
</tr>
<tr>
<td>edition</td>
<td>1934</td>
<td>A particular instance of a regular radio or television programme.</td>
</tr>
<tr>
<td>crewman</td>
<td>1937</td>
<td>A member of a group of people who work on and operate a ship, aircraft, etc., particularly one who is not an officer.</td>
</tr>
<tr>
<td>kicksorter</td>
<td>1947</td>
<td>A device for analysing electrical pulses according to amplitude.</td>
</tr>
<tr>
<td>kidvid</td>
<td>1955</td>
<td>Children’s television or video entertainment.</td>
</tr>
<tr>
<td>mask</td>
<td>1956</td>
<td>A patterned metal film used in the manufacture of microcircuits to allow selective modification of the underlying material.</td>
</tr>
<tr>
<td>software</td>
<td>1958</td>
<td>The programs and other operating information used by a computer.</td>
</tr>
<tr>
<td>application</td>
<td>1959</td>
<td>A program or piece of software designed to fulfil a particular purpose.</td>
</tr>
<tr>
<td>freeware</td>
<td>1982</td>
<td>Software that is available free of charge.</td>
</tr>
<tr>
<td>spreadsheet</td>
<td>1983</td>
<td>An electronic document in which data is arranged in the rows and columns of a grid and can be manipulated and used in calculations.</td>
</tr>
</tbody>
</table>

reuse, whereas our novelty hypothesis predicts high-novelty meanings should be expressed by compounding. Using two operationalizations of communicative need, our results provided evidence for least effort. Nonetheless, we did not find evidence that the novelty of a new meaning predicts word-formation strategy.

The lack of evidence for the novelty hypothesis may have originated from two limitations. The first one is the size of our sense dataset, which only contains 109 English senses. In our analysis of relative novelty, we assumed that the senses we collected are representative of the full range of novelty, so that the preference for expressing high-novelty meanings with compounding may distinguish compound and reuse meanings. However, it may be the case that most meanings in our dataset have relatively low novelty. In this case, our hypothesis does not distinguish the two strategies, since both compounding and reuse may transparently express the emerging meaning. Another possibility is that high novelty favours compounding as well. Previous work suggests that the plausibility of a concept is crucial in compound interpretation (Costello & Keane, 2000). If high novelty implies implausibility, then expressing a high-novelty concept with compounds may harm communication, making compounding no more efficient than reuse.

Although we motivated communicative need and novelty (or the inverse of semantic similarity) as separate predictors of strategy choice, these two factors may jointly shape the label of an emerging meaning. For example, consider meanings with low need and high novelty, and observe that the word forms used to express them will also be infrequent and semantically opaque. Since infrequent forms can be more easily forgotten and replaced (Bybee, 2006) but transparent forms are more easily retained (Floyd & Goldberg, 2021; Brunigian & Folk, 2012), infrequent and opaque existing forms may be more likely replaced by more transparent compound words. Another possibility is that when a high-need emerging meaning tends to co-occur with high-need and similar existing meanings, reused similar words may actually harm informativeness and become an unfavourable strategy (Karjus, Blythe, Kirby, Wang, & Smith, 2021).

Our methodology builds on recent work in natural language processing that utilizes large historical corpora to examine changes in word meaning (e.g., Ryskina, Rabinovich, Berg-Kirkpatrick, Mortensen, & Tsvetkov, 2020; Hu et al., 2019). In this line of work, most closely related to ours is the study by Ryskina et al. (2020). Their work showed that emerging meanings can be differentiated from existing meanings by the density and rate of change in communicative need within their semantic neighbourhoods. Our work extends their analysis by examining the difference between meanings expressed via reuse and compound meanings.

**Conclusion**

We presented an initial study on how strategies are chosen to express emerging meanings, a topic that has not been investigated rigorously in previous work on word formation. We connected strategy choice with existing theories of communicative and cognitive efficiency, and found that a pressure for least effort predicts the word formation strategy used to express new meanings. Future work may extend this study by refining the theoretical framework, considering other word formation strategies such as morphological derivation, and testing it with historical data at a larger scale.

**Acknowledgments**

We would like to thank anonymous reviewers for constructive comments on an earlier version of the paper. AX is funded partly by the U of T-UoM IRTG program. This work was supported by NSERC Discovery Grant RGPIN-2018-05872 and ARC Future Fellowship FT19010020.
References


730


