# Word formation supports efficient communication: The case of compounds

**Aotao Xu**[1,2] **(a26xu@cs.toronto.edu)**
**Charles Kemp**[3] **(c.kemp@unimelb.edu.au)**
**Lea Frermann**[2] **(lea.frermann@unimelb.edu.au)**
**Yang Xu**[1,4] **(yangxu@cs.toronto.edu)**

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

**Abstract**

Compounding is a common type of word formation extensively studied in linguistics and cognitive psychology. A growing line of research suggests that the lexicon supports efficient communication by balancing informativeness and simplicity. We propose that the formation of novel compounds reflects a similar tradeoff between informativeness and word length. We formalize this hypothesis in information-theoretic terms and develop a computational procedure to evaluate our hypothesis on English noun compounds that emerged over the past century. We find that attested compounds achieve more efficient tradeoffs between informativeness and word length than do alternative word forms. Our work demonstrates how word formation and compositionality can be connected with information-theoretic approaches to the design of the lexicon.

**Keywords:** the lexicon; word formation; compounding; compositionality; efficient communication

## Introduction

Compounding refers to a process of word formation in which speakers create novel form-meaning pairings to fill lexical gaps (Lehrer, 1970) by combining two or more existing free morphemes (Brinton & Traugott, 2005), e.g., *smartphone* is a combination of *smart* and *phone*. Compounding is not only one of the most common processes of novel word formation across the world's languages (Algeo, 1980; Wu & Yarowsky, 2018), but is also a prominent example of lexical compositionality in natural language. For these reasons, the analysis of compounds has garnered considerable interest in linguistics (Blutner, 1998; Jackendoff, 2010), cognitive psychology (Medin & Shoben, 1988; Costello & Keane, 2000; Barsalou, Simmons, Barbey, & Wilson, 2003), and computational studies of language (Mitchell & Lapata, 2010; Reddy, McCarthy, & Manandhar, 2011; Yazdani, Farahmand, & Henderson, 2015; Salehi, Cook, & Baldwin, 2015; Marelli & Baroni, 2015). Here we present a framework to investigate the formation of English compounds through the lens of general communicative principles.

One known factor constraining word formation is that the word form should deliver its underlying meaning while avoiding redundancy. Štekauer (2005) proposes that two "universal, contradictory tendencies" underlie word formation: "economy of speech" and "explicitness of expression". In their study of affixation, Marelli and Baroni (2015) suggest that the head of a derived word should be close to the full word in meaning to facilitate interpretation for the listener. Lieber (2004) describes the Redundancy Restriction, which states that affixes containing semantic information already available in the head word should not be added. Costello and Keane (2000) also argue that the speaker should carefully choose the constituents so that they are both necessary and sufficient for signifying the intended meaning. However, to our knowledge there exists no work that comprehensively examines these general principles in the historical formation of novel compounds.

Our theoretical starting point is a growing line of research suggesting that the lexicon is structured to support efficient communication, which we briefly review here focusing on two aspects: word meaning and word length. Recent work provides evidence that word meanings efficiently trade off complexity (or the opposite of simplicity) against informativeness (Regier, Kemp, & Kay, 2015; Zaslavsky, Kemp, Regier, & Tishby, 2018). Within a semantic domain, the complexity of a set of semantic categories is based on its size or description length, though the word forms that label each category are typically not considered. A related line of work is rooted in the study of Zipf (1949), who hypothesized that more frequent words tend to be short in form due to the need to communicate successfully while minimizing effort. More recent work suggests that word length in natural languages may not be optimized solely with respect to frequency. For instance, Pimentel, Nikkarinen, Mahowald, Cotterell, and Blasi (2021) shows that morphological composition can result in longer word forms. Similarly, the principle of uniform information density (Jaeger, 2006; Levy & Jaeger, 2007) predicts that words should be long in unpredictable contexts to make optimal use of the communication channel (Piantadosi, Tily, & Gibson, 2011). Here we ask whether novel word formation is shaped for efficient communication, and as an initial case study we focus on English noun compounds that emerged over the past century.

We hypothesize that the formation of novel compound words should near-optimally trade off informativeness, a measure of the ease of interpretability of a word, and word length, a measure of the effort in uttering a word (Zipf, 1949). These two dimensions trade off against each other when adding morphemes and increasing the length of a novel form allows its meaning to be specified more precisely.

To test this hypothesis, we develop a computational framework that extends the existing line of work on efficient com-



In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.* 

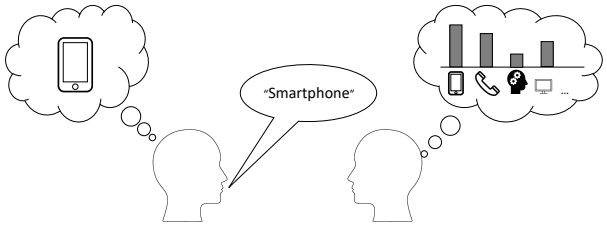


Figure 1: An illustration of the communicative scenario at the heart of our framework. A speaker (left) intends to express an emerging referent by coining a new compound word by combining *smart* and *phone*. A listener (right) infers the intended meaning upon hearing the form. Grey bars signify the probabilities of possible meanings.

munication (Regier et al., 2015; Kemp, Xu, & Regier, 2018; Zaslavsky et al., 2018; Piantadosi et al., 2011) to general principles for word formation. We illustrate this framework in Figure 1. Given an emerging intended referent such as "a multi-functional mobile phone", the speaker might choose to express that referent by either reusing an existing word form (e.g., *phone*) or coining a new term (e.g., *smartphone*). The listener in turn needs to reconstruct the intended meaning given the form uttered by the speaker. We hypothesise that novel compounds allow a near-optimal tradeoff between utterance length and reconstruction of the intended meaning. To evaluate this hypothesis, we develop a computational procedure for testing attested English compounds against a repertoire of plausible alternate word forms.

## Computational framework

We describe our computational framework in three steps. First, we explain the basic assumptions. Second, we formulate the two competing dimensions that form the bases of our efficiency hypothesis. Third, we formulate the hypothesis drawing on these dimensions.

### Assumptions

For simplicity, we assume the spaces of forms and meanings are discrete spaces. Let $\mathcal{L}_t = \{(w_1, m_1), ..., (w_q, m_q)\}$ be the lexicon of a language at time $t$ consisting of form-meaning pairs. In our work, we analyze how speakers use elements of $\mathcal{L}_t$ to express meanings with attested coinages in a future lexicon $\mathcal{L}_{t+1}$.

We start from a simple communicative scenario involving a speaker and a listener (Figure 1). Suppose at time $t$, both speaker and listener share the same lexicon $\mathcal{L}_t$. The scenario begins with a novel target meaning $m \in \mathcal{L}_{t+1} - \mathcal{L}_t$ which the speaker wishes to convey. To do so, the speaker chooses a form $w \in \mathcal{L}_t$ or creates a compound based on $\mathcal{L}_t$ (i.e., $w \in \mathcal{L}_t^*$). The form is observed by the listener who attempts to reconstruct the intended meaning of $w$.

Let $P_S(M|W)$ and $P_L(M|W)$ represent the speaker's and the listener's respective probabilistic interpretation of word forms at time $t$. We assume $P_S$ and $P_L$ are identically distributed, except that for the speaker $m$ is the only intended meaning of $w$ (i.e., $P_S(m|w) = 1$), whereas for the listener the meaning of $w$ remains uncertain. We also restrict the support of each distribution to the meanings attested in $\mathcal{L}_t$ and $\mathcal{L}_{t+1}$.

### Formulation of communicative cost

We formulate informativeness by a measure of communicative cost. Intuitively, an informative word should yield a low communicative cost. The communicative cost of a word form $w$ with respect to the target meaning $m$ is the amount of information lost when the listener reconstructs the meaning from $w$. A common distortion measure in the efficient communication literature is the KL divergence between the speaker's intended message and the listener's reconstruction (Regier et al., 2015; Zaslavsky et al., 2018). We thus define communicative cost as the KL divergence between $P_S(M|w)$ and $P_L(M|w)$:

$$\begin{aligned} & D_{KL}(P_S(M|w)||P_L(M|w)) \\ &= \sum_{m' \in M} P_S(m'|w) \log \frac{P_S(m'|w)}{P_L(m'|w)} \\ &= -\log P_L(m|w) \end{aligned} \tag{1}$$

The final equality follows from the simplifying assumptions above: the only time $P_S(m'|w)$ is positive is when $m' = m$. Thus the KL divergence is equivalent to a surprisal term, which captures how much information about the target meaning $m$ is lost when the listener tries to reconstruct it from the word form $w$.

We formalize the listener's probabilistic interpretation of $w$, $P_L(M|w)$ using the similarity choice model (Luce, 1963; Nosofsky, 1986), which calculates the probability of a response $m_j$ given a stimulus $w_i$ using the following equation:

$$P(m_j|w_i) = \frac{sim(m_j, w_i)}{\sum_k sim(m_k, w_i)} \tag{2}$$

where $sim(m_j, w_i)$ is the similarity between $m_j$ and $w_i$ in some psychological space, and the denominator sums over some set of responses. We define similarity using the Gaussian decay function (Nosofsky, 1986):

$$sim(m_j, w_i) = \exp(-d_{ij}^2) \tag{3}$$

where $d_{ij}$ is the Euclidean distance between $x_i$ and $x_j$, the representations of $w_i$ and $m_j$ in the psychological space.

In our case, every response is a meaning $m_j \in \mathcal{L}_{t+1} \cup \mathcal{L}_t$, and the stimulus is a potential word form $w_i \in \mathcal{L}_t^*$. We instantiate the psychological space using word embeddings. Since every $m_j$ corresponds to a word $w_j \in \mathcal{L}_{t+1} \cup \mathcal{L}_t$, we set $x_j$ as the embedding of $w_j$. We use a composition function (Mitchell & Lapata, 2010; Yazdani et al., 2015) to represent potential word forms $w_i = w^1 ... w^n \in \mathcal{L}^n$:

$$x_i = f(v_1, v_2, ..., v_n) \tag{4}$$

where $v_k$ is the embedding of constituent $w^k$, and $f$ is some composition function.

### Formulation of complexity

Complexity reflects the effort required to utter a word. There are at least three measures that reflect speaker effort: word frequency which is correlated with processing ease (Papesh & Goldinger, 2012), well-formedness which predicts ease of articulation (Kawasaki & Ohala, 1980), and word length which yields the number of sounds to be produced. Word length has been consistently shown to correlate with the amount of information conveyed by the form (Piantadosi et al., 2011; Lewis & Frank, 2016). In our communicative scenario, the listener relies on information provided by sub-lexical components of the form $w$ to reconstruct the speaker's intended meaning. As the number of these components contributes to length, we use the length of $w$, $len(w)$, as our measure of complexity.

### Efficiency hypothesis

Under our theoretical framework, communicative cost and word length compete against each other. If more sub-lexical components with information useful for guessing $m$ are conveyed to the listener, the speaker will spend more effort, and vice versa. Extending previous work on efficient communication (Piantadosi et al., 2011; Kemp et al., 2018; Zaslavsky et al., 2018), we hypothesize that communicative cost and word length should optimally trade-off against each other in the process of word formation. That is, for some trade-off parameter $\beta$, the word form $w$ should optimize the following objective:

$$\arg\min_{w} -\log P_L(m|w) + \beta len(w) \tag{5}$$

Specifically, for an attested compound $w$ with meaning $m$, our hypothesis predicts it will near-optimally trade off between its form complexity, $len(w)$, and its communicative cost, $-\log P_L(m|w)$.

## Materials and methods

Here we describe how we operationalized the framework. We first describe how we obtained historical lexicons $\mathcal{L}_t$ and calculated communicative cost and complexity. We then describe how we tested our hypothesis by comparing historically attested compounds against alternative word forms.

### A dataset of emerging compounds

As our source of lexicons and attested compounds, we used the Historical Thesaurus of English, HTE (Kay, Roberts, Samuels, & Wotherspoon, 2017), and the Large Database of English Compounds, LADEC (Gagné, Spalding, & Schmidtke, 2019).

**HTE.** The HTE provides 793,734 word senses along with their word form and dates of first and last appearance in historical records. This allowed us to define $\mathcal{L}_t$ for every novel meaning $m \in \mathcal{L}_{t+1}$. We only included entries that are fully alphabetic or are bigrams separated by a space or hyphen, and removed proper nouns.

**LADEC.** The LADEC is a database of 8,957 adjective-noun and noun-noun closed English compounds. The list of compounds is based on the Brown, CELEX and COCA corpora, as well as phrases provided by Costello, Veale, and Dunne (2006). In addition to a separation into head and modifier, every compound is labeled with a meaning predictability judgement. Meaning predictability is obtained by asking human participants how predictable a compound's meaning is from its parts on a scale of 0 to 100. We retrieved every entry's year of emergence as this compound's earliest emerging sense in the HTE.

In our experiments, we defined a novel meaning $m$ at year $t$ as the word meaning of a compound in LADEC that emerged at year $t+1$. For the same $m \in \mathcal{L}_{t+1}$, we defined $\mathcal{L}_t$ as all entries that existed at year $t$ according to the HTE. We analyzed compounds that emerged after 1900 and whose constituents are in the corresponding $\mathcal{L}_t$, obtaining a historical set of 230 compounds. The size of lexicon $\mathcal{L}_t$ ranges from 145k to 170k.

### Quantification of communicative cost and complexity

We now quantify the two competing dimensions of communicative cost and complexity, assuring scalability of their evaluation to a large set of alternate word forms.

**Communicative cost.** We instantiate the composition function $f$ in Equation 4 as an additive function (Mitchell & Lapata, 2008):

$$x_i = \sum_{k=1}^{n} v_k \tag{6}$$

which has proved widely successful despite its simplicity, and is highly scalable (Shen et al., 2018). We tested two kinds of word embeddings as our distributional semantic model: 1) pre-trained Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013) and 2) pre-trained subword-informed fastText embeddings (Bojanowski, Grave, Joulin, & Mikolov, 2017), both trained on Common Crawl (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018). While fastText can embed every word using character n-grams, Word2Vec embeddings are available for 215 (out of 230) LADEC compounds.

We need to compute Equation 2 for a large number of word forms, such that computing the denominator becomes very expensive. We thus considered the following simplified similarity choice model:

$$P(m_j|w_i) \propto sim(m_j, w_i) \tag{7}$$

That is, we assume that the denominator is roughly a constant. We validated both full and simplified similarity choice models using human judgements of meaning predictability in LADEC by setting $\mathcal{L}_{t+1} \cup \mathcal{L}_t$ as all HTE entries existing in year 2000 and all entries in LADEC.

**Complexity.** We used two measures of word length: 1) orthographic length as number of characters and 2) phonemic length. For phonemic length, we used the CMU Pronouncing Dictionary (Lenzo, 2014) which contains 133,854 entries. After lower-casing and intersecting with the HTE, the size of lexicon $\mathcal{L}_t$ spans between 33k and 35k. To compute the phonemic length of a compound not in the dictionary, we

took the sum of the phonemic lengths of its constituents, obtaining phonemic transcriptions for 220 LADEC compounds.

### Procedures for testing near-optimality

We hypothesized attested compounds are near-optimal with respect to the two measures defined in the previous section. Here we define near-optimality and the statistical tests that assess the extent to which our hypothesis holds.

For a given target meaning $m$, the optimality of a form is defined by Equation 5. However, searching over even $\mathcal{L}_t \cup (\mathcal{L}_t \times \mathcal{L}_t)$ is intractable, so we resort to exhausting a large, plausible subset of alternate forms. We used three subsets of $\mathcal{L}_t^*$: 1) reusing all forms in $\mathcal{L}_t$, 2) a set of forms that are related to the attested form for $m$, and 3) greedy search. To create 2), we first selected the $k = 5$ closest neighbours of the head word of the attested compound for $m$, including the head word itself; then we took the Cartesian product between this set and $\mathcal{L}_t$. To create 3), we first greedily selected the top $k = 5$ forms in $\mathcal{L}_t$ that minimize the objective function in Equation 5. We set the trade-off parameter $\beta = \{0, 0.005, 0.025, 0.05\}$ as determined in preliminary experiments (larger $\beta$ reduces the optimal form to a single letter or short acronym). We then took the Cartesian product between this set and $\mathcal{L}_t$. The final size of the subsets varies from 300k to 3m.

We approximated the theoretically optimal set of forms by using the forms that exist on the Pareto frontier of this large sample. Inspired by work in multi-objective evolutionary algorithms, we used the non-dominated (ND) rank which is used to quantify the fitness of an individual in a population along multiple objective functions (Jensen, 2003; Tian, Wang, Zhang, & Jin, 2017).[1] We defined the near-optimality of a form as its ND rank within the whole set of alternatives.

To assess the overall near-optimality of attested compounds, we tested whether the distribution of ND ranks over the set of attested compounds 1) is significantly higher on average than samples of alternatives and 2) has a significantly non-zero skew towards lower ranks, suggesting attested compounds are highly likely to have above average rank. We assessed the first hypothesis via a permutation test that randomly swapped every attested form with an alternate form generated for the same target meaning 100,000 times. For a more focused comparison, we also randomly swapped every form with a compound created by replacing the constituents of the attested form with their $k = 5$ nearest neighbours in $\mathcal{L}_t$.

## Results

We first validate our model of communicative cost against human judgement. We then test our efficiency hypothesis using historical English compounds and interpret the results. For the second part, we present results for fastText and orthographic length only due to space limitations, but we achieved similar results using Word2Vec and phonemic length.

### Evaluation of communicative cost

We evaluated our model of communicative cost against human judgements of compound meaning predictability provided by LADEC (Gagné et al., 2019). For each compound, we computed four types of communicative cost by taking the product of {fastText, Word2Vec} embeddings and {full, simplified} similarity choice models. Using fastText and the full model, the Pearson correlation between communicative cost and meaning predictability is $-0.408$, $p < 0.001$, $N = 8299$; using Word2Vec and the full model, the correlation is $-0.408$, $p < 0.001$, $N = 7085$; using fastText and the simplified model, the correlation is $-0.412$, $p < 0.001$, $N = 8299$; using Word2Vec and the simplified model, the correlation is $-0.399$, $p < 0.001$, $N = 7085$.[2] We observe a statistically significant correlation between our model and meaning predictability in all cases, providing empirical justification for our model of communicative cost. Moreover, we observe performance based on full and simplified similarity choice models are comparable.

For a more careful comparison of the two choice models, we correlated communicative costs given by the two choice models. Using fastText, the Pearson correlation is 0.955, $p < 0.001$, $N = 8299$; using Word2Vec, the correlation is 0.631, $p < 0.001$, $N = 7085$. We observe the simplified model is strongly correlated with the full model when using fastText, albeit less so when using Word2Vec. For this reason, we used the simplified similarity choice model in our analyses.

### Evaluation of efficiency hypothesis

We assessed the near-optimality of historical attested compounds that emerged during the 20th century. The mean rank of the attested compounds is 23812.98 ($p < 0.001, n = 230$). The moment coefficient of skewness (Doane & Seward, 2011) of the distribution is 2.36 ($p < 0.001, n = 230$), indicating a significant right skew. Figure 2 shows this distribution along with the rank distribution of a permuted sample. We observe that relative to a rank distribution of randomly sampled forms, the attested distribution concentrates around high ranks. In a more focused comparison, we examine whether these attested compounds are more optimal than near-synonym forms. Similar to the previous comparison, the mean rank of the distribution is significantly higher than the mean rank of near-synonym sets ($p < 0.001, n = 230$).

To test whether both dimensions contribute to the overall near-optimality of attested compounds, we performed an additional analysis where we repeated the first set of comparisons between attested compounds against the general set of alternatives by controlling for each dimension. Specifically, for each attested compound, we generated alternatives using

[1] In our case, a word form dominates another form if it is 1) shorter or more informative and 2) not worse in the other dimension than the latter. Given that the population of forms is partitioned into equivalent classes in which no form dominates another, the ND rank of a form is the number of classes whose members dominate the form. We computed ND ranks with the log-linear ND sorting algorithm for 2 objectives by Jensen (2003).

[2] The difference in sample size is due to intersecting with embedding vocabularies. Correlations do not change significantly if the same subset of LADEC was used.

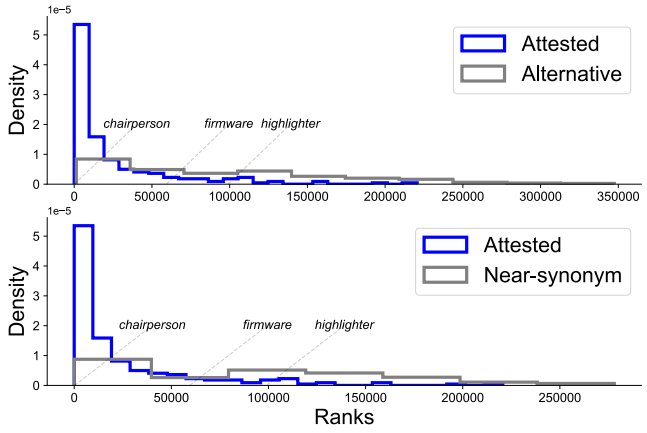


Figure 2: Comparisons of ND ranks between attested compounds and the full set of alternatives and a near-synonym subset. The ranks of a sample of compounds are annotated.

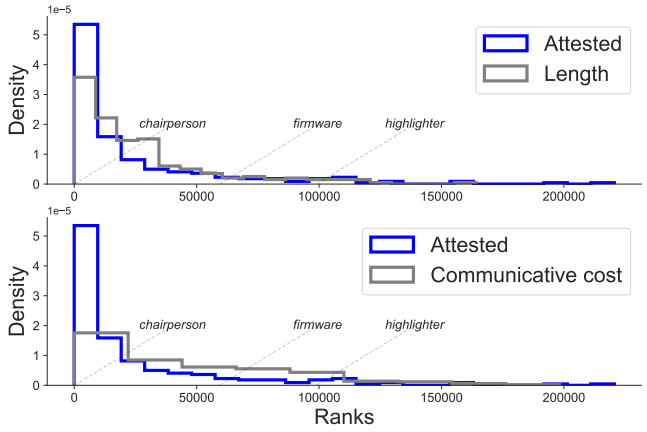


Figure 3: Comparison of ND ranks between attested compounds and alternatives controlling for word length (top) and communicative cost (bottom). The ranks of a sample of compounds are annotated.

the same procedure, and then discarded all forms that are longer or have higher communicative cost than the attested compound. The results are summarized in Figure 3. When controlling for length, the mean rank of the attested distribution is 23812.98 ($p = 0.0004, n = 230$), and the moment coefficient of skewness is 2.36 ($p < 0.001, n = 230$); when controlling for communicative cost, the mean rank of the attested distribution is 23812.92 ($p < 0.0001, n = 230$), and the moment coefficient of skewness is 2.36 ($p < 0.001, n = 230$). We observe that the location and shape of the attested rank distribution is robust when controlling for either dimension. These results suggest that both dimensions contribute to the near-optimality of attested compounds.

Figure 4 compares a sample of attested compounds to alternate word forms with respect to the respective target meanings. Qualitatively, a word form is near-optimal if it is close to the optimal frontier. Within this sample, we observe some

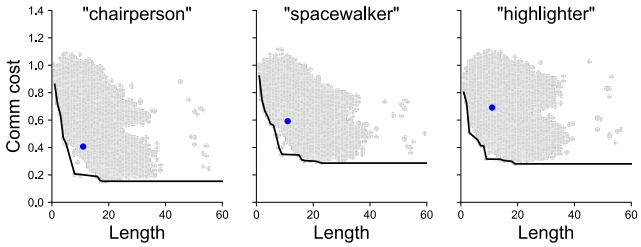


Figure 4: Qualitative comparison between attested compounds and alternatives for a selected sample of all 230 compounds. Target meanings are shown above each plot; grey dots correspond to alternate forms, and blue dots correspond to attested compounds. Black lines indicate the optimal frontiers obtained by interpolating optimal forms. The y-axes are proportional to the number of bits lost in communication.

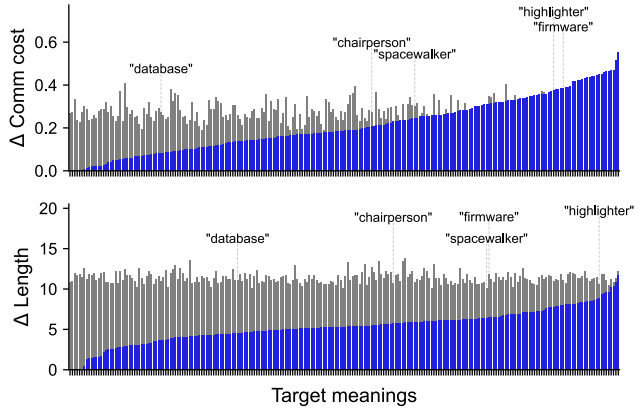


Figure 5: Distance of all 230 attested compounds (blue) and averages over alternative forms (grey) to the Pareto frontier along the dimensions of communicative cost (top) and complexity (bottom). In each subplot, each pair of vertically aligned bars corresponds to a target meaning.

attested compounds (e.g., *chairperson*) are very close to the frontier, but the others can be relatively far from optimal (e.g., *highlighter*); these correspond to the center and tail of the attested distributions in Figures 2 and 3. Figure 5 summarizes the location of the attested compounds along each dimension. All 230 attested compounds are closer to the frontier than alternatives along the length dimension, but the same tendency is relatively weaker for communicative cost as only 174 are closer to the frontier than alternatives on average.

Table 1 shows the optimal set of word forms with respect to two target meanings ("database" and "firmware"). We observe that long optimal forms tend to be more semantically transparent and vice versa (e.g., with respect to "database", *searchable-data* vs. *db*). This suggests that length and communicative cost trade off against each other, and that our method captures intuitions of informativeness for alternate word forms in addition to attested ones. Table 2 shows a sample of near-synonym alternatives. Here we see a clear trend

| |
|---|
| **database** a, db, data, data-db, list-data, data-query, archive-query, searchable-data |
| **firmware** a, xi, mem, cw-os, kit-os, web-rom, otas-rom, piezo-rom, karoro-rom, pedanty-rom, decoding-rom, algorithm-rom, flash-hardware, resetter-kernel, updater-software, kernite-microcode, kernelly-microcode, polyphone-microcode, caracoling-microcode, dressership-microcode, petticoating-microcode, compatibility-microcode |

Table 1: Pareto sets with respect to two target meanings (database and firmware). The alternate forms are sorted by length, and forms consisting of two constituents are shown using a dash.

| |
|---|
| **data-base**, data-ground, data-core, statistics-structure, analysis-structure, information-structure, analysis-core, information-base, analytic-score, information-core, statistics-core, analytics-ground, information-basing |
| client-ware, software, company-ware, **firm-ware**, steadfast-ware, client-pottery, company-whiteware, firm-pottery, company-pottery, soft-whiteware, soft-stoneware, steadfast-wares, firm-stoneware, steadfast-stoneware |

Table 2: Near-synonym set for database and firmware. The target meaning is represented by its attested form in English. Every row is sorted by ND rank in descending order. Forms consisting of two constituents are shown using a dash.

that attested compounds tend to avoid redundant, uninformative morphemes (e.g., *firm-ware* vs. *company-whiteware*). Taken together, these comparisons provide evidence for our hypothesis that attested compounds should near-optimally trade-off word length and communicative cost.

## Discussion

Existing literature on word formation proposes that novel words should be informative while avoiding redundancy (Lieber, 2004; Costello & Keane, 2000). By comparing attested compounds to alternatives and showing that the latter tend to be redundant, our work corroborates earlier theories of word formation. However, our theory goes beyond redundancy avoidance by also predicting that even if two potential constituents are similarly informative, the shorter one is preferable (e.g., *data-base* vs. *information-base*).

Our formulation of communicative cost is related to independently derived formalisms. By defining meaning using formal semantics, Blutner (1998) proposes a similar formulation of communicative cost. Blutner's proposal applies more generally to all types of utterances and is not derived from a communicative scenario. Intuitively, one might expect that the form-meaning systematicity of a language (Monaghan, Shillcock, Christiansen, & Kirby, 2014; Pimentel, McCarthy, Blasi, Roark, & Cotterell, 2019) relates to the communicative cost of a form.[3] By taking the average of Equation 1, we see that communicative cost is inversely proportional to the mutual information between meaning and form, $I(M;W)$, an information-theoretic definition of systematicity (Pimentel et al., 2019). This connection mirrors a link between language use and systematicity previously found in controlled experiments (Nölle, Staib, Fusaroli, & Tylén, 2018).

We note that our efficiency hypothesis is based on an one-shot communicative scenario where the need to communicate a specific novel meaning arises. Thus our hypothesis primarily pertains to synchronic word formation and not historical language change, even though we used historical data of compounds to recreate communicative scenarios. Nonetheless, historical changes may have confounded our results. For example, the communicative cost of certain words may have changed over time due to semantic change (e.g., *db* used to signify "decibel" in the early 1900s, but has now acquired the sense "database"). One potential way to control for this factor is to repeat the analyses using embeddings trained on historical corpora (Hamilton, Leskovec, & Jurafsky, 2016; Dubossarsky, Hengchen, Tahmasebi, & Schlechtweg, 2019).

Although we focused on word length and informativeness, our theory does not preclude other factors known to restrict the plausibility of compounds. For instance, the frequency of the constituents predicts the ease of compound processing (Juhasz, Starr, Inhoff, & Placke, 2003) and the character-level bigram probability at the boundary of constituents affects compound parsing (Gagné et al., 2019), which may explain why certain short and informative alternative forms (e.g., *list-data* for database) are never attested.

## Conclusion

We presented a formal framework for connecting two areas of research that have had little overlap so far: communicative efficiency of the lexicon, and the formation of novel word forms. Using an English dataset of historical compounds over the past century, we provided evidence that emerging lexical compounds support efficient communication by trading off informativeness against word length.

Our work helps to explain why certain word forms are chosen to fill lexical gaps ahead of other logically possible alternatives, and our framework offers an information-theoretic account for one important function of lexical compositionality: to support efficient communication about emerging items. We are optimistic that this framework will provide a foundation for future analyses that account for other processes of word formation in English and other languages.

[3]For example, in English the head word of a compound tends to signify its word class (Jackendoff, 2010), which helps reduce the uncertainty of the compound's meaning.

## Acknowledgments

We would like to thank Aida Ramezani, Jade Yu, Katie Warburton, and Zhewei Sun 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, SSHRC Insight Grant #435190272, ARC Future Fellowship FT19010020, and by an Ontario ERA Award to YX.

## References

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

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, *7*(2), 84–91.

Blutner, R. (1998). Lexical pragmatics. *Journal of semantics*, *15*(2), 115–162.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

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

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

Costello, F. J., Veale, T., & Dunne, S. (2006, July). Using WordNet to automatically deduce relations between words in noun-noun compounds. In *Proceedings of the COLING/ACL 2006 main conference poster sessions* (pp. 160–167). Sydney, Australia: Association for Computational Linguistics.

Doane, D. P., & Seward, L. E. (2011). Measuring skewness: a forgotten statistic? *Journal of statistics education*, *19*(2).

Dubossarsky, H., Hengchen, S., Tahmasebi, N., & Schlechtweg, D. (2019, July). Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 457–470). Florence, Italy: Association for Computational Linguistics. doi: 10.18653/v1/P19-1044

Gagné, C. L., Spalding, T. L., & Schmidtke, D. (2019). LADEC: the large database of english compounds. *Behavior research methods*, *51*(5), 2152–2179.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016, August). 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)* (pp. 1489–1501). Berlin, Germany: Association for Computational Linguistics. doi: 10.18653/v1/P16-1141

Jackendoff, R. (2010). *Meaning and the lexicon: the parallel architecture 1975-2010*. OUP Oxford.

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Unpublished doctoral dissertation, Stanford University Stanford, CA.

Jensen, M. T. (2003). Reducing the run-time complexity of multiobjective EAs: The NSGA-II and other algorithms. *IEEE Transactions on Evolutionary Computation*, *7*(5), 503–515.

Juhasz, B. J., Starr, M. S., Inhoff, A. W., & Placke, L. (2003). The effects of morphology on the processing of compound words: Evidence from naming, lexical decisions and eye fixations. *British journal of psychology*, *94*(2), 223–244.

Kawasaki, H., & Ohala, J. J. (1980). Acoustic basis for universal constraints on sound sequences. *The Journal of the Acoustical Society of America*, *68*(S1), S33–S33.

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

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

Lehrer, A. (1970). Notes on lexical gaps. *Journal of Linguistics*, *6*(2), 257–261.

Lenzo, K. (2014). *The CMU pronouncing dictionary (version 0.7 b)*. Carnegie Mellon University.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, *19*, 849.

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195.

Lieber, R. (2004). Combinability and the correspondence between form and meaning. In *Morphology and lexical semantics* (p. 154–177). Cambridge University Press. doi: 10.1017/CBO9780511486296.007

Luce, R. D. (1963). Detection and recognition. *Handbook of mathematical psychology*, 103–189.

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

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, *20*(2), 158–190.

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

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018, May). Advances in pre-training distributed word representations. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Mitchell, J., & Lapata, M. (2008, June). Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT* (pp. 236–244). Columbus, Ohio: Association for Computational Linguistics.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, *34*(8), 1388–1429.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130299.

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104.

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

Papesh, M. H., & Goldinger, S. D. (2012). Pupil-blah-metry: Cognitive effort in speech planning reflected by pupil dilation. *Attention, Perception, & Psychophysics*, *74*(4), 754–765.

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

Pimentel, T., McCarthy, A. D., Blasi, D., Roark, B., & Cotterell, R. (2019, July). Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1751–1764). Florence, Italy: Association for Computational Linguistics. doi: 10.18653/v1/P19-1171

Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., & Blasi, D. (2021, June). 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* (pp. 4426–4438). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.350

Reddy, S., McCarthy, D., & Manandhar, S. (2011, November). An empirical study on compositionality in compound nouns. In *Proceedings of 5th international joint conference on natural language processing* (pp. 210–218). Chiang Mai, Thailand: Asian Federation of Natural Language Processing.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. *The Handbook of Language Emergence*, 237–263.

Salehi, B., Cook, P., & Baldwin, T. (2015, May–June). A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 977–983). Denver, Colorado: Association for Computational Linguistics. doi: 10.3115/v1/N15-1099

Shen, D., Wang, G., Wang, W., Min, M. R., Su, Q., Zhang, Y., ... Carin, L. (2018, July). Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 440–450). Melbourne, Australia: Association for Computational Linguistics. doi: 10.18653/v1/P18-1041

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

Tian, Y., Wang, H., Zhang, X., & Jin, Y. (2017). Effectiveness and efficiency of non-dominated sorting for evolutionary multi-and many-objective optimization. *Complex & Intelligent Systems*, *3*(4), 247–263.

Wu, W., & Yarowsky, D. (2018, May). Massively translingual compound analysis and translation discovery. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Yazdani, M., Farahmand, M., & Henderson, J. (2015, September). Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1733–1742). Lisbon, Portugal: Association for Computational Linguistics. doi: 10.18653/v1/D15-1201

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

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