A computational analysis of compound expressions for emerging concepts

by

Aotao Xu

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Graduate Department of Computer Science
University of Toronto
School of Computing and Information Systems
University of Melbourne

Abstract

A computational analysis of compound expressions for emerging concepts

Aotao Xu

Doctor of Philosophy

Graduate Department of Computer Science

University of Toronto

School of Computing and Information Systems

University of Melbourne

2026

New compounds enable the expression of emerging concepts and ideas, making them highly prevalent in the lexicons of many languages around the world. In this thesis, I present a large-scale, computational analysis of compound expressions that builds on insights from computational cognitive science. The analysis investigates two issues that remain underexplored in the extensive literature on compounds. First, it is unclear how compound labels compete with reuse items, which are labels for emerging concepts that correspond to existing words instead of new combinations. Second, non-literal compound labels are relatively common, but there are few accounts regarding their creation.

This thesis consists of three studies. In the first study, I propose that compounds and reuse items reflect opposing pressures for informativeness and ease of use, and expressions for emerging concepts are jointly shaped by these pressures to support efficient communication. I formulate this proposal in computational terms, and I use it to analyze compounds and reuse items that emerged in English, French, and Finnish during the last century. I show that these attested labels of emerging concepts are more efficient than hypothetical labels, but literal labels are better explained by the present account than non-literal labels. In the second study, I extend the previous study by examining preferences for lexicalization strategies. I show that communicative need predicts a preference

for using new compounds over reusing existing words, which is consistent with a drive for ease of use in communication. In the third study, I propose that considering analogy in addition to composition provides a better account of non-literal compound labels. I formulate a computational model that integrates these cognitive processes to predict the compound label of a given concept, and I show that the model more accurately reconstructs both literal and non-literal compound labels that are attested in English, Chinese, and German than simpler baseline models.

Taken together, my work establishes a new connection between compound expressions and computational analyses of language in cognitive science. The approaches developed in this thesis provide a foundation for future computational analyses of compound expressions and other aspects of lexical evolution.

To my parents, whose toil and sweat have enabled my education.

Acknowledgements

I would like to express my gratitude to everyone who has contributed to the creation of this thesis. First and foremost, I am greatly indebted to my supervisors, Yang Xu, Lea Frermann, and Charles Kemp. I am very grateful for your advice and support over the past years, without which I could not have completed this thesis. In particular, I want to thank Yang for introducing me to cognitive science research back when I was an undergrad, as well as for joining weekly Toronto-Melbourne meetings in the evening.

I would like to thank my thesis examiners, Shane Steinert-Threlkeld, Richard Futrell, and Guillaume Thomas for positive comments and suggestions. I would also like to thank my Melbourne committee chairs, George Buchanan and Udaya Parampalli, for general suggestions during progress review meetings.

Through a joint education placement, I had the privilege of overlapping with communities based in different universities. I would like to thank the Language, Cognition, and Computation group at the University of Toronto, the NLP group at the University of Melbourne, and the Complex Human Data Hub (CHDH) for the chance to learn from so many brilliant minds. I would also like to thank the wonderful members of CHDH for fostering a strong sense of community; this list of people includes Andy Perfors, Simon Cropper, Elizabeth Davie, Frank Mollica, Simon De Deyne, Temuulen Khishigsuren, Katie Warburton, Andrew Wang, Jie Sun, Raina Zhang, Manikya Alister, Allegra Micomonaco, Jonathan Januar and many more.

Finally, I would like to express deep gratitude to my family and friends for offering emotional and social support along this journey, as it has greatly boosted my resilience.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Context of the study

Compounds are generally known as words that are combinations of two or more existing words (Jespersen, 1942, p. 134). This type of linguistic expression is highly common in our day-to-day experiences. For instance, a person who is glancing at the current headlines would likely come across *trade war* or *election debate*. Important aspects of life in the recent pandemic are also captured through compounds, such as *lockdown* and *essential worker*. More generally, compounds constitute a core part of human language. In an analysis of English words that emerged during the last century, it was shown that 30-40% of the recorded new words are compounds (Algeo & Algeo, 1991, p. 14). The process of creating compounds, also known as compounding, is productive in many languages around the world (Bauer, 2011; Schlücker, 2019), including pidgins (Plag, 2006) and constructed languages (e.g., Toki Pona; Lang, 2014).

One of the key motivations for the prevalence of compounding is that it enables the lexicon to express emerging concepts and ideas. If relatively new entities are seen as extensions of existing categories, compounding provides a simple solution for creating labels that reflect the original categories: for instance, cars that are primarily powered

by an electric motor can be simply expressed as *electric car*. If the new concept needs to be expressed in a non-literal way, compounding also provides a solution for creating distinct, non-literal expressions, as in *blue-collar* and *white-collar*. However, although compounding offers a large number of possible labels, the lexicon is a finite inventory that assigns one or few specific labels to each emerging concept. This implies that over time, very specific combinations of existing words are selected as conventional labels, despite the existence of many alternative ways to create compounds as well as non-compound labels, such as individual existing words (e.g., *mouse* as a pointing device) or derived words (e.g., *reactor* as a noun version of *react*).

In cognitive science, it has often been argued that language is shaped by pressures from communication and learning (e.g., Bates & MacWhinney, 1989; Christiansen & Chater, 2008; Gibson et al., 2019; K. Smith, 2022). Recently, a large body of work has formulated and tested this hypothesis in computational terms, showing that these pressures may explain various aspects of syntax and the lexicon, including word and morpheme order (Hahn, Degen, & Futrell, 2021), word length (Piantadosi, Tily, & Gibson, 2011), semantic categories (Regier, Kemp, & Kay, 2015), and word meaning extension (Ramiro, Srinivasan, Malt, & Xu, 2018). In this thesis, I build on these computational studies together with their underlying theoretical perspective to examine why the lexicon uses certain attested compounds as the conventional labels of historically emerging concepts as opposed to other possible expressions.

1.2 Motivation for the study

While much of the extensive literature on compounds has focused on the processing and representation of compounds by humans or machines (e.g., Lieber & Štekauer, 2011b; Nakov, 2013), the creation of compounds and the concepts that they express have primarily been examined in linguistic analyses of attested compounds. The most common

approach so far is to propose constraints on the intended relation between compound constituents, which limit how words can be combined to express a concept (e.g., Levi, 1978; Jackendoff, 2010; Lieber, 2016; Levin, Glass, & Jurafsky, 2019). For example, Levin et al. (2019) propose that the intended relations in compound names for artifacts tend to highlight events of their creation or use; this predicts that speakers would prefer the term *chocolate cake* to the alternate term *dark cake* as the label for the category of cakes that are made of chocolate. At a theoretical level, the existence of constraints has been associated with a pressure for informative communication, since constraints on how constituents may relate to each other help listeners to narrow down the potential ambiguity of new compounds (Downing, 1977; Levi, 1978).

However, these traditional accounts have focused on compound labels with a constituent that identifies the intended concept in a literal way (e.g., *chocolate cake* is a type of cake), and there are few accounts on the creation of non-literal compound labels (e.g., *seahorse*, *white-collar*). Despite the fact that their intended concepts are not explicitly identified by one of the constituents, non-literal compound labels often appear to be semantically motivated (Bauer, 2010; Benczes, 2015). For instance, the constituents of *seahorse* reflect the organism's shape and habitat, even though they do not explicitly indicate that the compound refers to a type of fish. On the other hand, a separate line of work suggests that non-literal compound labels created via analogy are constrained by their analogues (e.g., Klégr & Čermák, 2010; Mattiello & Dressler, 2018). Because they have only examined compounds with clear analogues, it is unclear the extent to which analogy-based accounts apply to non-literal compound labels in general.

In addition, there is no comprehensive account that applies to both compounding and word reuse, a common strategy for lexicalizing concepts that involves the recycling of existing words (Ramiro et al., 2018). Traditional accounts of compounding treat it in isolation from other strategies for expressing emerging concepts, as evident by their reliance on the relations between compound constituents. Nonetheless, there exist two

kinds of concept-based approaches that jointly consider multiple lexicalization strategies. The first approach analyzes concepts as sequences of semantic categories and considers how the categories may be expressed by words or morphemes (e.g., Štekauer, 2005; Antoniová, 2020); this has been applied to compounding and word class conversion (e.g., *walk* as a noun) but not to more general types of reuse (e.g., *antenna* as a transducer). The second approach examines the semantic associations that underlie how an individual concept is lexicalized across languages (e.g., Blank, 2003; Norcliffe & Majid, 2024), but so far this approach has only been applied to specific concepts or semantic domains.

1.3 Main contributions

This thesis presents a large-scale, computational analysis of compound expressions from the perspective that language is shaped by pressures from communication and learning. In cognitive science, previous work has explained various aspects of lexical items by using information-theoretic approaches (e.g., Mollica et al., 2021; Hahn, Mathew, & Degen, 2022) and computational modelling (e.g., Arndt-Lappe, 2011; Ramiro et al., 2018). However, these existing computational approaches have not been able to explain why the lexicon uses specific combinations of morphemes to label certain lexicalized concepts as opposed to other expressions. My work establishes a new connection between compound expressions and computational analyses of language in cognitive science by demonstrating how existing information-theoretic and modelling-based approaches can be extended to account for attested compound expressions in the lexicon.

From a theoretical point of view, my work offers an investigation of two issues that remain underexplored in the extensive literature on compounds. First, it is unclear how compound labels compete with the products of word reuse, which is one of the most common strategies for expressing emerging concepts besides compounding (Algeo & Algeo, 1991; Ramiro et al., 2018). Second, there are few accounts on the creation of non-literal

compound labels, even though the creation of this type of compound is productive in many languages around the world (Scalise, Fábregas, & Forza, 2009; Bauer, 2010). I show that the competition among compound labels and reuse items can be in part explained by a drive for efficient communication, and that non-literal compound labels can be predicted by modelling the cognitive processes involved in compound interpretation.

1.4 Thesis structure

This thesis is structured as follows. In Chapter 2, I first briefly discuss the definition of compounds and provide a comprehensive outline on existing accounts of compounding, which have primarily been based on linguistic analyses. The rest of this chapter reviews the computational approaches that I build on in the later chapters.

Chapter 3 presents a unified account of compounding and word reuse by building on the idea that language is shaped to support efficient communication. I propose that compounds and reuse items reflect opposing pressures for informativeness and ease of use, respectively, such that attested expressions achieve an efficient tradeoff between the two pressures. I formulate the proposal in computational terms, and I use it to analyze reuse items and compounds that emerged in English, French, and Finnish during the last century. I show that attested labels are more efficient than hypothetical labels, which suggests that these expressions are shaped by pressures from communication. Nonetheless, while this account applies to both literal and non-literal compound labels, literal compound labels tend to be more efficient than non-literal labels on average.

Chapter 4 further considers the competition among compounds and reuse items. Although the previous chapter suggests this competition can be explained by an efficient tradeoff between opposing communicative pressures, it remains unclear when a lexicalization strategy might be preferred over another. I hypothesize that communicative and cognitive economy might predict the choice between word reuse and compounding for

expressing an emerging concept. I test this hypothesis by developing a computational analysis of English word meanings that emerged over the last century, and I find that strategy choice may be explained partly by a pressure for least effort.

Chapter 5 returns to the creation of non-literal compound labels and proposes that considering analogy in addition to composition provides a better account of these labels. I formulate a computational model that integrates these cognitive processes of compound interpretation to predict the compound label of a given concept, and I show that the model more accurately reconstructs both literal and non-literal compound labels that are attested in English, Chinese, and German than simpler baseline models.

Finally, Chapter 6 summarizes the findings of the present thesis and how they contribute to research on compounding. I also discuss how the approaches developed in this thesis may be extended to future computational analyses of compound expressions and other aspects of lexical evolution.

Chapter 2

Background

2.1 Defining compounds

There is an extensive literature on compounds spanning theoretical, psychological, and computational research, but to consider its insights and limitations, we must first make sure we are covering the same linguistic construction. In fact, a well established view is that creating a standard, universally accepted definition of compounds remains an open question. For example, after reviewing the literature on compounds in different languages, Lieber and Štekauer (2011a) conclude that "there are (almost) no reliable criteria for distinguishing compounds from phrases or from other sorts of derived words". Similarly, in their discussion of compounds and multiword expressions (MWEs) in European languages, Finkbeiner and Schlücker (2019) suggest that "a cross-linguistically valid definition of compounds and the demarcation from MWEs may be impossible". Bauer (2017) points out that despite the superficial simplicity of compounding as a word formation process, the literature offers "no overall agreement on such basic issues as the definition of a compound".

Although there is no standard definition of compounds, it is still possible to outline the typical conditions invoked in existing definitions. Besides the most basic criterion that

a compound consists of two constituent words, conditions in compound definitions may be categorized as phonological, morphological, and syntactic (Lieber & Štekauer, 2011a; Finkbeiner & Schlücker, 2019). First, the most typical phonological condition is that stress is placed in a distinctive position. English compounds are usually left-stressed; for example, the compound *gréen card*, which refers to a permit, has a different stress pattern from the phrase *green cárd*, which refers to a card that is green (Finkbeiner & Schlücker, 2019). Second, morphological inflection is carried by a single constituent. English noun compounds are pluralized by modifying the second constituent; for example, the plural form of *fire truck* modifies *truck* instead of *fire*. Last, the constituents are syntactically inseparable, which means that neither constituent is allowed to interact with other words in a sentence independently of the compound. With respect to the compound *blackbird*, one cannot modify individual constituents as in *this is a very blackbird* or *this is a black ugly bird* (Lieber & Štekauer, 2011a).

These typical conditions are generally applicable to English and other Germanic languages (Schlücker, 2019; Booij, 2019), but they often fail to include word combinations that are traditionally seen as compounds in other languages (see Appendix, Section A.1 for examples). There are two opposing approaches to resolving the lack of a standard set of conditions that consistently apply across languages. The first approach is to define compounds on "a language-specific basis, since languages vary with respect to the realization of their morphological features and the use of morphologically-proper units" (Ralli, 2013). Under this approach, the set of compounds across languages is the union of forms that are seen as compounds in specific languages (e.g., including all forms in Appendix, Section A.1). The opposite approach is to create a single crosslinguistic definition by using "shared-core" conditions; the point is to "single out the great majority of constructions that have been called 'compound' " (Haspelmath, 2025). Here, only a subset of the union of all language-specific compounds are seen as crosslinguistic compounds, but all forms in the subset satisfy the same core conditions (e.g., syntactic inseparability).

In this thesis, I define compounds by formulating a version of the language-specific approach. Specifically, a compound in a certain language is a combination of two existing open-class words, which satisfies additional language-specific conditions. Since it is relatively inclusive, this definition allows us to build on most existing accounts of compounds. The details on language-specific conditions will be specified in later chapters and the appendix, depending on the languages involved in data analysis.

2.2 Linguistic analyses of compounds

In this section, I review existing accounts of compounding which have primarily taken the form of linguistic analyses of attested compounds. Before the review, Section 2.2.1 discusses basic terms recurrent in existing studies of compounds which will also be referenced in the later chapters. By dividing these accounts based on their main approaches, the rest of the chapter reviews traditional accounts in Section 2.2.2, concept-based accounts in Section 2.2.3, and analogy-based accounts in Section 2.2.4. The discussion focuses on their limitations with respect to the creation of non-literal compound labels and the competition among compounds and reuse items.

2.2.1 Basic terminology

2.2.1.1 Endocentric and exocentric compounds

The distinction between endocentric and exocentric compounds is recurrent in existing accounts of compounding (e.g., Marchand, 1969; Bauer, 2017). Based on the classic definition by Bloomfield (1933), a compound is endocentric if it is a hyponym of one of its constituents, and otherwise it is exocentric. For example, *cellphone* is endocentric because it belongs to the category of phones, and *electric car* is endocentric because it belongs to the category of cars. On the other hand, *blue-collar* and *white-collar* are exocentric since they denote membership in certain occupations and neither constituent

is a hypernym; *seahorse* is also exocentric for similar reasons.

The meaning of exocentric compounds can be derived from figurative (or non-literal) interpretations of compound constituents, which may be achieved through at least two strategies (Bauer, 2010, 2017; Benczes, 2015; Jackendoff, 2016). In the first strategy, the meaning is derived by interpreting one of the constituents in a non-literal way and modifying it with the other constituent. For example, the term *phone neck* refers to "pain in the neck caused by using the phone" (Bauer, 2010, p. 174), in which *neck* is metonymically interpreted as neck pain and modified by *phone* to specify the cause of pain. In the second strategy, the meaning is derived by interpreting the compound in a literal way and extending its literal interpretation. For example, the term *birdbrain* refers to "person whose brain is similar to that of a bird" (Jackendoff, 2016, p. 35), in which the literal brain of a bird is extended to denote a person with a similar brain.

In the following, I will equate endocentric compounds with literal compound labels and equate exocentric compounds with non-literal compound labels. This is because endocentric compounds contain a hypernym that identifies the intended concept in a literal way (e.g., a cellphone is a phone), whereas exocentric compounds must involve some form of non-literal interpretation.[1] As we will see, most existing accounts of compounding have focused on endocentric compounds, and there are few accounts on creating exocentric compounds to label concepts and ideas.

2.2.1.2 Head constituents

Intuitively, the head of a compound is the constituent that is considered more important than the other constituent, and there are different ways to determine the head (e.g., Zwicky, 1985; Scalise & Fábregas, 2010). In his discussion of compound heads, Bauer (2017) notes that there are two types of head that are common in the literature. The

[1]Endocentric compounds may also involve figurative interpretations (e.g., *ghost town*), but here "non-literal compounds" is exclusively used as an informal label for exocentric compounds.

Structure	Compound	Language	Gloss	Translation
$[A + A]_N$	*dà-xiǎo*	Chinese	big-small	size
$[V + V]_N$	*bagnasciuga*	Italian	soak-dry	shoreline
	subibaja	Spanish	climb-descend	see-saw
	yapboz	Turkish	construct-destroy	jigsaw puzzle
	cái-feng	Chinese	cut-sew	tailor
$[A + V]_N$	*guǎng-gào*	Chinese	wide-announce	advertisement

Table 2.1: Exocentric compounds that do not have a grammatical head, adapted from Scalise et al. (2009, p. 54).

first type is known as semantic head, which is a constituent that is a hypernym of the compound. The second type is known as grammatical head, which is a constituent that shares its word class with the compound. The grammatical head is almost always the same as the semantic head but the reverse is not true (Scalise & Fábregas, 2010). For example, *cake* in *chocolate cake* is both the semantic head and the grammatical head, whereas *horse* in *seahorse* is the grammatical head but not the semantic head.

It is straightforward to determine the head for endocentric compounds, since every endocentric compound contains a semantic head by definition, but it is often difficult to determine the head of an exocentric compound. While the notion of semantic heads is not directly applicable because exocentric compounds are not hyponyms of their constituents, the notion of grammatical heads only applies to a strict subset of exocentric compounds. It is common to find compounds with no semantic head or grammatical head: Scalise et al. (2009) finds that noun compounds that are combinations of A + A, V + V, or A + V are common in several languages that are not Germanic; some examples of these compounds are shown in Table 2.1. These compounds are often seen as headless, which means that no constituent can be regarded as more important than the others.

To discuss the literature on compounding in the current chapter, I will follow the typical notion of heads (Bloomfield, 1933; Bauer, 2017). I will also follow standard terminology and refer to the non-head constituent as the modifier. In Appendix, Section A.2, I develop an extended notion of semantic heads for exocentric compounds, and in later chapters, I will refer to the constituents of exocentric compounds that correspond

to this extended notion as non-literal heads (e.g., *neck* in *phone neck*).

2.2.2 Traditional accounts

Although existing accounts of compounding treat both the syntactic and the semantic aspects of compounds (e.g., see Štekauer & Lieber, 2005), I focus on the semantic aspects as the rest of the thesis examines the role of compounds in expressing concepts. Traditional semantic analyses of compounds aim to describe constraints on the intended head-modifier relations in attested compounds (e.g., Levi, 1978; Jackendoff, 2010). Following ten Hacken (2016), I classify these analyses into two main approaches, which consist of the relation-based approach and the constituent-based approach.

The relation-based approach aims to describe the relations that are typically encoded in attested compounds and distinguish them from atypical relations (e.g., Levi, 1978; Warren, 1978; Jackendoff, 2010; Schlücker, 2016). Levi (1978) provides a classic account of English endocentric compounds that is representative of this approach. Perhaps the most well-known aspect of her account applies to noun-noun and adjective-noun combinations, for which she argues that head-modifier relations typically belong to one of nine meta-relations known as Recoverably Deletable Predicates (RDPs); Table 2.2 shows the proposed RDPs along with examples. Her account further predicts that non-RDP relations are explicitly encoded in the compound. For example, the concept of "fish-eating dinosaurs" should not be expressed by *fish dinosaur* because the intended relation between the constituents is not an RDP. In more recent work, Jackendoff (2010) integrates the general idea of RDPs into a generative account of head-modifier relations. This account is able to produce fine-grained relations by composing information encoded in the constituents with basic functions (or predicates). Crucially, the list of basic functions is finite and it serves to constrain the space of possible head-modifier relations.

The constituent-based approach aims to derive constraints on head-modifier relations from the meanings of compound constituents (e.g., Allen, 1978; Lieber, 2016; Levin et

RDP label	Alternative label(s)	Example
CAUSE	causative	*tear gas*
HAVE	possessive, dative	*gunboat*
MAKE	productive, constitutive, compositional	*snowball*
USE	instrumental	*steam iron*
BE	essive, appositional	*soldier ant*
IN	locative (spatial or temporal)	*field mouse*
FOR	purposive, benefactive	*arms budget*
FROM	source, ablative	*solar energy*
ABOUT	topic	*tax law*

Table 2.2: RDPs and examples of noun compounds, adapted from Levi (1978, p.76-77).

al., 2019). This approach underlies most syntactic and semantic analyses of synthetic compounds, which have a constituent that is derived from a verb (Bauer, 2017, p. 79). It has been shown that the head-modifier relations encoded in these compounds are often constrained by the verb base (e.g., Lieber, 1983; Spencer, 1991; Bauer, 2017) as well as interactions between the semantic features of other elements in the compound (Lieber, 2016). For instance, while *taxi* is the direct object of *drive* in *taxi driver*, *city* is the subject of *employ* in *city employee* since the referent of *-ee* is non-volitional and serves as the object of the verb. This approach has also been applied to compounds beyond cases that involve a verb base. By using a large-scale analysis of attested compounds in English, Levin et al. (2019) show that noun compounds denoting artifacts tend to invoke head-modifier relations that highlight the events of their creation or use, whereas noun compounds denoting natural kinds tend to invoke relations that highlight their essence (e.g., perceptual properties).[2] For instance, this predicts that because the head of *coconut cake* is an artifact, the compound is more suitable for expressing cakes made of coconuts as opposed to cakes that superficially look like coconuts.

Because the proposed constraints place limits on the mappings between compound forms and compound meanings, these accounts are relevant to explaining why the lexicon expresses emerging concepts with specific attested compounds as opposed to other

[2]Similar observations on compound labels for artifacts and natural kinds have also been made in earlier studies (Downing, 1977; Wisniewski & Love, 1998).

possible combinations. However, a major limitation is that these accounts typically focus on endocentric compounds and treat exocentric compounds as exceptions (e.g., Levi, 1978; Allen, 1978). In particular, the proposed constraints on head-modifier relations are based on literal interpretations of compound constituents, and it is unclear if they are applicable to exocentric compounds in which the constituents are related to the intended meaning in a non-literal way (see Section 2.1.1.1). Moreover, because they focus on head-modifier relations in compounds, these accounts cannot explain why the lexicon expresses a concept via an attested compound as opposed to other types of expressions (e.g., derived words or reuse items).

2.2.3 Concept-based accounts

An alternative approach to compounding examines the patterns among how speakers or languages find a form to express a particular concept. Unlike traditional accounts of compounding, most concept-based accounts examine how concepts are expressed by compounds alongside other types of expressions. These accounts may be generally categorized into one of two approaches, which consist of the model-based approach and the typological approach.

The model-based approach is grounded in a theoretical model of word formation that assigns word forms to target concepts (Štekauer, 2005, 2016). The model involves two sets of processes; one set decomposes the concept into a sequence of semantic categories, and the other set assigns forms to these categories (see Table 2.3 for examples). One major aspect of this approach is the existence of constraints on the category sequence, which limits how concepts can be mapped to word forms. Antoniová (2020) provides the most comprehensive study of these constraints in the context of compounding, showing that only 87 out of 2,704 possible sequences are attested in English noun-noun compounds. Nonetheless, because Štekauer (2016) treats exocentric compounds as the output of individual semantic categories, it is unclear how the constraints on category sequences can

Compound	Sequence of categories	Sequence of morphemes
workplace	(Agent) - Action - Location	*0* - *work* - *place*
dragnet	(Agent) - Action - Object	*0* - *drag* - *net*
weekend	Temporal - (State) - Temporal	*week* - *0* - *end*
wedding party	Process - (State) - Process	wedding - *0* - *party*

Table 2.3: Examples of noun compounds decomposed into semantic categories, adapted from Antoniová (2020, p. 329). In the second column, each item corresponds to a semantic category, and each category maps to a morpheme in the third column (0 indicates the empty string).

explain the internal structure of attested exocentric compounds. Another aspect of the model-based approach is that it applies to other forms of word formation, since a category may be mapped to morphemes or the empty string in addition to words. However, it has only been applied to account for conversion (e.g., *walk* as a noun) but not to other types of reuse items (e.g., *mouse* as a pointing device).

The second concept-based approach falls within the literature on lexical typology, which aims to uncover and explain frequent patterns through which concepts are lexicalized across languages (Talmy, 1985; Blank, 2003). Typological accounts that consider both compound forms and their meanings have primarily focused on the semantic associations between the compound and its constituents (e.g., Tagliavini, 1949; Koch, 2008; Urban, 2012; Tjuka & List, 2024). For example, Blank (2003) discusses expressions for "pupil" from more than 100 languages and the associations that they encode (e.g., "ball" and "pupil"). He argues that certain associations are much more frequent than others because speakers consider them to be cognitively salient. Similarly, Norcliffe and Majid (2024) examine complex perception verbs, such as the Kaluli term *mun dabuma* (lit. odour-hear) for smell, from a balanced sample of 100 languages. The authors find that certain concepts (e.g., "hear" and "smell") are more frequently associated by the constituents of these verbs than others. Since they consider multiple lexicalization strategies jointly, these accounts may be relevant to explaining the competition among compounds and other types of expressions. However, the scope of these typological studies is typically limited to particular concepts or semantic domains.

Compound	Literal translation	Meaning
scharrel-kip	scratch-chicken	'free range chicken'
scharrel-vlees	scratch-meat	'free range meat'
scharrel-ei	scratch-egg	'free range egg'
scharrel-melk	scratch-milk	'free range milk'

Table 2.4: Subfamily of Dutch compounds with an ecological meaning, adapted from Booij (2010).

2.2.4 Analogy-based accounts

In linguistics, the cognitive process of analogy has long been invoked to account for morphologically complex forms and their meanings (de Saussure, 1916; Bauer, 2001; Arndt-Lappe, 2015). Although it has been peripheral in traditional accounts of compounding, analogy is being emphasized in a growing number of studies on the creation and interpretation of compounds (e.g., Ryder, 1994; Booij, 2010; Mattiello & Dressler, 2018, 2022).

Following Mattiello and Dressler (2018), an analogical compound is a new compound that 1) shares a constituent with one or more existing compounds such that the shared constituent is in the same position across these compounds and 2) places a new word in the non-shared position. This notion encompasses the products of two processes that are often differentiated in the literature. The first process is known as local analogy, through which a new compound is created by modifying a single existing word: for instance, *software* was created by using *hardware* as a model and modifying the first constituent.[3] The second process forms a new compound by modifying a productive pattern formed by multiple compounds: for instance, Dutch compounds in Table 2.4 formed the pattern *scharrel* + *X*, which was responsible for giving rise to the new compound *scharrel-wijn* (lit. scratch-wine; 'eco-wine') (Booij, 2010).

Similar to the traditional accounts, semantic analyses of analogical compounds have suggested constraints on the mappings between the forms of these compounds and their

[3]See entry for *software* in the OED, where the etymology section states "soft adj. + ware n.3, in senses 1a and 2 [the computing sense] after hardware n." (Oxford University Press, 2025).

intended concepts. First, the shared constituent among compounds from the same productive pattern tends to develop a bound meaning that shifts away from its original meaning, and the bound meaning is inherited by analogical compounds based on the pattern (Klégr & Čermák, 2010; Booij, 2010; Booij & Hüning, 2014). For instance, the pattern formed by *blue-collar* and *white-collar* gave rise to the new term *green-collar*.[4] Here, the pattern *X* + *collar* developed the meaning of signifying membership in some profession, which is distinctive from its original neck-related meanings. Moreover, the semantic relation between the constituents of an analogical compound tends to resemble the relations that exist in its analogues (Mattiello & Dressler, 2018). In the example involving *X* + *collar*, the first constituent is always related to the second by providing a salient indicator of the intended profession (e.g., office workers used to be associated with wearing white-collar shirts).

As shown by these examples of analogical compounds, analogy-based accounts of compounding are able to explain the creation of certain exocentric compounds. However, a major limitation with these accounts is that they have only examined compounds with clear analogues (i.e., productive patterns or single-word models), and it is unclear if the analogy-based constraints may apply to a wider range of attested compounds. Another limitation is that compounding can create combinations that share no constituent with any existing compound and thus do not involve analogy, whereas traditional accounts may generalize to these cases if they are endocentric.

2.3 Computational analyses of lexical items

In this section, I review existing computational analyses of lexical items in cognitive science, focusing on the approaches that I build on in the later chapters. Section 2.3.1 reviews a body of computational work suggesting that the lexicon is shaped to support

[4]This term signifies membership in the green industry.

efficient communication. Section 2.3.2 reviews studies that use computational modelling to understand the mechanisms that may underlie the emergence of new word meanings or new word forms.

2.3.1 Efficiency-based accounts

Recently, a growing line of work examines linguistic structures in information-theoretic terms and suggests that these structures are shaped to support efficient communication (e.g., Regier et al., 2015; Kemp, Xu, & Regier, 2018; Gibson et al., 2019; Futrell & Hahn, 2022). Efficiency-based accounts have been shown to explain various aspects of language, but here I review existing accounts that focus on the lexicon or morphology. These accounts may be further classified depending on if they explain the structure of word meanings or the structure of word forms.

2.3.1.1 Structure of word meanings

One class of efficiency-based accounts examines how semantic categories carve up a semantic space (e.g., colour) and aims to explain why languages have the semantic categories they do (e.g., Regier et al., 2015; Kemp et al., 2018). The core proposal is that an attested category system should be informative by supporting accurate transfer of information from speakers to listeners and simultaneously simple to minimize the amount of cognitive resources used for representing it. While an informative system requires a large amount of resources to specify intended messages in full detail, a simple system leaves many aspects of the messages underspecified. Therefore, informativeness and simplicity must trade off against each other, and it is argued that attested systems are shaped by the need for efficient communication if they achieve near-optimal tradeoffs.

Although ideas related to informativeness and (cognitive) simplicity have also been discussed in earlier work (e.g., Rosch, 1978), efficiency-based accounts of semantic categories formulate both notions in information-theoretic terms. In the classic formula-

tion (Kemp & Regier, 2012; Regier et al., 2015), information loss (or inverse of informativeness) is measured by the surprisal of an intended message and how often it is expressed, and simplicity is defined as the size of the description for a system given some description language (e.g., a set of rules). This formulation defines a two-dimensional space of possible category systems in which attested systems can be compared against both baseline and theoretically optimal systems. One of the main advantages of this approach is that it is highly general, and it has been applied to explain attested category systems in various semantic domains, including kinship (Kemp & Regier, 2012), colour (Regier et al., 2015), containers (Y. Xu, Regier, & Malt, 2016), seasons (Kemp, Gaby, & Regier, 2019), and numerals (Y. Xu, Liu, & Regier, 2020).

Zaslavsky, Kemp, Regier, and Tishby (2018) propose an alternative formulation of informativeness and simplicity which is grounded in the information bottleneck (IB) principle (Tishby, Pereira, & Bialek, 2000). Relative to the classic formulation, the IB formulation is more closely aligned with Shannon's original communication model (Shannon, 1948) and requires explicit models of the speaker's strategy for encoding messages and the listener's strategy for decoding utterances. The other major point of divergence is that under the IB formulation, the measure of complexity (or inverse of simplicity) is always the mutual information between messages and utterances. For the semantic domain of colour, Zaslavsky et al. (2018) show that optimal systems generated by the IB formulation are more similar to attested systems than optimal systems generated by the classic formulation. Subsequent work has applied the IB formulation to account for the meanings of other elements of the lexicon, including grammatical markers (Mollica et al., 2021) and spatial demonstratives (Chen, Futrell, & Mahowald, 2023).

From a methodological point of view, this body of work demonstrates how the efficiency of word meanings may be rigorously formulated by using ideas from information theory and computationally evaluated on large-scale, crosslinguistic datasets. However, since their primary goal is to examine attested systems of semantic categories, these

accounts have generally treated category labels as abstract symbols and defined communicative efficiency in isolation from the structure of word forms. The account by Mollica et al. (2021) is an exception to this characterization as it connects the efficiency of word meanings to the length of grammatical markers.

2.3.1.2 Structure of word forms

Word length is one of the most well-studied properties of word forms that have been related to communicative efficiency. In particular, an established statistical law states that word length and word frequency are inversely correlated, which is argued to be the result of balancing between opposing pressures in communication (Zipf, 1949; Bentz & Ferrer-i Cancho, 2016; Kanwal, Smith, Culbertson, & Kirby, 2017). Assuming that longer words take more effort to produce, the speaker effort in word production is lower on average if meanings that need to be frequently expressed are labelled by short words; on the other hand, ambiguity is minimized if every meaning corresponds to a distinct word, which necessitates the use of long words because there is a limited number of ways to combine phonemes into words of a certain length. By using ideas from data compression, it can be shown that a lexicon that balances between these pressures in an optimal way makes word length a non-increasing function of word frequency (Pimentel, Nikkarinen, Mahowald, Cotterell, & Blasi, 2021; Mollica et al., 2021; Ferrer-i Cancho, Bentz, & Seguin, 2022). Nonetheless, Pimentel et al. (2021) show that the lexicons of seven languages do not optimize word length as much as possible largely due to phonotactic and morphological constraints on word forms.

Frequency-based studies of word length have been expanded by considering other factors in communication. Piantadosi et al. (2011) provide an account of word length by taking into account linguistic context, predicting that the information content of a word in context negatively correlates with its length. Across eleven European languages, the authors show that this correlation is overall stronger than the correlation between

frequency and length. However, Meylan and Griffiths (2021) show that comparisons between frequency and information content as predictors of word length are sensitive to design choices on text preprocessing. In a separate account, Mahowald, Dautriche, Gibson, and Piantadosi (2018) propose that to facilitate production and comprehension, frequent words should be more phonologically probable and similar to other words. By analyzing word forms from 96 languages, the authors find support for their proposal even after controlling for word length.

Morpheme ordering is another aspect of word forms that has been shown to reflect efficiency in communication. Building on earlier work in language processing, Hahn et al. (2021) propose that the ordering of linguistic elements in a sequence reflects a tradeoff between the pressure to minimize the memory capacity required to represent preceding elements and the opposite pressure to minimize the surprisal of the next element, which captures the amount of effort that listeners require to process it. The authors define both dimensions in information-theoretic terms, so that in a two-dimensional space where the axes correspond to memory and surprisal, the efficiency of an ordering may be determined by the area under the tradeoff curve. Hahn et al. (2022) apply this approach to morpheme orderings in six languages with rich agglutinative morphology, finding that attested orderings are more efficient than most possible orderings and that crucial properties of attested morpheme orderings emerge in hypothetical orderings that are created by optimizing the memory-surprisal tradeoff.

2.3.2 Modelling-based accounts

Another general approach for analyzing lexical items focuses on the emergence of new items and casts this linguistic process as a prediction problem. Here, I selectively review two bodies of work that have taken this approach and used computational modelling to understand the mechanisms that underlie the emergence of new meanings in existing words and the emergence of new words.

2.3.2.1 Word meaning extension

One recent line of computational work examines word meaning extension, which is the process through which existing words obtain new meanings (e.g., Y. Xu et al., 2016; Ramiro et al., 2018; Sun, Zemel, & Xu, 2021). These studies treat each existing word as a category of meanings, and whether a new meaning attaches to a particular word becomes a categorization problem. Building on accounts of categorization from cognitive psychology (e.g., Rosch, 1975; Nosofsky, 1986), this line of work treats different categorization models as distinct mechanisms of extension and examines the extent to which these mechanisms can account for attested extensions. For example, Ramiro et al. (2018) propose a series of computational models that predict the order in which the attested senses of a word emerged over time. These models predict that an existing word gains new senses that are similar to a privileged existing sense, similar to all existing senses on average, or maximally similar to any of the existing senses. By evaluating these models on $5,000$ English words and their attested sense extensions that emerged during the past millennium, the authors find that the last mechanism provides the best account for the historical emergence of sense extensions.

More recent studies have combined categorization models with machine learning methods to acquire a similarity space that may improve the accuracy of the models. Sun et al. (2021) examine the extension of the conventional senses of a word to slang senses. To predict attested extensions, the authors primarily apply a nearest neighbour model and a prototype model that is based on the central tendency of a category (Reed, 1972), which are combined with neural encoders that capture the semantic similarity between slang and conventional senses through contrastive learning (Le-Khac, Healy, & Smeaton, 2020). Yu and Xu (2023) examine how existing words extend to new senses in linguistic contexts. The authors explore a similar prototype model and a variant of the generalized context model (Nosofsky, 1986), and they propose a learning algorithm that transforms the similarity space so that it reflects the relations between new and existing

senses in different types of extensions (e.g., metonymy and metaphor).

In sum, these studies provide a computational account of how emerging concepts are lexicalized by reusing existing words since they suggest that attested extensions can be predicted by categorization models from cognitive psychology. However, this line of work has not directly addressed lexicalization strategies beyond reuse. In particular, some studies have used this approach to examine the selectional restrictions of English verbs and adjectives over time (Yu & Xu, 2021; Grewal & Xu, 2021; Yu & Xu, 2025), which are arguably relevant to an account of V + N combinations (e.g., *pickpocket*) and A + N combinations (e.g., *solar energy*). However, the same approach has not been applied to an analysis of attested combinations in the English lexicon.

2.3.2.2 Word formation

In theoretical accounts of morphology, the role of analogy in creating new word forms has been subject to extensive debate (Bauer, 2001; Arndt-Lappe, 2015). In particular, proponents of rule-based accounts have argued that the general notion of analogy is too permissive in generating new forms to distinguish attested forms from unattested forms.[5] One of the responses to this argument was to formulate analogy-based accounts in computational terms. Given some linguistic element as input (e.g., a root), analogy-based accounts that are computational treat the emergence of specific modifications to the input (e.g., structural change in past tense formation) as a classification problem. To quantitatively compare against rule-based predictions, these accounts make use of computational models that predict specific modifications (or classes) to an input based on its similarity with comparable items in the lexicon.

There are three models that have been widely applied in analogy-based studies of morphology. The first model is Skousen's analogical model of language (Skousen, 1989;

[5]Bauer (2001, p.76) states that the principle underlying the general notion of analogy is that "any new form can be created as long as there is a suitable pattern for it to be formed on."

Daelemans, Gillis, & Durieux, 2013). This model uses an algorithm to select a set of exemplars based on the features of the input, such that the selected exemplars are similar to the input and have frequent combinations of features; moreover, selected exemplars similar to each other also tend to have the same class. To make predictions, the model selects the most frequent class or a random sample based on class frequencies in the set. Next, the second model is the Tilburg memory based learner (Daelemans, Zavrel, van der Sloot, & van den Bosch, 2004). The core of this model is a standard k-nearest neighbour model that is combined with algorithms for feature weighting. On the other hand, the third model is the generalized context model (Nosofsky, 1986). Unlike the first two models, this model assigns a score to a class by using the similarity between the input and all exemplars associated with that class instead of more specific selections of exemplars (e.g., the nearest neighbour).

Analogy-based accounts of morphology have applied computational models to study inflection in a considerable number of studies, but their application to word formation has been relatively limited (e.g., see discussion by Arndt-Lappe, 2015). With regards to derivational morphology, these models have been used to examine nominalization in English and Spanish (Eddington, 2006; Arndt-Lappe, 2014; Plag, Kawaletz, Arndt-Lappe, & Lieber, 2023; Hofmann, Weissweiler, Mortensen, Schütze, & Pierrehumbert, 2025), diminutive formation in Dutch and Spanish (Daelemans, Berck, & Gillis, 1997; Eddington, 2002), English negative prefixes (Chapman & Skousen, 2005), English comparatives (Elzinga, 2006), and English adjectival suffixes (Arndt-Lappe, 2023). With regards to compounding, analogy-based models have been applied to predict the use of linking elements in Dutch and German compounds (Krott, Baayen, & Schreuder, 2001; Krott, Schreuder, Baayen, & Dressler, 2007) and the position of word stress in English noun-noun compounds (Arndt-Lappe, 2011). So far, analogy-based accounts of word formation have not applied computational models to examine the selection of the base in derived words or the head and modifier words in compounds.

2.4 Models of compound interpretation

Since I emphasize the role of the listener in the later chapters, in this section I review the literature on computational models of compound interpretation. Section 2.4.1 reviews approaches that represent the meaning of a compound via head-modifier relations similar to those reviewed in Section 2.2.2. Section 2.4.2 reviews approaches that represent the meaning of a compound through real-valued vectors in semantic space.

2.4.1 Predicting semantic relations

A body of computational work assumes that the meaning of a compound is represented by the relation between its constituents and aims to predict the intended relation of a target compound form. The relations are usually obtained from a pre-defined list (e.g., Table 2.2), but they may also take the form of a paraphrase which explicates the connection between the two constituents and shares the same head with the target compound (e.g., Nakov & Hearst, 2006; Hendrickx et al., 2013). This body of work aims to create models that correctly assigns the intended head-modifier relation to a compound.

In natural language processing (NLP), earlier approaches for compound interpretation have relied on classic machine learning models and features derived from dictionaries or corpus statistics (e.g., Lauer, 1995; Kim & Baldwin, 2005; Girju, 2007; Ó Séaghdha & Copestake, 2009; Tratz & Hovy, 2010b). For example, Kim and Baldwin (2005) propose a nearest-neighbour model based on the taxonomic distance between a pair of heads or a pair of modifiers in WordNet (Fellbaum, 1998). Ó Séaghdha and Copestake (2009) adopt the distributional hypothesis and use the co-occurrence statistics of individual constituents and word pairs as features. They present kernel-based methods that are suitable for these features and use them to classify the target compound into one of six relations. Tratz and Hovy (2010b) implement a maximum entropy classifier with

a very large number of binary features derived from WordNet, Roget's Thesaurus, the substrings in a compound, and n-grams in which the constituents of a compound co-occur. Notably, the authors find that features based on words in WordNet definitions had the biggest overall impact on model accuracy.

More recent NLP models for relation-based compound interpretation have increasingly relied on neural networks (e.g., Dima & Hinrichs, 2015; Fares, Oepen, & Velldal, 2018; Ponkiya, Patel, Bhattacharyya, & Palshikar, 2018; Shwartz & Waterson, 2018). In particular, these studies often place an emphasis on learning continuous representations for target compounds and their potential paraphrases. For example, Ponkiya et al. (2018) apply long-short term memory networks (LSTMs; Hochreiter & Schmidhuber, 1997) to encode target compounds and potential paraphrases, respectively, so that paraphrases can be ranked based on their similarity with the target. Liu et al. (2022) propose a more complex method that learns the similarity between a target compound and a potential paraphrase by using contrastive learning (Le-Khac et al., 2020) and encoders based on contextualized language models. In a multitask setup, Shwartz and Dagan (2018) propose a model that encodes paraphrases that contain a missing head, modifier, or relation (e.g., juice [relation] apple) with a bidirectional LSTM and predicts different kinds of missing elements with multilayer perceptrons (MLPs).

Models for relation-based compound interpretation have also been proposed in other areas of research. Gagné and Shoben (1997) propose the Competition Among Relations In Nominals (CARIN) model, which scores how likely a head-modifier relation is the correct interpretation of a target compound based on the frequency of the relation in existing compounds that share a modifier with the target. Costello and Keane (2000) propose the C^3 model, which is a search algorithm that scans for the best paraphrase for a target compound based on scores derived from three pragmatic constraints. These constraints require that concepts intended by speakers are plausible to listeners and that new compounds are made to be informative with respect to the intended concepts

while avoiding redundancy.[6] Note that the C^3 model can also produce paraphrases that are headed by a non-constituent (e.g., "a fish with a horse-shaped head" for the target *seahorse*), but it is nonetheless included in the current section as a more general solution to problems tackled by the other models.

The models from this body of work generally inherit the limitations in traditional accounts of compounding that were discussed in Section 2.2.2. First, with the exception of the C^3 model, these models assume that the meaning of a compound is fully captured by its constituents and a head-modifier relation, but this assumption does not apply in the case of exocentric compounds. Second, the representation of word meanings via constituents and relations applies to compounds and certain similar constructions (e.g., genitive + N; Bauer & Tarasova, 2013) but cannot be readily applied to account for the meanings of other types of expressions, such as the extensions of morphologically simple words (e.g., *mouse* as a device).

2.4.2 Vector-based composition functions

Another body of computational work assumes that word meanings are represented by vectors in semantic space and that the meaning of a compound can be obtained from composing the vectors of its constituents. This body of work aims to create composition functions that can produce new vectors for novel word combinations such that the new vectors capture certain semantic properties of the combinations.

In cognitive science, this notion of composition has been investigated in research on prototype theory. While earlier work suggests that the mental representation of a known concept consists of a central exemplar known as the prototype (e.g., Rosch, 1975), E. E. Smith, Osherson, Rips, and Keane (1988) propose the selective modification

[6] Note that Costello and Keane (2000) use the label "informativeness" to refer to redundancy avoidance. In the present thesis, the same label refers to the degree of information transfer in communication, which is more closely related to the diagnosticity constraint in the C^3 model.

model to account for the prototypes of novel adjective-noun combinations. Their model represents the prototype of a concept by using a list of attributes (e.g., colour) and a list of values (e.g., red) for every attribute; each attribute has a degree of diagnosticity and each value has a degree of salience.[7] The prototype of the combination is obtained by using the degrees of diagnosticity and salience of the adjective prototype to modify the noun prototype. Beyond adjective-noun combinations, Hampton (1991) describes the creation of prototypes for novel coordinative compounds (e.g., *pet fish* is a fish that is also a pet), but he does not formulate this process in computational terms.

In NLP and computational linguistics, a large number of studies have combined composition functions with distributional semantic models (e.g., Guevara, 2010; Mitchell & Lapata, 2010; Reddy, McCarthy, & Manandhar, 2011; Lazaridou, Vecchi, & Baroni, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Mitchell and Lapata (2010) propose a general framework that summarizes two classes of common composition functions. The first class consists of additive composition functions, which apply a linear transformation to each of the input vectors and return the sum of the transformed vectors. The second class consists of multiplicative composition functions, which involve multiplying both input vectors with a single tensor. The authors show that vectors produced by various composition functions are able to capture human judgments of phrase similarity far better than simply representing a combination by its head word. In addition to inferring vector representations for novel combinations, these classes of functions have been widely used in psycholinguistics to derive measures for predicting human behavioural data on the processing of both existing and novel word combinations (e.g., Vecchi, Marelli, Zamparelli, & Baroni, 2017; Günther & Marelli, 2016; Marelli, Gagné, & Spalding, 2017; Hsieh, Marelli, & Rastle, 2025).

[7]Observe that this prototype representation can be restated in terms of vectors. As a result, the modification model may be seen as a composition function since it takes two prototypes as input and produces one prototype as output.

Later studies in NLP have proposed more complex composition functions based on neural networks (Socher, Huval, Manning, & Ng, 2012; Yazdani, Farahmand, & Henderson, 2015; Dima, 2015; Dima, de Kok, Witte, & Hinrichs, 2019). For example, Yazdani et al. (2015) propose composition functions based on MLPs or quadratic transformations of the input vectors. By applying them to model human ratings of compositionality, the authors compare these functions to additive composition functions, finding that the quadratic-based function performs best but is relatively close to both the additive and MLP-based functions in performance. Dima et al. (2019) propose the transformation weighting model, which transforms the input vectors with multiple non-linear functions and returns a weighted combination of the outputs of these functions. The authors compare the new model to seven composition functions from earlier work based on their accuracy in reconstructing phrase embeddings obtained from word2vec (Mikolov, Sutskever, et al., 2013), and they show that the new model achieves the best overall performance.

Relative to head-modifier relations which form the basis of the models in the last section, vector-based representations appear to be less interpretable and much more detached from traditional accounts of compounding (see Section 2.2.2). However, the combination of composition functions with distributional semantic models offers two major advantages. First, distributional semantic models provide representations of both compounds and non-compounds in semantic space, which is highly useful for a computational analysis of both compounds and reuse items. Second, distributional semantic models are able to create meaning representations by leveraging historical and multilingual text data, such that the representations reflect patterns of word usages in different time periods and languages (e.g., Y. Xu & Kemp, 2015; Hamilton, Leskovec, & Jurafsky, 2016; Yamada et al., 2020).

Chapter 3

Word reuse and combination support efficient communication

The contents of this chapter are based on the following publication:
Xu, A., Kemp, C., Frermann, L., & Xu, Y. (2024). Word reuse and combination support efficient communication of emerging concepts. Proceedings of the National Academy of Sciences, 121(46), e2406971121.

3.1 Introduction

The human lexicon is not static but evolves over time. A central role of the lexicon in this evolutionary process is to lexicalize novel ideas, therefore serving as an adaptive system which supports the symbolic encoding and communication of emerging concepts (Deacon, 1997; Pinker & Bloom, 1990). The most common strategies of lexicalization involve reusing and combining existing words in the lexicon (Marchand, 1969; Algeo, 1980; Brinton & Traugott, 2005; Ramiro et al., 2018), although other strategies such as borrowing (e.g., *tofu*) and coinage (e.g., *quark*) also exist. Reuse refers to using the form of an existing word to express something new. For instance, *mouse* was reused to describe a "small device that is moved by hand across a surface to control the movement of the cursor on a

computer screen" (Oxford University Press, 2023). Combination refers to concatenating two or more existing words to form a new word, typically known as a compound (e.g., *armchair* combines *arm* and *chair* to express a new type of chair).[1] Word reuse and combination are often viewed and studied separately as two distinct aspects of lexical evolution. Here, we present an information-theoretic framework that accounts for both strategies through the lens of efficient communication.

Word reuse has been traditionally discussed in the context of historical semantic change (e.g., Traugott & Dasher, 2001) and word meaning extension (e.g., Williams, 1976). This body of work aims to identify regularities in how words take on new meanings over time. More recent studies using large-scale historical and cross-linguistic data have suggested that words tend to take on new meanings that are semantically related to existing ones (Y. Xu et al., 2016; Ramiro et al., 2018), and the processes of word meaning extension reflect cognitively economic ways of expanding the referential range of existing words in the lexicon (Srinivasan & Rabagliati, 2015; Y. Xu, Duong, Malt, Jiang, & Srinivasan, 2020). However, this line of research focuses almost exclusively on word reuse (or extension) and treats it in isolation from word combination.

Word combination has been commonly studied in the literature on word formation and morphology (e.g., Štekauer & Lieber, 2005). In particular, existing accounts focusing on noun-noun compounds have suggested that compound interpretation involves selecting from a systematic list of predicate relations which in turn constrains possible noun combinations produced by speakers (e.g., Levi, 1978; Lieber, 1983; Levin et al., 2019). An alternative approach appeals to more functionally motivated principles arguing that compounds should be semantically transparent while shorter forms are preferred if compound constituents are redundant (Downing, 1977; Dressler, 2005; Costello & Keane, 2000). These principles have also been discussed in psycholinguistic work showing that

[1] In this chapter, the term "word combination" is used interchangeably with "compounding".

the semantic relatedness between a novel compound and its constituents predicts its acceptability, unless the relatedness is too high (Günther & Marelli, 2016). We believe these functional principles might be equally applicable to explaining word reuse. However, to our knowledge there is no unified account that characterizes both word reuse and combination (c.f. Blank, 2003; Grzega, 2011).

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

Here we extend existing efficiency-based accounts that assume speakers and listeners use the same static lexicon by considering communicative interactions during the spread of linguistic innovations (e.g., Labov, 2011; Milroy & Milroy, 1985). In particular, we consider the labels of novel concepts that are yet to be encoded in the lexicon of a large number of language users. We propose that when speakers communicate novel concepts to these language users, word reuse and combination reflect a fundamental tradeoff between speaker effort and information loss (or the inverse of informativeness): on the one hand, speakers can minimize their effort by reusing short words that underspecify intended concepts; on the other hand, information loss is minimized if speakers use relatively long word forms that combine existing words in an informative way. We hypothesize that as speakers repeat these communicative interactions, their encoding of novel concepts is

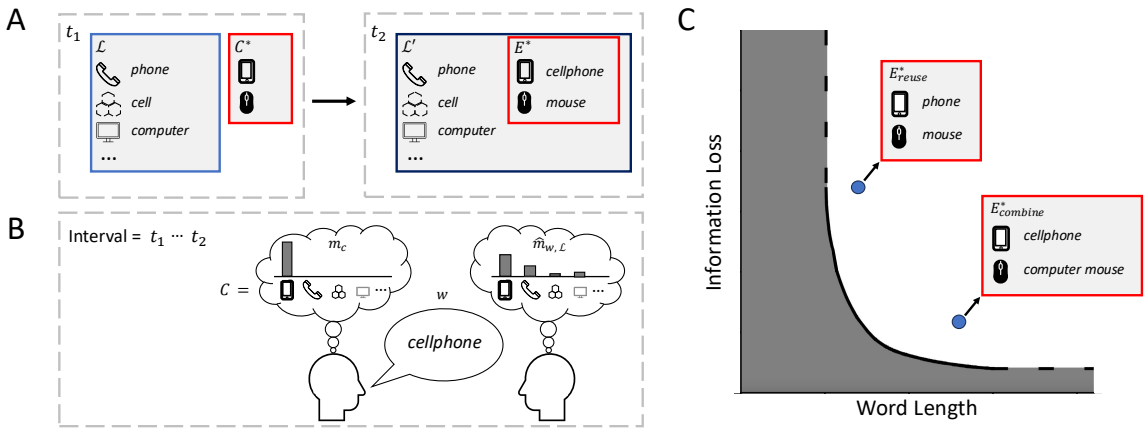

Figure 3.1: Illustration of our theoretical proposal. Panel (A) illustrates the lexicalization of emerging concepts using examples from English during the historical interval 1980-2000. The existing lexicon $\mathcal{L}$ and the set of emerging concepts $\mathcal{C}^*$ at time t_1 are illustrated on the left. At a later time t_2, the attested encoding of the novel concepts E^* enters the expanded lexicon $\mathcal{L}'$, which are shown on the right. Panel (B) illustrates the two opposing pressures in a communicative interaction taking place before t_2. Here, the speaker intends to convey the emerging concept "cellphone" to a listener whose lexicon does not yet have a word for expressing it, and grey bars illustrate probability distributions over a universe of concepts $\mathcal{C}$ that capture uncertainty regarding the intended concept. Our proposal focuses on the pressure for minimizing the length of the utterance, and the pressure for minimizing information loss, or the difference between the speaker and listener distributions over concepts. Panel (C) illustrates possible encodings of the novel concepts in Panel (A). Each point corresponds to the average length and information loss of an encoding of the novel concepts, and the shaded area corresponds to costs that are not attainable. We propose that word reuse and combination reflect a tradeoff between these two costs, and that both attested reuse items and attested compounds achieve tradeoffs that are relatively efficient. Here, the example encodings are simplified to contrast reuse and combination, and in reality an encoding can consist of both strategies.

shaped by the pressure to minimize speaker effort and the opposing pressure to minimize information loss, such that the word length and informativeness of both attested reuse items and attested compounds efficiently trade off against each other. We illustrate this idea in Figure 3.1.

Our proposal is consistent with other functional accounts of word reuse and combination. Previous accounts have separately suggested that combinations should be informative (Lieber, 2004; Clark & Berman, 1984; Downing, 1977) and reuse represents an economical strategy (Štekauer, 2005; Piantadosi, Tily, & Gibson, 2012). Computational studies of reuse items (Ramiro et al., 2018; Brochhagen, Boleda, Gualdoni, & Xu, 2023)

and compounds (Günther & Marelli, 2016; Vecchi et al., 2017; Pugacheva & Günther, 2024) have shown that semantic transparency is preferred across both strategies, which is consistent with a preference for using informative word forms to reduce communicative error. In recent work on language evolution, a pressure for informativeness is often considered necessary for compositional or subword structure to emerge in the lexicon (e.g., Kirby, Tamariz, Cornish, & Smith, 2015; Carr, Smith, Cornish, & Kirby, 2017). Crucially, prior studies have shown that a pressure for informativeness may be sufficient for the emergence of subword structure when speakers have to communicate novel meanings to a large community with limited shared history (Raviv, Meyer, & Lev-Ari, 2019a, 2019b). Here, we build on these existing studies to offer a unified functional account of word reuse and combination.

In the following, we first specify our theoretical proposal in formal terms. We then evaluate our efficiency-based proposal against a large historical dataset of reuse items and compounds attested in English, French, and Finnish over the past century. Last, we discuss the implications of our results and avenues for future work.

3.2 Computational formulation of theory

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

We illustrate the setting for this scenario in Figure 3.1A. Here, we consider an encoding of concepts as a set of form-concept pairs (or mappings), and we treat the lexicon as an encoding of lexicalized concepts which may incorporate novel pairs over time. Suppose

the existing lexicon $\mathcal{L}$ consists of form-concept pairs known by all speakers in the speech community at time t_1, and the set $\mathcal{C}^*$ contains novel concepts emerging at t_1 but not encoded in the existing lexicon. We assume that the eventual attested encoding of novel concepts, denoted by E^*, spreads among speakers until the expanded lexicon $\mathcal{L}' = \mathcal{L} \cup E^*$ has been acquired by all speakers at time t_2. In the current study, we assume that forms in E^* always reuse or combine forms that exist in $\mathcal{L}$.

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

3.2.1 Model of communication

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

We describe our model using an illustration of the interaction in Figure 3.1B. The speaker's mental representation is a speaker distribution m_c over a universe of concepts $\mathcal{C}$. In general m_c could capture speaker uncertainty, but we assume that m_c corresponds to a single intended concept c and picks it out with certainty. The intended concept c is drawn from a need distribution $p(c|\mathcal{L}')$ which captures the frequency with which different concepts are communicated (e.g., Kemp et al., 2018; Zaslavsky et al., 2018), and here we assume the speaker only communicates concepts encoded in her lexicon. To express c, the speaker selects a form w according to her production policy $p(w|c, \mathcal{L}')$ which captures the frequency of using specific forms in her lexicon to communicate an intended concept.

In turn, the listener uses w and his lexicon to deterministically construct his mental representation which is a listener distribution $\hat{m}_{w,\mathcal{L}}$ that aims to reconstruct m_c.

We define the listener distribution by using a variant of prototype-based categorization models (e.g., Rosch, 1975; Ramiro et al., 2018).[2] In our model, the listener treats each form w as the label of a category of concepts, which is represented by the category prototype q_w. The listener uses the category to construct a distribution, so that the probability of concept c is high if it is semantically similar to q_w. We specify this distribution via the similarity choice model (Luce, 1963; Nosofsky, 1986):

$$\hat{m}_{w,\mathcal{L}}(c) \propto \exp\left\{-\gamma d(c, q_w)\right\} \tag{3.1}$$

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

3.2.2 Communicative costs

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

[2] Although they are often used in related work (e.g., Zaslavsky et al., 2018; Mollica et al., 2021), Bayesian listeners are not applicable in the current context because the dyad does not share the same lexicon.

Labov, 2011; Milroy & Milroy, 1985; Del Tredici & Fernández, 2018).

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

$$\mathbb{E}[l(W)|\mathcal{L}'] = \sum_{c,w} p(c, w|\mathcal{L}')l(w) \tag{3.2}$$

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

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

$$\mathbb{E}[D(M||\hat{M})|\mathcal{L}', \mathcal{L}] = \sum_{c,w} p(c, w|\mathcal{L}')h(\hat{m}_{w,\mathcal{L}}(c)) \tag{3.3}$$

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

3.2.3 Efficiency of attested encodings

Our proposal can be specified by considering how much the attested encoding E^* contributes to these communicative costs. This contribution can be summarized by a single

objective function obtained from combining and simplifying Equations 3.2 and 3.3:

$$L_\beta[E^*|\mathcal{L}] = \mathbb{E}[D(M||\hat{M})|\mathcal{L}', \mathcal{L}] + \beta\mathbb{E}[l(W)|\mathcal{L}'] \tag{3.4}$$

$$\propto \sum_{(c,w)\in E^*} p(c, w|\mathcal{L}') \cdot (h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w)) \tag{3.5}$$

where $\beta \geq 0$ is a tradeoff parameter. As in previous efficiency-based approaches (Zaslavsky et al., 2018; Mollica et al., 2021), optimizing Equation 3.5 for every β produces a Pareto frontier that specifies the space of possible encodings of $\mathcal{C}^*$, which we illustrate in Figure 3.1C. We hypothesize that the attested encoding is shaped by competing pressures for minimizing speaker effort and information loss, which predicts that it will be relatively close to the Pareto frontier in this space of possible encodings.

In Appendix, Section B.1, we provide a more detailed derivation of Equations 3.2-3.5. We also provide a discussion on the limitations of previous efficiency-based accounts of the lexicon in terms of accounting for attested combinations of words or morphemes.

3.3 Results

To test our proposal, we instantiated our scenario using Princeton WordNet (Fellbaum, 1998) and its multilingual extensions (Bond & Foster, 2013). These WordNets are conceptually organized dictionaries that map language-specific, orthographic forms to a common set of word senses or lexicalized concepts. We focused on English, French, and Finnish because they have the largest WordNets among alphabetical languages in terms of the size of their sense inventory (Bond & Foster, 2013). For each language and one of five consecutive intervals over the past century, we instantiated emerging concepts as WordNet senses, and we instantiated their attested encoding and the existing lexicon as sets of form-sense pairs. We identified whether an English form-sense pair is an emerging reuse item or compound by using its first citation in the Historical Thesaurus of En-

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

Table 3.1: Examples of reuse items (R) and compounds (C) that emerged in the past century; sense definitions have been truncated for brevity

glish (Kay, Roberts, Samuels, & Wotherspoon, 2017), and we inferred whether a pair is existing based on its estimated frequency in historical text (Davies, 2002; Michel et al., 2011). We implemented the same components for French and Finnish by relabelling emerging and existing senses in the English data with a language-specific form that was attested in historical French or Finnish text (Michel et al., 2011; National Library of Finland, 2014). For tractability, we ignored linking constituents in compounds and we used compounds that have exactly two constituents. More details on data processing are provided in Section 3.5.

To show that our approach applies to both reuse and combination, we controlled for differences in sample size between strategies by instantiating each attested encoding such that it only contains reuse items or compounds. Across our target intervals, we analyzed 518 reuse items and $2,828$ compounds in English, 529 reuse items and 409 compounds in French, and 510 reuse items and 645 compounds in Finnish; sample sizes for specific intervals are provided in Appendix, Section B.2. In Table 3.1, we show examples of reuse items and compounds that make up these encodings.

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

Table 3.2: Samples of near-synonyms created for attested labels

Given the existing lexicon, we computed the average length and information loss incurred by a speaker communicating with an encoding of novel concepts as specified in Equation 3.5. We used the orthographic length of word forms in our subsequent analyses as a proxy for production effort. To estimate information loss, we implemented the listener distribution in Equation 3.1 in three parts. We first represented each concept or word sense by embedding the text of its WordNet definition (see Table 3.1 for examples) with a sentence encoder (Reimers & Gurevych, 2019), and we followed approaches that construct the prototype as an average of the existing senses of a word or its constituents (Reed, 1972; Mitchell & Lapata, 2008). For the sensitivity parameter, we set $\gamma = 10$ based on the informativeness of existing words. These costs were computed for each form-sense pair in the encoding, and then averaged according to their need and production probabilities estimated from historical form-sense frequencies. We specify our implementation in Section 3.5.

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

3.3.1 Average-case efficiency

In our first analysis, we compared attested items to optimal encodings on the Pareto frontier and two sets of baseline encodings. The first baseline consists of alternate encodings created from replacing the label of each item in an attested encoding with a near-synonym; examples of near-synonyms are shown in Table 3.2. To probe the space

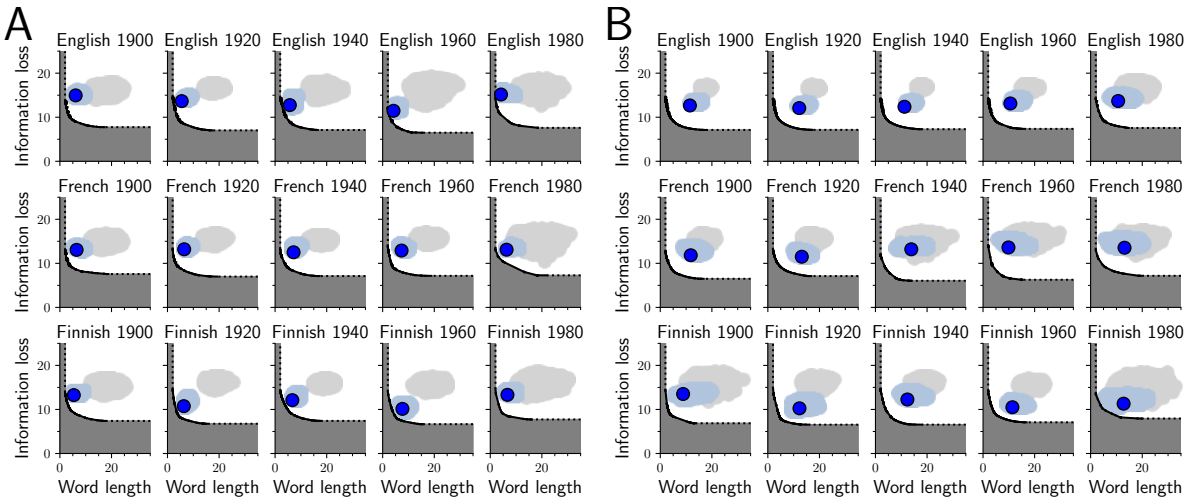

Figure 3.2: Illustration comparing (A) attested reuse items and (B) attested compounds to the constructed baselines and the Pareto frontier. Every point corresponds to an encoding of emerging concepts for a specific language and interval. Attested cases are marked in blue, near-synonym baselines in light blue, and random baselines in grey. Black solid lines in the bottom left show the estimated Pareto frontier, and the shaded areas show costs that are not attainable.

of all possible alternatives, we created a second baseline by replacing each attested label with a string uniformly sampled from labels in the existing lexicon and their combinations. Details on estimating the Pareto frontier and creation of baseline encodings are specified in Section 3.5.

Figure 3.2 summarizes the comparisons between attested and alternative encodings. Each Pareto frontier shows the optimal tradeoff that can be achieved by any encoding of the emerging concepts, and intuitively, the closeness of an encoding to the frontier approximates its efficiency. Across strategies, intervals and languages, we observe that both attested encodings (blue) and near-synonym baselines (light blue) tend to be closer to the frontier and more efficient than random baselines (grey), and attested encodings tend to be more efficient than both baselines. By construction, attested and near-synonym encodings tend to be shorter than random baselines because it is more likely to sample a combination of long words than a single short word. The fact that attested labels also dominate near-synonyms indicates the relative efficiency of attested labels does not arise solely due to chance and the prevalence of longer word forms.

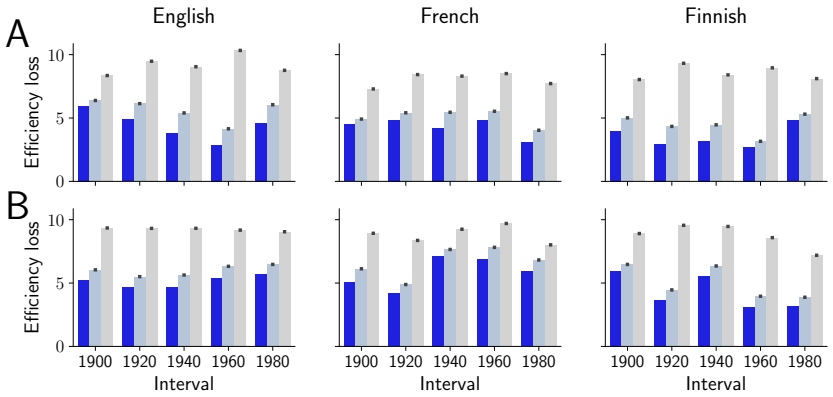

Figure 3.3: Efficiency loss of attested encodings for (A) reuse items and (B) compounds relative to the average loss of baselines. Attested loss is marked in blue, and the average loss of near-synonym and random baselines is marked in light blue and grey, respectively. Error bars show bootstrapped 95% confidence intervals.

Figure 3.3 compares attested encodings against baselines using a quantitative measure of efficiency loss (see Section 3.5), which overall confirms our qualitative observations. In Appendix, Section B.4, we show that attested reuse items and compounds remain more efficient than the constructed baselines under different implementations of our scenario of lexical evolution. First, we show our results are robust if we use a uniform distribution over attested items and different values for the sensitivity parameter. Second, we show the results hold up across different communication channels via an implementation that represents word forms using phonemes instead of letters. Third, we describe an analysis based on historical embeddings (Mikolov, Chen, Corrado, & Dean, 2013; Hamilton et al., 2016) to address the concern that our approach may be biased by using contemporary embeddings to study historical change. Last, we show these findings are robust to alternative datasets of lexicalized concepts by considering another English dictionary and by assuming one-to-one correspondence between concept and form.

3.3.2 Item-level variation

In Figure 3.2, we observe non-trivial gaps between Pareto frontiers and attested encodings. This suggests that the communicative efficiency of attested encodings could be

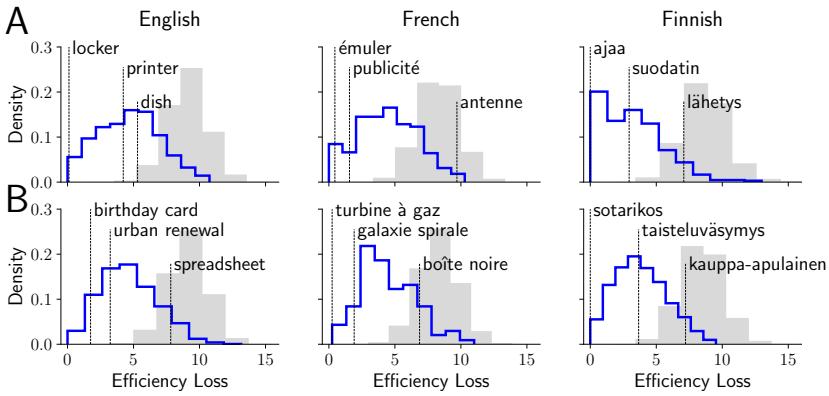

Figure 3.4: Efficiency loss of individual attested items for (A) reuse and (B) compounding and randomly sampled labels. The distributions for attested and random are marked in blue and grey, respectively. Examples in Table 3.1 are annotated.

improved by replacing some attested reuse items and compounds with more efficient forms. We thus compared individual items to optimized forms by using the same implementation of our scenario, except we replaced full encodings with singletons that contain individual items.

Figure 3.4 shows efficiency losses for individual items, which measure their deviation from optimized forms, and each distribution is aggregated over all time intervals. As in the previous analysis, the item-level loss is approximated by the distance between attested items and Pareto frontiers, which is illustrated in Figure 3.5. We observe that attested items tend to be much closer to optimized forms (loss = 0) than most randomly sampled labels (grey), but nonetheless the right tails of attested items overlap with random distributions. In Appendix, Section B.5, we show that item-level loss based on orthographic forms strongly correlates with item-level loss based on phonemic forms. The variation from near-optimal to near-random among attested reuse items and compounds reveals that some items are more strongly shaped by our proposed tradeoff than others.

Here we characterize this variation using two well-studied subclasses of lexical items. Endocentric compounds are the most well-studied subclass of compounds that are defined by the relation between intended and existing constituent concepts (e.g., Downing, 1977; Jackendoff, 2010). The head word of an endocentric compound encodes a superordinate

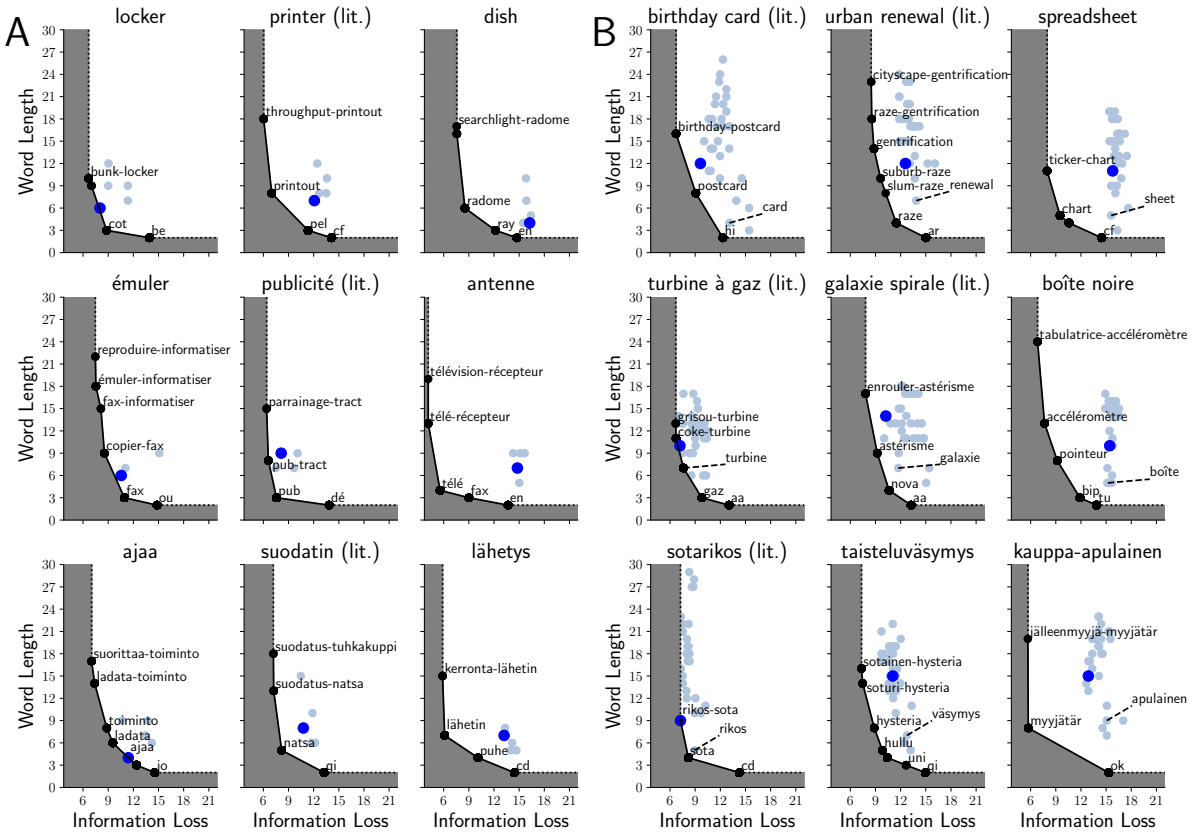

Figure 3.5: Item-level illustration for (A) attested reuse items and (B) attested compounds. Headers correspond to the examples in Table 3.1, with additional marking for literal items (lit.). Each dark blue dot corresponds to an attested form. Black dots correspond to the item-level Pareto frontier, and light blue dots correspond to the near-synonym set generated for this item; the size of markers for attested items is larger than the size of other markers for improved visibility. A sample of optimal labels and compound head words are shown as text. Note that the axes are swapped relative to Figure 3.2 and the x-axis is truncated so there is more space to display optimal labels.

category of the intended concept and is a literal expression of this concept (e.g., *birthday card* is a *card*), in contrast to non-literal (or exocentric) compounds (e.g., *blue-collar*). Similarly, a subclass of reuse items often expresses an intended concept that is more narrow than an existing sense of the reused word (e.g., Bloomfield, 1933), for instance the modern use of *car* as a motorized vehicle refers to a narrower set of concepts compared to its original meaning of *wheeled cart*. Across strategies, these literal expressions may be more efficient because they are more transparent to the listener than non-literal ones. For example, in Figure 3.5, we observe that literal reuse items and endocentric compounds

tend to be more efficient than their non-literal counterparts (e.g., *birthday card* vs *dish* or *antenne*).

To test this hypothesis, we leveraged the WordNet taxonomic hierarchy to classify reuse items and compounds into literal and non-literal cases. We performed a quantitative analysis that compares the efficiency loss of literal and non-literal items; we supplemented attested reuse items with additional data by using the head words of attested compounds, since the original English WordNet does not explicitly encode literal items (Miller, 1998) (see Section 3.5 for details). We find that in French and Finnish, literal reuse items are significantly more efficient than non-literal items ($t(527) = 4.70$, $p < .001$; $t(508) = 6.37$, $p < .001$), and the same trend holds between literal and non-literal head words in English ($t(2793) = 23.60$, $p < .001$), French ($t(396) = 7.32$, $p < .001$), and Finnish ($t(631) = 9.40$, $p < .001$); endocentric compounds also tend to be more efficient on average in English ($t(2826) = 17.26$, $p < .001$), French ($t(407) = 3.68$, $p < .001$), and Finnish ($t(643) = 7.58$, $p < .001$). We illustrate these comparisons in Appendix, Section B.5. These results suggest our efficient tradeoff proposal applies more strongly to labels that encode novel concepts in a literal way across both strategies.

In Appendix, Section B.5, we explore the variation in efficiency among attested reuse items and compounds in two further analyses. In the first analysis, we used taxonomic distance measures (Z. Wu & Palmer, 1994; Leacock, Chodorow, & Miller, 1998) as a continuous version of the literal and non-literal distinction. In line with our findings above, we found that taxonomic distance measures are positively correlated with efficiency loss across all languages and across both strategies. In the second analysis, we investigated whether frequent items are closer to Pareto frontiers since this implies less total efficiency loss. We did not find that frequency differentiates variation in item-level loss. This may be due to frequency effects in lexicalization beyond the scope of our account. We return to other factors that underlie lexical evolution in Section 3.4.

We demonstrate our findings with examples in Figure 3.5. Along each Pareto frontier,

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

3.3.3 Strategy comparison

In Figure 3.2, we also observe that reuse-based encodings and combination-based encodings tend to occupy different neighbourhoods in the space of possible encodings. To compare differences between the two strategies, we compared informativeness and word length between all attested reuse items and compounds, aggregated across intervals for each language. We show the statistics in Appendix, Section B.5, finding that on average, attested reuse items tend to be shorter than attested compounds across all three languages, and attested compounds tend to be more informative than attested reuse items in English and French. This mirrors existing proposals that cast reuse as an economical lexicalization strategy (Štekauer, 2005; Piantadosi et al., 2012) and compounding as an informative strategy (Downing, 1977; Clark & Berman, 1984).

3.4 Discussion

We have presented evidence that word reuse and combination are shaped by competing pressures of informativeness and length minimization that affect the lexicalization of emerging concepts. We formulated this view in information-theoretic terms, and we tested our proposal using large-scale resources over history and across languages. We found that both attested reuse items and attested compounds that emerged over the

past century in English, French, and Finnish are more efficient than random and near-synonym baselines, and that literal items are generally more efficient than non-literal items across both strategies.

Our work makes several contributions to efficiency-based accounts of language. First, our work establishes a new connection between efficiency-based accounts and word formation. Previous accounts have focused on form length and meanings (e.g., Mollica et al., 2021) and morpheme ordering (Hahn et al., 2022), but not the specific combination of certain morphemes as opposed to others in compositional words. In particular, one account suggests that the presence of subword structures in word forms hinders efficient length minimization (Pimentel et al., 2021). We show that in the case of compounding, these structures may nonetheless support efficient communication in settings where the speaker and listener do not fully share the same lexicon. Second, another line of investigation uses ideas from information theory to show that existing form-meaning mappings support accurate reconstruction of intended meanings under cognitive constraints (e.g., Regier et al., 2015; Zaslavsky et al., 2018). Our work shows that information-theoretic formulations of efficiency can also account for generalizations to novel concepts not yet encoded in the lexicon. Last, previous work has also examined the tradeoff between simplicity and informativeness in the lexicon, but only within restricted domains (Kemp & Regier, 2012; Y. Xu et al., 2016; Zaslavsky et al., 2018; Y. Xu, Liu, & Regier, 2020; Denić, Steinert-Threlkeld, & Szymanik, 2020; Steinert-Threlkeld, 2021; Zaslavsky, Maldonado, & Culbertson, 2021; Chen et al., 2023). Our computational framework offers a promising venue for extending simplicity-informativeness analyses toward the broader lexicon beyond individual semantic domains.

Although we focused on the lexicalization strategies of reuse and compounding, the functional principles invoked by our theory are general and our framework can be applied to other types of word formation. For instance, derived words are also highly productive (Algeo, 1980) and previous work has suggested that they are subject to a

preference for informativeness and brevity (Lieber, 2004; Marelli & Baroni, 2015). Since derived words are combinations of existing words and morphemes, one way to extend our approach might be to incorporate a representation of affix meanings and model their combination with stem words (Marelli & Baroni, 2015; Westbury & Hollis, 2019). Derivational morphemes tend to be shorter than free morphemes and might yield more efficient strategies for expressing emerging or novel concepts; for example, the derived word *physicist* is more compact than the compound *physics scientist*. As a result, derivation may sit at a midpoint between economy of production and informativeness, flanked by the strategies examined in the current study along the Pareto frontier of efficiency.

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

Bybee, 1998; Pagel, Atkinson, & Meade, 2007; Lieberman, Michel, Jackson, Tang, & Nowak, 2007; Hamilton et al., 2016). Frequent lexical innovations may outcompete alternative, more efficient innovations that express the same concepts but emerged later and were used less frequently by speakers.

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

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

it with computational models of compound interpretation (e.g., Mitchell & Lapata, 2008; Tratz & Hovy, 2010a; Nakov, 2013) to further differentiate genuinely inefficient labels (e.g., homonyms and opaque compounds) from other non-literal cases.

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

3.5 Materials and methods

Treatment of data

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

We first instantiated $\mathcal{L}$ and E^* using the English WordNet (Fellbaum, 1998) for each interval. Before setting up the components, we standardized word forms in the dataset to facilitate estimation of frequency and word length. We assigned form-sense pairs to $\mathcal{L}$ if their frequencies during $[t_1 - 20, t_2]$ exceeded certain thresholds, based on token frequencies from the Google Ngrams corpus (English 2020 version; Michel et al., 2011) and sense frequencies estimated using the Corpus of Historical American English (COHA; Davies, 2002) and a state-of-the-art word sense disambiguation algorithm, EWISER (Bevilacqua

& Navigli, 2020). We obtained E^* by first collecting reuse items and compounds with first citations in $[t_1, t_2]$ according to the Historical Thesaurus of English (Kay et al., 2017), and we then took a coarse-grained approach that assumed these items emerged at t_1 and have entered the lexicon by t_2. Last, we processed both $\mathcal{L}$ and E^* to ensure they are disjoint. We provide a full description of this data processing pipeline in Appendix, Section B.2.

We instantiated the same components for each interval using French and Finnish WordNets (Sagot & Fišer, 2008; Lindén & Carlson, 2010). Here, we assumed concepts encoded in the English lexicon $\mathcal{L}$ are also encoded in the French or Finnish lexicon $\mathcal{L}$ for the same interval, and we implemented $\mathcal{L}$ by labeling these concepts with forms that are attested in historical sections of the Google Ngrams corpus (French 2020 version; Michel et al., 2011) and the Newspaper and Periodical Corpus of the National Library of Finland (FNC; National Library of Finland, 2014). We also assumed the set of emerging concepts $\mathcal{C}^*$ is the same as the English set for the same interval, and we implemented E^* by pairing each $c \in \mathcal{C}^*$ with one of its French or Finnish forms if the form is in $\mathcal{L}$ or combines forms in $\mathcal{L}$. We provide a full description of this data processing pipeline in Appendix, Section B.2.

Need and production distributions

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

In the case of English, we estimated the first term using token frequencies in historical texts that appeared during $[t_1, t_2]$ in the English Google Ngrams corpus (Michel et al., 2011). We defined the first term as $p(w|\mathcal{L}') \propto f_w$, where f_w is the frequency of the form w. We estimated the second term in two steps. If w is a unigram, we reused sense frequencies based on text that appeared during $[t_1, t_2]$ in COHA (Davies, 2002), so that given sense c

with frequency $f_{c,w}$, we have $p(c|w, \mathcal{L}') \propto f_{c,w}$. Otherwise, we used a uniform distribution over concepts in $\mathcal{L}'$ that are labeled by w since our sense disambiguation method did not apply to open compounds. Last, we applied add-one smoothing to form-concept pairs in $\mathcal{L}'$.

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

Prototype model

Here, we specify our variant of prototype-based categorization models in Equation 3.1. We first represented each $c \in \mathcal{C}$ by embedding its definition using a state-of-the-art sentence encoder, Sentence-BERT (Reimers & Gurevych, 2019), yielding a semantic vector for each concept c. Although WordNet definitions were compiled during the past century (Fellbaum, 1998), the encoder was trained on contemporary natural language data, and we view these vectors and the semantic space as an approximation of listener representation of concepts in our target intervals. Throughout this study, we used cosine distance for $d(\cdot, \cdot)$ following Reimers and Gurevych (2019).

Since the word w is either an existing word in the listener lexicon $\mathcal{L}$ or a combination of existing words, we defined the prototype $q_{w,\mathcal{L}}$ in two parts. If w is in $\mathcal{L}$, we extended the model in Reed (1972) and defined the prototype as a weighted average of category exemplars, i.e., the concepts encoded by w in $\mathcal{L}$; alternatively, if w is a combi-

nation of N constituents, we used the additive composition function (Mitchell & Lapata, 2008) to recursively define a composite prototype that combines the prototypes of its constituents (e.g., E. E. Smith et al., 1988):

$$q_{w,\mathcal{L}} = \begin{cases} \sum_c p(c|w,\mathcal{L})c & \text{if } w \in \mathcal{L} \\ \sum_i q_{w_i,\mathcal{L}} & \text{else if } w = w_1 ... w_{\mathrm{N}} \in \mathcal{L}^{\mathrm{N}} \end{cases} \tag{3.6}$$

Here the expression $w \in \mathcal{L}$ implies $(c, w) \in \mathcal{L}$ for at least one $c \in \mathcal{C}$, and $p(c|w, \mathcal{L})$ was estimated from the relative frequencies of items in $\mathcal{L}$. In Appendix, Section B.3, we validated our embedding space and construction of prototypes using datasets of English word similarity and compound meaning predictability.

Intuitively, the parameter γ simulates how much the listener prefers to infer the most transparent interpretation of a word. In Appendix, Section B.3, we show that the average information loss incurred over communicating existing concepts (i.e., part of the omitted term in Equation 3.5) is minimized when $\gamma \in [15, 20]$. This suggests a reasonable range for γ is $(0, 15]$ since if γ is too low then the listener does not distinguish among concepts, and if γ is too high then the listener will incur high information loss whenever the word form expresses an extended sense. For this reason, in the main text we set $\gamma = 10$. We note that our argument is based on the information loss of existing words, and we leave more fine-grained modeling of the listener distribution for compounds for future work.

Estimating the Pareto frontier

For each tradeoff parameter $\beta = 0, 0.01, ..., 10$, we computed the optimal encoding E^*_β that encodes concepts in $\mathcal{C}^*$ and minimizes Equation 3.5. We assume that need probabilities are constant with respect to how concepts are encoded. In this case, each novel concept independently contributes to the overall cost in Equation 3.5, and this optimization is equivalent to finding a form w for each $c \in \mathcal{C}^*$ such that w jointly minimizes word length

and surprisal for a certain β. That is, we want to optimize the following item-level objective over existing words and possible combinations:

$$L_\beta[w|c, \mathcal{L}] = h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w) \tag{3.7}$$

For tractability, we greedily selected a first string $u \in \mathcal{L}$ that optimizes Equation 3.7, and we greedily selected a second string $u' \in \mathcal{L}$ or the empty string such that the concatenation of the selected strings minimizes Equation 3.7. The final concatenation is the approximately optimal form for c and this form-concept pair is added to E^*_β.

Baseline encodings

We constructed near-synonym and random encodings as baselines for every attested encoding E^*. To construct a near-synonym encoding, we started by constructing a near-synonym set for every form-sense pair in E^*. Suppose that the form contains modifier w and syntactic head u; we assumed all English and Finnish compounds are right-headed and all French compounds are left-headed due to their relative prevalence (Lieber, 2011; Hyvärinen, 2019; Van Goethem & Amiot, 2019). For the constituent w, we selected the top $k = 5$ forms, among all existing forms $x \in \mathcal{L}$ that are closest to w in terms of the cosine distance between $q_{w,\mathcal{L}}$ and $q_{x,\mathcal{L}}$ but are not antonyms of w in WordNet. We repeated this procedure for u, except we also made sure the possible word classes of the generated constituents overlap with the possible word classes of u. The near-synonym set is defined as $\{xy : x \in S_u, y \in S_w\} \cup S_u$, where S_w and S_u are the forms generated for w and u, respectively. We then created a near-synonym encoding by replacing the form in each attested form-sense pair with a random sample from its near-synonym set. We constructed random encodings for each E^* similarly by replacing every attested label with a form uniformly sampled from forms in $\mathcal{L}$ and their combinations. For every E^*, we created $100,000$ near-synonym and random encodings, respectively.

Efficiency loss

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

$$\epsilon = \min_{\beta} \Big(L_\beta[E^*|\mathcal{L}] - L_\beta[E^*_\beta|\mathcal{L}] \Big) \tag{3.8}$$

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

Literal and non-literal items

We classified a reuse item (c, w) as a literal item if the novel concept c is a hyponym of an existing sense of w, and otherwise we classified the pair as a non-literal item. We classified a compound item (c, w) as an endocentric compound if c and the head word constitute a literal item, and otherwise we classified the item as an exocentric compound; we made the same assumptions on head positions as in our construction of near-synonyms. Since Princeton WordNet was made to avoid linking a sense and its hyponyms to the same word (Miller, 1998), there are few literal reuse items for English (N = 3) and we supplemented attested reuse items for all languages with additional data by replacing the compound in each attested compound-sense pair with its head word. A small number of endocentric compounds with head positions different from our assumption were not used in data augmentation. We provide hypernyms of literal items from Table 3.1 in Appendix, Section B.5.

Data availability

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

Chapter 4

Predicting strategy choice in lexicalization

The contents of this chapter are based on the following publication:
Xu, A., Kemp, C., Frermann, L., & Xu, Y. (2023). Predicting strategy choice in word formation: A case study of reuse and compounding. In Proceedings of the 45th Annual Meeting of the Cognitive Science Society.

4.1 Introduction

In this chapter, we continue to examine the competition among reuse items and compounds. The previous chapter suggests that this competition reflects a tradeoff between informativeness and word length, and that efficient labels are preferred for lexicalizing emerging concepts over inefficient labels. However, this account cannot explain why certain concepts are lexicalized by compounds but not by equally efficient existing words, even if they differ in terms of informativeness and length (e.g., *gas turbine* and *turbine*). Here, we ask what factors may predict the strategy for expressing an emerging concept, and we address this question with a computational analysis of historical cases of reuse and compounding in English.

We begin our investigation by adopting a functionalist perspective which suggests that language reflects communicative and cognitive efficiency (e.g., Zipf, 1949; Ramiro et al., 2018). This view has received empirical support from previous studies on historically attested cases of word meaning extension (Y. Xu, Malt, & Srinivasan, 2017; Ramiro et al., 2018), conventionalized complex words (A. Xu, Kemp, Frermann, & Xu, 2022), and loan words (Monaghan & Roberts, 2019), but these different lexicalization strategies have typically been examined in isolation. We extend these previous studies by suggesting that communicative and cognitive efficiency might also predict the strategy choice for expressing an emerging meaning, and in this initial work we focus on the strategy choice between reuse and compounding.[1]

One earlier 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 et al., 2022; Mollica et al., 2021; Pimentel et al., 2021). This least effort account has also been supported by experiments on artificial language learning (Kanwal et al., 2017) and repeated reference games (Krauss & Weinheimer, 1964; Hawkins et al., 2023). 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, the lengths of their labels should be constrained by the corresponding communicative need, so that the lexicon remains relatively compact in a Zipfian sense. In particular, since the plausible compounds for expressing a given mean-

[1]In general, the lexicalization of an emerging concept may involve multiple strategies. For example, *car* and *automobile* are synonymous; the former is a case of reuse because it used to refer to horse-drawn carriages whereas the latter is a complex word. However, in the present study we focus on concepts that are lexicalized by a single strategy.

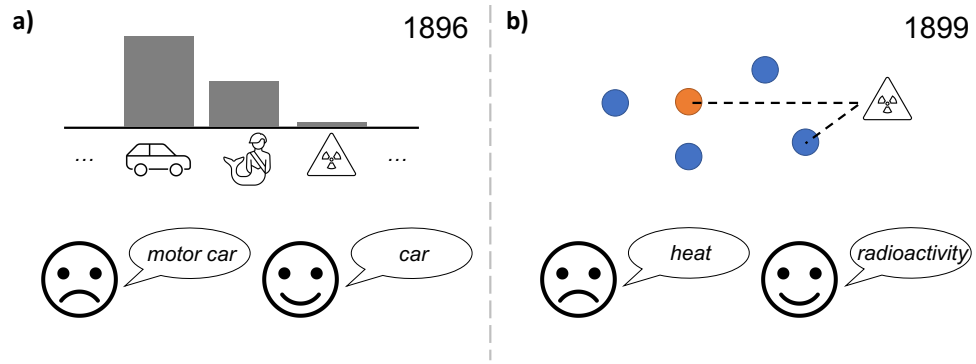

Figure 4.1: Illustration of our a) least effort hypothesis and b) novelty hypothesis. In panel a), grey bars represent the distribution of communicative need for different concepts, and the speaker prefers to reuse the form "car" for a meaning with high communicative need. In panel b), each dot or symbol corresponds to a meaning in similarity space; red and blue dots correspond to existing meanings and the meanings of possible compounds, respectively. Here the speaker prefers a compound form for a meaning that is far from the meanings expressed by existing words. In both panels, the upper right corner shows the year of first citation of the emerging concept according to the Oxford English Dictionary (OED).

ing 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 psychology and linguistics (Rosch, 1975; Lakoff, 1987; Geeraerts, 1997). Using computational models and large-scale historical data, this line of work suggests that the lexicon tends to express novel meanings with existing words that are semantically related (Ramiro et al., 2018; Brochhagen et al., 2023), which is argued to reflect ease of learning (Srinivasan, Al-Mughairy, Foushee, & Barner, 2017; Floyd & Goldberg, 2021). However, certain emerging concepts (e.g., "radioactivity" or "nuclear winter") may be highly novel relative to the concepts expressed 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 there is a large number of possible compounds and some are likely to be more semantically related to novel meanings than existing words.

We illustrate our hypotheses in Figure 4.1. 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 prefer compounding to reuse. In the following, we first introduce our datasets and operationalization of need and novelty. We then evaluate the least effort and novelty hypotheses separately, and we conduct a predictive analysis to test whether the two factors can jointly predict the attested lexicalization strategy of meanings that emerged over the recent history of English.

4.2 Data

To investigate strategy choice in lexicalization, we used three sources of data. First, we used large historical datasets of English text and word frequency. Second, we used an English dictionary with records of word sense definitions for compounds and polysemous words. To analyze historically emerging meanings, we identified a set of compounds and reuse items that emerged in the last century.

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 the Google Ngrams corpus (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 the Ngrams corpus.

Dictionary data. We primarily used a collection 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)[2]. This dataset consists of 3,220 frequent

[2]Note that this is different from the OED.

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

Table 4.1: Examples of OD sense definitions and their year of emergence according to the OED.

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[3], and 2) we obtained $43,482$ COHA lemmas and their definitions that appear in archived webpages of OD[4]. In total, this provided us with a set of $77,359$ senses for $30,760$ words. Example sense definition are shown in Table 4.1.

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.[5] We identified this subset by using sense definitions that contain culturally salient keywords (Cook, Lau, McCarthy, & Baldwin, 2014). To identify these keywords, we manually selected salient words that have changed the most in frequency

[3] `https://www.lexico.com/`

[4] `https://archive.org`

[5] Data can be found at `https://github.com/johnaot/strategy-prediction-data`

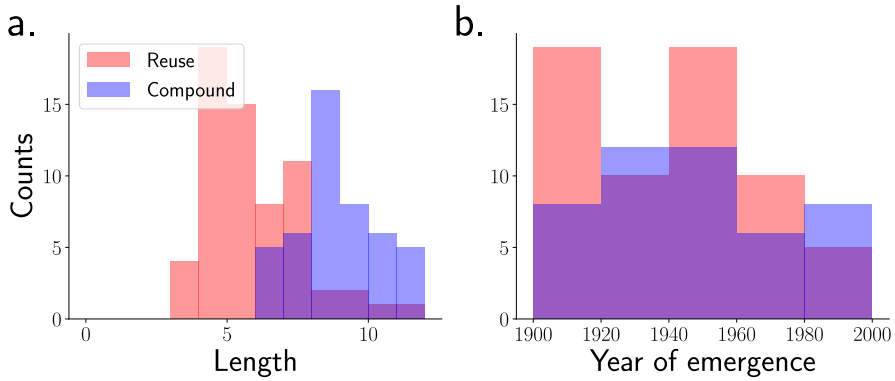

Figure 4.2: Descriptive statistics for our dataset of 20th-century word senses that are lexicalized by reuse and compounding.

aircraft, airport, broadcast, broadcasting, car, cinema, computer, data, electric, electrical, electronic, film, jazz, motor, phone, pilot, radio, software, television, video

Table 4.2: Set of cultural keywords.

since 1900 in COHA; our keywords are shown in Table 4.2. We shortlisted the senses of compounds from LADEC and the noun senses of polysemous words from Hu et al. (2019) since LADEC only provides noun compounds. After timestamping the shortlist by using the OED, we obtained 67 cases of reuse and 46 compounds, as well as their exact year of first citation in the OED. To guarantee positive communicative need[6], we filtered out senses that have zero frequency for two decades after their emergence. This left us with a final dataset of 63 meanings that are expressed via reuse, and 46 meanings that are expressed with compounds.

Figure 4.2 shows descriptive statistics for our dataset. Observe that the senses we collected emerged during the whole span of the 20th century, and that the orthographic (word) lengths of reuse items tend to be shorter than the lengths of compound words in the dataset.

[6]The main consideration here is related to presentation in Section 4.4, but regardless our results do not change significantly.

4.3 Computational methods

To evaluate our hypotheses on lexicalization strategies for emerging meanings, we first define the space of meanings obtained from our datasets. We then define the measures we used to quantify the factors involved in our hypotheses.

Meaning space. Let $M = \{m_1, m_2, ..., m_k\}$ be the set of meanings we used in our analyses. We defined each meaning m_i as a word sense that we obtained from OD. We represented each meaning as a 768-d vector by embedding its sense definition with Sentence-BERT (Reimers & Gurevych, 2019).[7]

We ensured that the embedding space reflects genuine semantic relatedness using three datasets of pairwise word similarity ratings: WordSim-353 (Finkelstein et al., 2002), SimLex-999 (Hill, Reichart, & Korhonen, 2015), and MEN (Bruni, Tran, & Baroni, 2014). We compared our sense embeddings based on sentence-BERT to a sense representation obtained by averaging the word2vec (Mikolov, Chen, et al., 2013; Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018) vectors of non-stopwords in sense definitions. We also compared these representations against word similarity directly obtained from pretrained word2vec embeddings. The last representation solely serves an upper bound in the context of our word similarity validation task, and it cannot be applied in our main analyses which rely on sense-level representations.

To evaluate these representations, we computed the spearman correlation of human ratings and the cosine distances between word meanings. For sense-level representations, we represented each word meaning by averaging the embeddings of the word's senses. The evaluation results are summarized in Appendix, Section B.3. We observe that Sentence-BERT embeddings are significantly better than the sense-level word2vec baseline, possibly because the former is able to better capture the dependencies among

[7]The model was trained on a corpus of contemporary texts, but here we make the simplifying assumption that every word sense corresponds to the same meaning across time.

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. We approximated the communicative need of a newly emerged meaning through its observed frequency after its earliest known occurrence in a diachronic corpus. Suppose that the meaning m_i emerged in year t, and suppose w is its attested label. Let $p_x(w)$ be the relative frequency of w in year x obtained from the Ngrams corpus, and let $p_x(m|w)$ be the proportion of tokens of w that express sense m in year x according to 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 specific time window X:

$$f(m) = \frac{1}{|X|} \sum_{x \in X} p_x(m|w) p_x(w) \tag{4.1}$$

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

Relative novelty. To quantify the novelty of an emerging meaning m_i relative to existing meanings, we first defined the existing lexicon $\mathcal{L}_t$. We started by dividing up the time period between 1880 and 2000 into consecutive intervals of 20 years, denoted $I_1, ..., I_6$. We then identified the list of senses that existed in I_{i-1} if year $t \in I_i$. Specifically, we automatically selected every sense m such that 1) its attested label w is polysemous and it satisfies $p_x(m|w) \geq 0.1$ for some $x \in I_{i-1}$, or 2) otherwise its attested label w occurs at least 10 times during I_{i-1}. We used these existing senses to define two versions of the lexicon: 1) the set $\mathcal{L}_t^{\text{sense}}$ which consists of all existing senses, and 2) the set $\mathcal{L}_t^{\text{word}}$

which contains word meanings obtained by averaging the existing senses of each word.

We measured the novelty of m_i relative to $\mathcal{L}_t$ via the cosine distance from m_i to $\mathcal{L}_t$:

$$d(m_i, \mathcal{L}_t) = \min_{m \in \mathcal{L}_t} \text{cos-dist}(m_i, m) \tag{4.2}$$

This measure characterizes the upper bound on the semantic similarity between m_i and any existing meaning. Intuitively, a low novelty score (i.e., small cosine distance to the nearest existing meaning) implies the opportunity to achieve greater cognitive efficiency via reuse, and vice versa[8].

4.4 Results

In this section, we first separately evaluate each of the two hypotheses we outlined previously. We then use the predictors motivated by these hypotheses to predict the historically attested choices of lexicalization strategy.

4.4.1 Evaluating the least effort hypothesis

Since the least effort hypothesis implies the lexicon prefers to express high-need meanings with reuse, we expected a difference between the average communicative needs of reuse meanings (e.g., "mouse") and compound meanings (e.g., "spreadsheet"). The results are illustrated in Figure 4.3. In both settings, reuse meanings tend to have higher communicative need compared to compound meanings. 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 $t(107) = 4.920$, $p < 0.001$. These results provide support for our least effort hypothesis.

[8]This is graphically illustrated in Figure 4.1b.

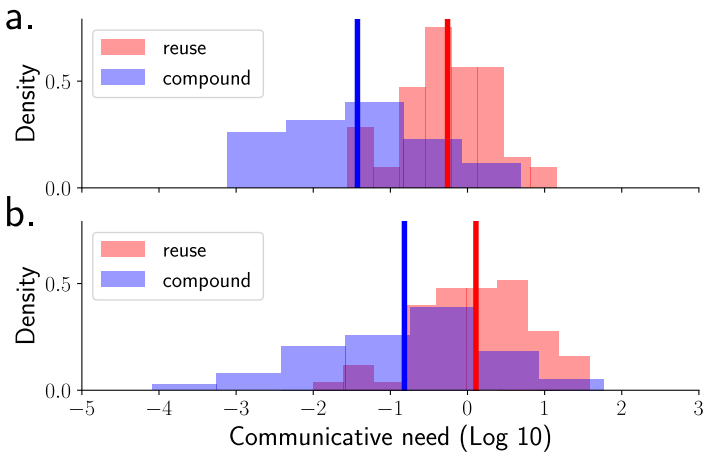

Figure 4.3: Communicative need estimated using a) historical data and b) contemporary data. Each vertical line indicates the mean over attested cases of reuse (red) or compounds (blue).

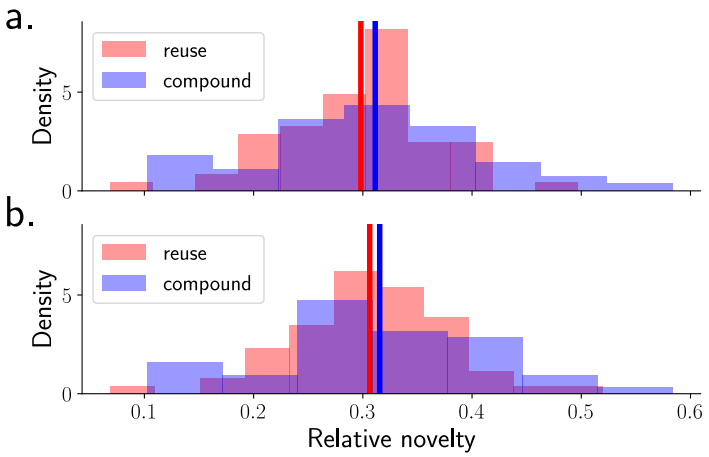

Figure 4.4: Relative novelty measured using a) the nearest existing sense and b) the nearest existing word. Each vertical line indicates the mean over attested cases of reuse (red) or compounds (blue).

4.4.2 Evaluating the novelty hypothesis

Similar to the previous analysis, we compared the average novelty of reuse and compound meanings to assess the novelty hypothesis. The results are illustrated in Figure 4.4. We observe that on average, the novelty of meanings expressed by compounds tends to be higher than the novelty of meanings expressed by reuse. We then compared the mean novelty of each group using t-tests. For novelty relative to existing senses, we obtained

Predictor	β	**Z statistic**	**p-value**
Intercept	-1.5353	-0.977	0.328
$f_{hist}(m)$	-2.8210	-3.488	< 0.001
$f_{now}(m)$	0.3194	0.618	0.537
$d_s(m, \mathcal{L}_t)$	3.7112	0.144	0.886
$d_w(m, \mathcal{L}_t)$	-7.6745	-0.309	0.757

Table 4.3: Summary of logistic regression results.

$t(107) = -0.801$, $p = 0.425$. For novelty relative to existing word meanings, we obtained $t(107) = -0.538$, $p = 0.592$. In both cases, we do not find evidence for our novelty hypothesis.

4.4.3 Predicting strategy choice from need and novelty

After testing our hypotheses individually, we explored whether communicative need and novelty can jointly predict attested lexicalization strategies. Since we focus on the choice between reuse and compounding, we formulated this task as a binary prediction problem. We proceeded by using a logistic regression model of the following form:

$$y(m) \sim \Big(1 + \exp(-(\beta_0 + \beta_1 f_{hist}(m) + \beta_2 f_{now}(m) + \beta_3 d(m, \mathcal{L}_t^{\text{sense}}) + \beta_4 d(m, \mathcal{L}_t^{\text{word}})))\Big)^{-1} \tag{4.3}$$

Here, $y(m)$ represents whether m is expressed as a reuse item or a compound, and the pair (f_{hist}, f_{now}) represents communicative needs estimated using historical and contemporary frequencies, respectively. We implemented the model using `statsmodel` (Seabold & Perktold, 2010).

To evaluate the model, we trained it on meanings that emerged before year 1950 ($n = 66$), and we tested it using meanings that emerged after year 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 the Wald test. The test results are summarized in Table 4.3. We observe that $f_{hist}(m)$ 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

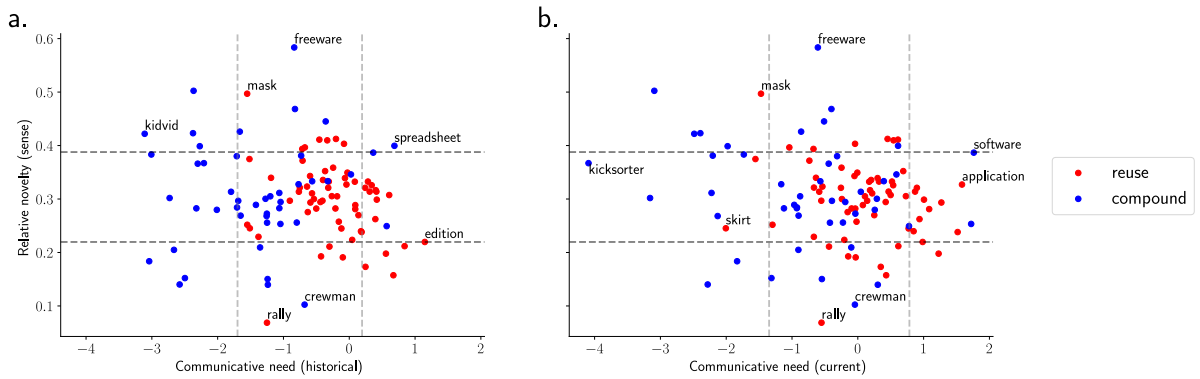

Figure 4.5: Scatter plots showing the communicative need and novelty of each emerging meaning. Each dot corresponds to an emerging meaning, and the extreme cases for each strategy are annotated. Dashed lines correspond to one standard deviation away from the mean over all meanings.

individual meanings in Figure 4.5; sense definitions for the labelled examples are shown in Table 4.1. Since the two novelty measures are highly collinear (Pearson correlation 0.978, $p < 0.001$, $n = 109$), we created the plot with respect to sense-level novelty only. In both panels of Figure 4.5, we observe that meanings 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 items, whereas in the other direction most meanings tend to be expressed as compounds. In contrast, meanings with novelty much lower than the mean do not appear to be dominated by a single strategy, and the same trend also applies to meanings with high novelty.

We also observe that the separation between the two strategies created by $f_{hist}(m)$ in Figure 4.5a is greater than the separation created by $f_{now}(m)$ in Figure 4.5b. This is demonstrated by the meanings at the extremes: for $f_{hist}(m)$, both the most needed and least needed meanings in the reuse group (*mask* and *edition*) have higher need than their counterparts in the compound group (*spreadsheet* and *kidvid*); however, for $f_{now}(m)$, the most needed meaning in the compound group (*software*) actually has higher need than its counterpart in the reuse group (*application*). This suggests that communicative need and strategy choice are sensitive to cultural changes over time.

4.5 Discussion

In this study, we sought to explain the strategy choice between reusing an existing word and coining a novel compound form. Building on previous functionalist accounts of the lexicon (e.g., Ramiro et al., 2018; Mollica et al., 2021), we hypothesized that pressures for communicative and cognitive efficiency may account for the choice between reuse and compounding. Specifically, our least effort hypothesis predicts that new meanings with high communicative need should be expressed via reuse, whereas our novelty hypothesis predicts high-novelty meanings should be expressed via compounding. Using two operationalizations of communicative need, our results provided evidence for the least effort hypothesis. Nonetheless, we did not find evidence that the novelty of a new meaning predicts lexicalization 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 word 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 meanings lexicalized by different strategies. However, it may be the case that most meanings in our dataset have relatively low novelty, making it difficult to distinguish the strategies. Another possibility is that high novelty disfavours 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 preferable than word reuse.

Besides communicative need and novelty, the context associated with an emerging meaning may also contribute to strategy choice. In particular, reuse may be preferable if the intended meaning tends to be predictable given the context in which it occurs, thereby reducing information loss in communication. For instance, within the context of programming, *queue* is an informative label because data structures are much more

likely than a line of people. On the other hand, in a retail store during the 1980s, using the same label for both landline phones and cellphones would likely lead to confusion. Future work should evaluate this hypothesis in formal terms and clarify its relation to existing accounts on the role of context in shaping word length (Piantadosi et al., 2011; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013) and colexification patterns (Karjus, Blythe, Kirby, Wang, & Smith, 2021; Brochhagen & Boleda, 2022).

Our methodology builds on recent work in natural language processing that utilizes large historical corpora to examine changes in the lexicon (e.g., Hu et al., 2019; Ryskina, Rabinovich, Berg-Kirkpatrick, Mortensen, & Tsvetkov, 2020). In this line of work, most closely related to ours is the study by Ryskina et al. (2020). Their work shows 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 examined similar factors, but we extended their analysis by examining the strategy choice for expressing emerging meanings.

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 lexical evolution. We connected strategy choice with existing theories of communicative and cognitive efficiency, and found that a pressure for least effort predicts the lexicalization strategy used to express new meanings. Future work may extend this study by refining its theoretical framework, by considering other lexicalization strategies such as morphological derivation, and by testing it with historical data at a larger scale.

Chapter 5

Modelling compounding via composition and analogy

The contents of this chapter are based on the following publication:
Xu, A., Kemp, C., Frermann, L., & Xu, Y. (to appear). Modelling compounding via composition and analogy across languages. In Proceedings of 47th the Annual Meeting of the Cognitive Science Society.

5.1 Introduction

In this chapter, we return to the issue of non-literal compound labels. In Chapter 3, our results suggest that compound labels for emerging concepts reflect an efficient tradeoff between informativeness and ease of use, but non-literal compound labels (e.g., *spreadsheet*, *blue-collar*) tend to be less efficient than literal compound labels (e.g., *birthday card*, *urban renewal*). Here, we explore how to provide a better account of non-literal compound labels by seeking to understand and model the cognitive processes that underlie compounding across languages.

We begin by considering traditional accounts of compounding that aim to identify regularities in the relations between compound constituents (e.g., Levi, 1978; Lieber,

1983; Jackendoff, 2010; Levin et al., 2019). For example, recall that Levi (1978) argues that head-modifier relations in noun compounds tend to belong to one of nine meta-relations, while other intended relations tend to be explicitly encoded in the compound (e.g., *eating* should not be removed from *fish-eating dinosaurs*). Moreover, Levin et al. (2019) show that English noun compounds denoting artifacts tend to invoke head-modifier relations highlighting events of their creation or use, whereas noun compounds denoting natural kinds tend to invoke relations highlighting their so-called essence (e.g., perceptual properties). The examples above show that traditional semantic analyses of compounding differ in scope and granularity. However, these accounts tend to focus on the semantic composition of compound constituents and ignore non-compositional processes involved in compounding.

An alternate approach examines the semantics of compounding by considering the general cognitive process of analogy (e.g., Ryder, 1994; Booij, 2010; Klégr & Čermák, 2010). In this view, one or more existing compounds that share the same constituent can form a productive pattern that can be modified to create new compounds; these new compounds are regarded as analogical compounds. Crucially, the shared constituent in a productive pattern tends to develop bound meanings that shift away from its original lexicalized meaning, which are then inherited by new compounds based on the pattern (Booij, 2010). For example, *blue-collar* and *white-collar* form a productive pattern that led to the creation of the new compound *green-collar* (Mattiello & Dressler, 2018). Here, the pattern X + *collar* signifies membership in a certain profession, which is distinct from compositional formations like *dog collar*. Compounding can create word combinations that share no constituent with any existing compound and thus do not involve analogy, but the non-compositional nature of analogical compounds suggests that analogy may nonetheless complement compositional accounts of compounding.

In this study, we formulate a computational account of compounding that combines both composition and analogy. Specifically, we hypothesize that speakers create new

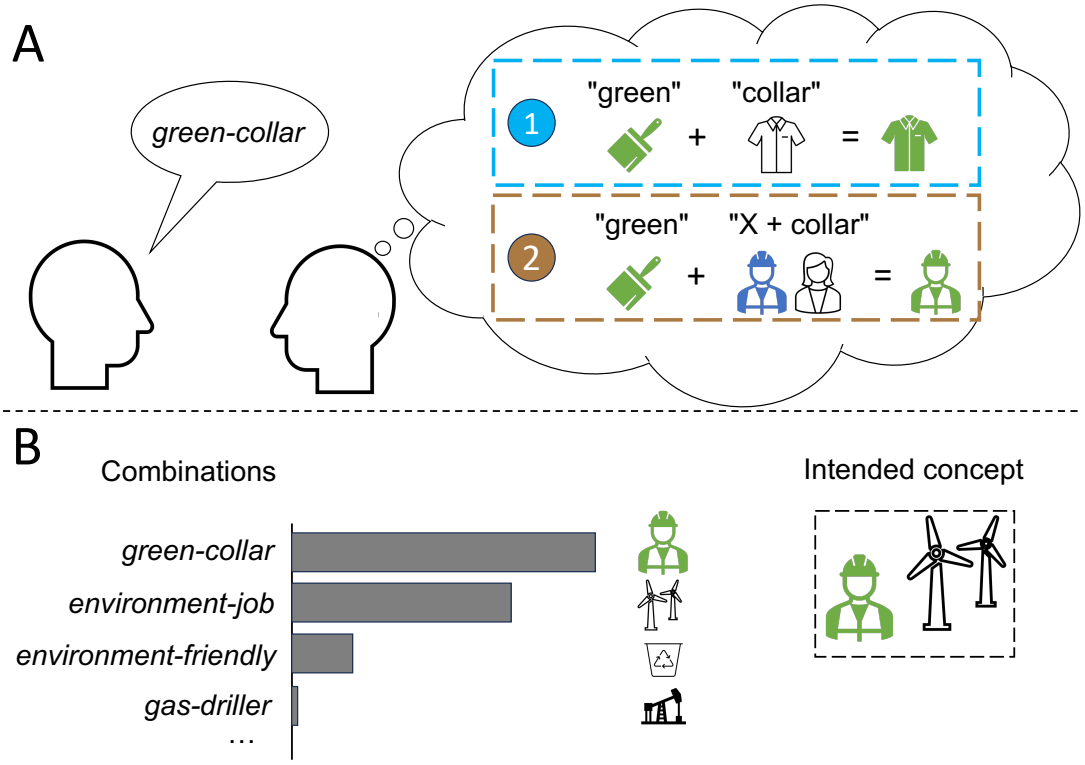

Figure 5.1: Analogy and composition in compounding. Panel (A) shows a listener that tries to interpret the compound *green-collar*. The listener chooses between 1) a compositional strategy and 2) an analogical strategy, in which the meaning of the pattern is based on *blue-collar* and *white-collar*. Panel (B) illustrates predictions on possible compound expressions for the concept "belonging to a green industry," which is illustrated on the right. Grey bars on the left illustrate the probability of choosing a compound, and the adjacent icons illustrate the corresponding listener interpretations.

compounds by inferring how word combinations may be interpreted by listeners. Listener interpretation may involve the composition of constituents, or in the case where one constituent overlaps with existing compounds that form a productive pattern, it may involve composing the bound meaning of the pattern with the other constituent. The speaker finally chooses a compound with an interpretation that is very similar to the intended concept. We illustrate this idea in Figure 5.1.

Our work builds on a line of computational work on vector-based semantic composition (e.g., Mitchell & Lapata, 2010; Dima et al., 2019). In this approach, word meanings are represented by real-valued vectors in semantic space, and semantic composition is

modelled by composition functions that take word vectors as input and produce a new vector as output. In addition to using them to infer the meaning of novel combinations, previous work has applied composition functions to derive measures to predict human behavioural data on the processing of both existing and novel combinations (Vecchi et al., 2017; Günther & Marelli, 2016; Marelli et al., 2017; Hsieh et al., 2025). Here we extend these studies by using vector-based composition to reconstruct attested combinations given their meanings.

Our work is related to but distinctive from existing computational models of analogy in morphology (e.g., Krott et al., 2001; Plag et al., 2023). With regards to compounding, analogy-based models have been applied to predicting the use of linking morphemes (Krott et al., 2001, 2007) and the position of word stress (Arndt-Lappe, 2011). For a new compound, these models typically compute the preference for an outcome (e.g., stress position) based on the similarity between the new item and a group of existing items in the lexicon that are associated with the outcome. In the current study, we propose a model that computes the probability of a possible combination for an intended concept by using the concept's similarity with the meanings of existing compounds that belong to a productive pattern.

In the following, we first formulate our account in computational terms. We then compare our model to simpler models by using them to predict the constituents of attested compounds in English, Chinese and German. Finally, we discuss the implications of our work and future directions.

5.2 Computational formulation

Our account assumes the following setting. Let $\mathcal{L}$ be the lexicon which contains the set of existing words, and let D be the set of existing compounds. Let C be a semantic space of real vectors, such that each word w corresponds to the semantic vector $c_w \in C$.

We define a compound family as a set containing existing compounds with the same head or modifier. We assume each family is partitioned into subfamilies, and each subfamily corresponds to a distinct productive pattern.[1] In this study, we focus on analogy with respect to compounds sharing the same head word (e.g., X + *collar*), and we leave modifier-based analogy for future work.

5.2.1 Probabilistic model of compounding

Given an intended concept $c \in C$ that is not expressed by compounds in D, we wish to predict an expression that consists of a head word $h \in \mathcal{L}$ and a modifier word $m \in \mathcal{L}$. We formalize this computational problem in probabilistic terms by applying Bayes rule:

$$p(m, h|c, D) \propto p(c|m, h, D)p(m|D)p(h|D). \tag{5.1}$$

There are three terms on the righthand side of Equation 5.1. The rightmost terms $p(m|D)$ and $p(h|D)$ represent the prior probability of selecting a specific modifier and head without considering the intended concept. The third term $p(c|m, h, D)$ is the likelihood that captures the similarity between c and listener interpretations of an observed combination. Importantly, all terms are dependent on statistical tendencies among existing compounds in the set D.

Likelihood. To formulate the likelihood, we first formalize listener interpretations via composition and analogy. Since we assumed meanings are represented in semantic space, a natural way to model composition is to use vector-based composition functions (e.g., Mitchell & Lapata, 2010). Let $f(\cdot, \cdot)$ be a general composition function that takes a pair of head and modifier and outputs a vector in C that represents the compositional interpretation. Similarly, let $g_i(\cdot|h)$ be a function that composes a modifier and existing

[1] For example, the subfamily containing *software* and *freeware* is separate from the subfamily containing *tableware* and *kitchenware*.

compounds in the i-th subfamily associated with head h, such that it outputs a vector that represents an analogy-based interpretation. These functions correspond to processes illustrated in Figure 5.1A.

We formulate the likelihood in Equation 5.1 by using the similarity between c and listener interpretations of m and h. Because a listener may choose between the compositional interpretation and an analogy-based interpretation, if available, we define the likelihood as an average of similarities:

$$p(c|m,h,D) \propto q_0 \text{sim}(c, f(m,h)) + \sum_i q_i \text{sim}(c, g_i(m|h)) \tag{5.2}$$

$$\text{sim}(c_1, c_2) \propto \exp(s \cdot \cos(c_1, c_2)) \tag{5.3}$$

where $0 \leq q_i \leq 1$ is the probability of choosing a specific interpretation, $s > 0$ is a sensitivity parameter, and $\cos(\cdot,\cdot)$ is cosine similarity. We choose to use cosine because it is standard in previous applications of composition functions (e.g., Mitchell & Lapata, 2010).[2]

We define the probability of invoking a specific interpretation by using the prevalence of each pattern among lexicalized compounds. Let N_h be the size of the family associated with h, and let $N_{h,i}$ be the size of its i-th subfamily. We define q_j by using family and subfamily sizes:

$$q_0 = \frac{\theta}{N_h + \theta},\ q_{i>0} = \frac{N_{h,i}}{N_h + \theta} \tag{5.4}$$

where $\theta > 0$ is a pseudocount that determines the frequency of invoking composition. Equation 5.4 is inspired by previous work showing that the ease of understanding combination w is sensitive to the frequency of head-modifier relations associated with the constituents of w (Gagné & Shoben, 1997; Maguire, Maguire, & Cater, 2010).

[2]We assume every c is on the unit sphere, which makes Equation 2 a mixture of von-Mises-Fisher distributions. In this case, the omitted normalizing constant is only a function of s.

Priors. In principle, any open-class word in the lexicon can be chosen as the head or modifier, but the distribution of compound family sizes also tends to be skewed. Previous work in psycholinguistics suggests that morphological family size facilitates processing (e.g., De Jong, Feldman, Schreuder, Pastizzo, & Baayen, 2002; Martín, Bertram, Häikiö, Schreuder, & Baayen, 2004), and thus we hypothesize that words corresponding to larger families tend to be preferred by speakers:

$$p(h|D) \propto N_h + \alpha,\ p(m|D) \propto N'_m + \beta \tag{5.5}$$

where $\alpha, \beta > 0$ are pseudocounts of compound families, and N'_m is the number of compounds in D with modifier m. Smaller pseudocounts increase the preferences for reusing a word from a large family, and vice versa.

5.2.2 Functions of compound interpretation

We now specify the composition function $f(\cdot, \cdot)$ and the function $g_i(\cdot|h)$ that captures analogical reasoning with respect to the i-th pattern associated with h. Given a modifier m and a head word h, we define their semantic composition using the simple additive function (Mitchell & Lapata, 2008):

$$f(m, h) = c_m + c_h \tag{5.6}$$

Because Equation 5.6 composes the lexicalized meanings of m and h, the resulting vector represents a compositional interpretation (e.g., *dog* + *collar*).

If the head h is associated with a compound family in D, the interpretation may also be obtained from analogy. Let $r_{h,i}$ be the bound meaning of the i-th pattern associated with head h (e.g., the job-related meaning of X + *collar*). We define the interpretation via analogy with respect to this pattern by composing its bound meaning with the modifier

word:

$$g_i(m|h) = c_m + r_{h,i} \tag{5.7}$$

In contrast to Equation 5.6, here the interpretation is based on a bound meaning at the pattern level.

The bound meaning $r_{h,i}$ should capture how existing compounds that form the corresponding pattern tend to relate to their modifiers.[3] Here, we represent relations in semantic space via vector differences (e.g., Rumelhart & Abrahamson, 1973; Vylomova, Rimell, Cohn, & Baldwin, 2016). Let $S_{h,i}$ be the i-th subfamily associated with head h, and m_w be the modifier of compound w. We define $r_{h,i}$ via the central tendency of these relations:

$$r_{h,i} = \frac{\mu_{h,i}}{\|\mu_{h,i}\|_2},\ \mu_{h,i} = \sum_{w \in S_{h,i}} c_w - \gamma c_{m_w} \tag{5.8}$$

where $\gamma \geq 0$ is an additional free parameter that governs the importance of the modifier in contributing to relational meaning. Equation 5.8 normalizes $r_{h,i}$ to ensure its magnitude is on par with c_m in Equation 5.7.

Note that in the case where a pattern is based on a single existing compound w, Equation 5.8 reduces to a single vector difference. If we plug this difference into Equation 5.7 and disregard the coefficients on each semantic vector, we obtain the parallelogram-model solution to a proportional analogy involving four concepts in semantic space (e.g., after observing *software*, the listener may infer its meaning by solving "hard" : "hardware" :: "soft" : ?).

[3]For example, X + *ware* may represent a computer program characterized by the modifier.

5.3 Materials and methods

In the current study, we evaluate a version of our model that assumes a single productive pattern per compound family.[4] To do so, we first compiled datasets of compound words drawn from English, Chinese, and German, which are commonly studied but show crosslinguistic differences in morphology (e.g., J. Xu & Li, 2014; Günther, Smolka, & Marelli, 2019). We then split each dataset into training and test sets, which simulate the lexicon, the set of existing compounds, and the concepts to be expressed via compounding.

Datasets. We compiled large datasets of English, Chinese, and German compounds. The English dataset consists of closed English compounds from the Large Database of English Compounds (Gagné et al., 2019) and entries that are labeled as etymologically obtained from compounding in a scrape of Wiktionary (W. Wu & Yarowsky, 2020); these compounds were intersected with Princeton WordNet (Fellbaum, 1998) after filtering out combinations involving stopwords and chemical names in WordNet. The Chinese dataset was obtained by using 2-character compounds in the Chinese Open WordNet (Wang & Bond, 2013), and words that are chemical names according to WordNet were removed. The German dataset was obtained from a large compilation of German compounds (Günther, Marelli, & Bölte, 2020), CELEX (Baayen, Piepenbrock, & Gulikers, 1995), and the same scrape of Wiktionary. We removed combinations involving stopwords or translations of English chemical names that were obtained from the Google Translate API. All compounds in the datasets contain exactly two constituents; we ignored linking elements in German and English compounds and we removed compounds containing other compounds.

In all languages, we implemented the semantic space using word embeddings trained

[4] All code used in analyses is available at `https://github.com/johnaot/Compounds-compose-analogy`

on Wikipedia (Mikolov, Chen, et al., 2013; Li et al., 2018; Yamada et al., 2020); all vectors were normalized. After ensuring the compounds and constituents map to vectors, we obtained 5,093 English compounds, 11,420 Chinese compounds, and 42,715 German compounds. Note that this implies the remaining Chinese character constituents also function as independent words.

Methods. For each language, we set the lexicon $\mathcal{L}$ to be the set of all constituent words in the compound dataset. The lexicon contained 2,646 words, 3,535 characters, and 9,487 words for English, Chinese, and German, respectively. We then randomly sampled at most 10,000 compounds from the dataset and applied 5-fold cross-validation to the subset. In all languages, we assumed all compounds are right-headed, except for Chinese verbs which we assumed are left-headed.[5]

We used the training set to estimate model parameters. For each training set, we randomly set aside 3/4 of compounds to initialize the set D. We used the remaining 1/4 as a development set to estimate the parameters via a limited grid search, searching over $\{10^n : n = -2, -1, .., 2\}$ for α and β, $\{x/(1-x) : x = 0.5, 0.4, 0.3, 0.2\}$ for γ, $\{1, 10, 100\}$ for θ, and $\{0.01^{-1}, 0.02^{-1}, ..., 1\}$ for s. For every combination of α, β, γ and θ, we first set s by maximizing the likelihood $\prod_{D'} p(c|m, h, D)$, where D' contains compounds in the development set, and then we ranked all possible combinations given the meaning of each attested compound in D' according to the score given by Equation 5.1. We chose the set of values that maximizes the mean reciprocal rank (MRR) obtained from the ranks of attested combinations.

We used the test set to evaluate the model by using Equation 5.1 to score each attested combination given its meaning. The full model was compared with simpler baselines: 1) a model that only involves composition, 2) a model that sets the priors to uniform, 3) a model that combines the previous two settings, 4) a prior-only model, and 5) a uniform

[5]Ceccagno and Basciano (2007) estimates that about 75% of Chinese compound verbs are left or two-headed. These words are usually combinations of V + V or V + N.

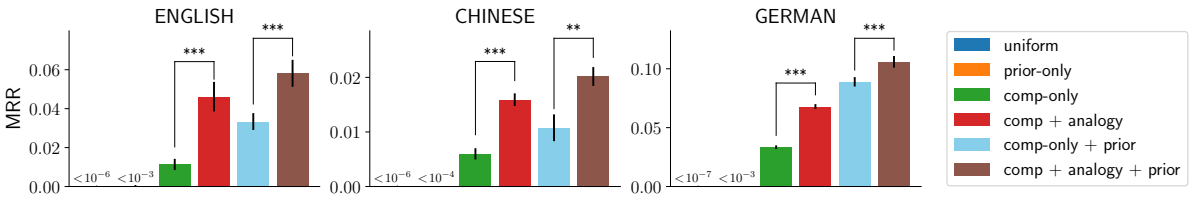

Figure 5.2: Comparing the full model against simpler variants across languages. Each bar shows the MRR averaged over five folds, and error bars indicate standard deviation across the folds. Stars indicate the p-values obtained from comparing models that involve both processes to composition-only models ("**" indicates $p < .01$ and "***" indicates $p < .001$).

distribution. The parameters of all baseline models were estimated in the same way as the full model based on the training set.

5.4 Results

We first evaluate the model by reconstructing an attested combination given its meaning. We then perform fine-grained analyses of bound meanings at the pattern level and across compound types.

Model evaluation. Figure 5.2 summarizes model performance through MRR. We observe that the full model obtained the best performance in all three languages. We also observe that the priors tend to improve model performance across languages and that the model with full likelihood but no prior is superior to composition-only models in English and Chinese. Similar observations can be made if we used top-k accuracy to evaluate the models.

Figure 5.3 shows the ranks predicted by the two models with priors and the (head-based) family size of every attested compound. We observe that the full model tends to dominate the simpler baseline across most family sizes. We also observe that predicted rank tends to improve with family size for both models in the cases of English and German, but this only applies to the full model in the case of Chinese. This is confirmed by the correlations between predicted rank and family size in Table 5.1, and it may be in part explained by the previous observation that including priors which encode family

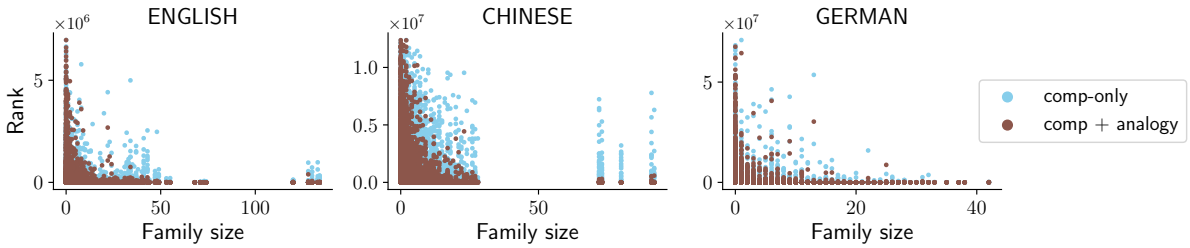

Figure 5.3: Comparing the full model and the composition-only model (including priors) in terms of attested-compound rank across family sizes. Predicted ranks lower in magnitude indicate better performance.

Language	Model type	Spearman ρ	p-value
English	comp-only	-0.276	$< .001$
	comp + analogy	-0.491	$< .001$
Chinese	comp-only	-0.058	$< .001$
	comp + analogy	-0.421	$< .001$
German	comp-only	-0.240	$< .001$
	comp + analogy	-0.424	$< .001$

Table 5.1: Correlation between predicted rank and family size.

Head	Lexicalized meaning NNs	Bound meaning NNs
front	rear, frontmost	downtown, condo
head	coach, helm	dopy, cruelness

Table 5.2: Samples of nearest neighbours (NNs) of the lexicalized and bound meanings of certain head words.

size information improves performance.

Bound meanings at the pattern level. Table 5.2 shows a sample of the neighbours of the lexicalized meanings and bound meanings of the same head words, which were retrieved from the vocabulary of our embeddings. We can see that X + *front* clearly developed a property-related meaning (as in *beachfront*) distinct from the original location-related meaning, and X + *head* obtained a meaning that is more related to personal character (as in *metalhead*). These English examples illustrate the semantic shift of pattern-level meanings from the lexicalized meanings of head words.

We hypothesized that compounds tend to inherit the bound meanings of productive patterns to a degree greater than their degree of inheritance from the lexicalized meanings of head words. We quantified the degree to which a specific meaning is inherited by measuring the cosine similarity between the former and the intended meaning

Language	Spearman ρ	p-value	N
English	-0.487	$< .001$	$4,187$
Chinese	-0.316	$< .001$	$8,625$
German	-0.413	$< .001$	$7,856$

Table 5.3: Correlation between cosine similarity and model predicted rank.

of a compound. Figure 5.4A shows the distribution of differences given by subtracting bound-meaning similarities from lexicalized-meaning similarities. We observe that bound meanings are on average much more similar to intended compound meanings in English ($t(4,186) = 29.57, p < 0.001$), Chinese ($t(8,624) = 109.40, p < 0.001$), and German ($t(7,855) = 13.01, p < 0.001$). A similar trend can be observed in Figure 5.4B. These results may in part explain the better performance of the full model since higher similarity between bound and intended meanings correlates with better predicted rank (see Table 5.3).

In Figure 5.4, we can also observe that differences between lexicalized and bound meanings in terms of their similarity with intended meanings tend to be smaller for German compounds relative to English compounds. This is consistent with an earlier empirical finding that German complex words tend to be more transparent than English ones (Günther et al., 2019) and earlier observations by linguists suggesting that the German lexicon is more motivated than the English lexicon (de Saussure, 1916; Ullmann, 1953).

Literal and non-literal head words. Recall that an advantage of analogy over composition is that it is capable of addressing certain non-compositional compounds (e.g., *green-collar*). We examined one type of non-compositional compound known as exocentric compounds, in which the head is not a hypernym of the compound (e.g., *seahorse* is not a *horse*). Specifically, we hypothesized that the full model is able to better account for exocentric compounds than the best-performing composition-only baseline. Following our methods in Chapter 3, we used WordNet (Fellbaum, 1998) to categorize English noun compounds into exocentric and non-exocentric (or endocentric) compounds.

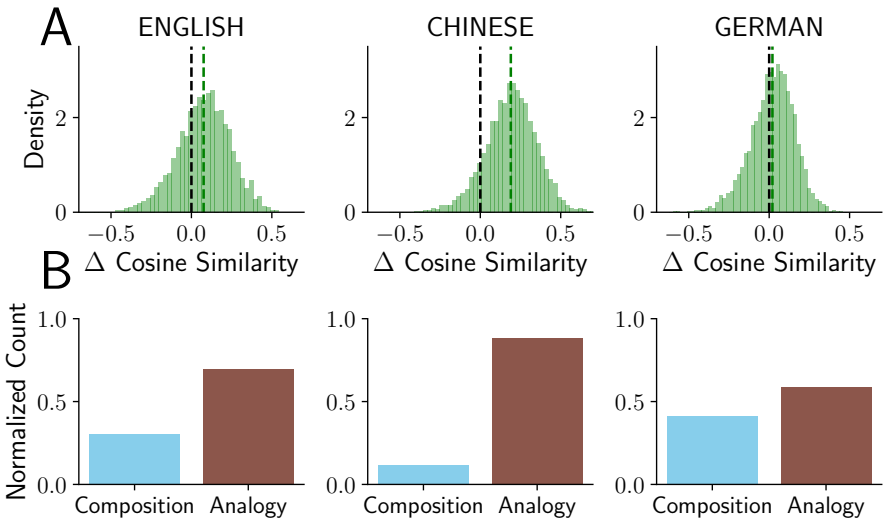

Figure 5.4: Semantic similarity between head and compound meanings. Panel (A) shows the distribution of differences in similarity between intended compound meanings and different meanings of the head; green lines show mean differences. Panel (B) shows the proportion of compounds that are more similar to lexicalized meanings and the proportion of compounds that are more similar to bound meanings.

Model type	Compound type	MRR	N
comp-only	endocentric	0.0482	1,962
	exocentric	0.0257	2,818
comp + analogy	endocentric	0.0660	1,962
	exocentric	0.0565	2,818

Table 5.4: Comparing model performance across English endocentric and exocentric noun compounds.

We compare model performance across compound types in Table 5.4. We observe that both models tend to perform better for endocentric compounds. Although we observe that the ranking of models remains the same across compound types, we also observe that the performance gap across types appears to be smaller for the full model.

We demonstrate the relative advantage of the full model with English examples in Table 5.5. Similar to a composition-only model, the full model is able to generate compositional expressions, which accounts well for compositional compounds like *taxicab* but not for non-compositional and non-analogical compounds like *mastermind*. However, the full model gains an advantage when an analogy-based interpretation of the attested combination is highly similar to its intended meaning. This can be demonstrated by the

Target	F. size	Rank	Top model predictions
taxicab	0	23	taxi-car, taxi-bus
mastermind	0	13k	evil-leader, shadow-leader
beachfront	3	1	beach-front, beach-land
metalhead	43	4	metal-boy, metal-man

Table 5.5: Examples of predictions from the full model. The second and third columns show the family size and predicted rank of the attested compound (target), respectively.

examples of *beachfront* and *metalhead*: using the bound meanings of X + *front* and X + *head*, the model assigned similar ranks to non-compositional attested combinations and hypothetical combinations that have similar interpretations but are compositional (e.g., *beach-land*, *metal-man*).

5.5 Discussion

We have presented a computational account of compounding that integrates the cognitive processes of composition and analogy. Using datasets of compounds from English, Chinese, and German, we tested computational models that reconstruct these compounds given their intended meanings. We showed that our full model, which integrates both processes, performs better than simpler baselines across all three languages.

Our work connects two separate lines of research on compounding by integrating the underlying cognitive processes. Specifically, traditional accounts focus on regularities among compositional relations between the constituents of attested compounds (e.g., Levi, 1978; Jackendoff, 2010), which correspond to the composition process, while a separate line of work appeals to the general cognitive process of analogy (e.g., Booij, 2010; Klégr & Čermák, 2010). In our account, a speaker creates a new compound by choosing a word combination that has an interpretation similar to their intended concept. Since compound interpretation may take place via the composition of the constituents or via analogical reasoning with respect to existing compounds, both processes are naturally integrated into our account.

Our account emphasizes the similarity between listener interpretations and intended

meanings, and thus is closely related to functionalist accounts of compounding which argue that new compounds are shaped by a pressure for informativeness (see Chapter 3). However, these accounts typically assume that listeners interpret compounds via some type of semantic composition, and thus they have limited applicability to the creation of non-compositional compounds. Our findings suggest that functionalist accounts may provide a better explanation for certain non-compositional compounds by considering analogical processes in compound interpretation.

The full model examined in this paper can be extended in various ways. For example, there are various composition functions that are more involved than our simple additive function (e.g., Mitchell & Lapata, 2010; Marelli et al., 2017), which may be able to better capture the regularities of compositional compound interpretation. Moreover, we made the simplifying assumption that every compound family corresponds to a single productive pattern. In contrast, existing semantic analyses of compounding that emphasize analogy tend to consider patterns that are subsets of a compound family or local analogies based on a single existing compound (e.g., Booij, 2010; Mattiello & Dressler, 2018). One way to model productive patterns in a more fine-grained manner is to use clustering to identify meaningful compound subfamilies, which can be straightforwardly integrated into our computational model.

Since we formulated our account in computational terms, we were able to validate the role of composition and analogy in compounding by reconstructing attested compounds. In contrast, theoretical accounts of morphology have often associated analogy with unpredictability (Arndt-Lappe, 2015). Although recent computational models have in part addressed this concern (e.g., Krott et al., 2001; Plag et al., 2023), to our knowledge we have presented the first computational account that specifies how analogy can predict the head and modifier in new compounds. In the future, our approach may be extended to analyze a more typologically diverse set of languages and datasets of recent neologisms.

Chapter 6

Conclusions and future directions

6.1 Thesis summary

In this thesis, I presented a large-scale, computational analysis of compound expressions from the perspective that language is shaped by pressures from communication and learning. Previous accounts of compounding have focused on literal compound labels and treated them in isolation from word reuse. By building on efficiency-based and modelling-based accounts of lexical items, I showed that the competition among compound labels and reuse items can be in part explained by a drive for efficient communication, and that non-literal compound labels can be predicted by modelling analogy and composition. My work therefore establishes a new connection between compounding and computational analyses of language in cognitive science.

In particular, I first presented a unified account of compounding and word reuse by building on the idea that language is shaped to support efficient communication. I proposed that compounds and reuse items reflect opposing pressures for informativeness and ease of use, respectively, such that attested expressions achieve an efficient tradeoff between the two pressures. I used a computational formulation of this idea to analyze reuse items and compounds that emerged in English, French, and Finnish during the last

century. I showed that attested labels are more efficient than hypothetical labels, which suggests that these expressions are shaped by pressures from communication. Nonetheless, while this account applies to both literal and non-literal compound labels, I also found that literal compound labels tend to be more efficient than non-literal labels.

The second study examined strategy choice in lexicalization. Although it predicts that efficient labels are preferred over inefficient ones, the first study does not explain why certain concepts are expressed by compounds but not by equally efficient existing words. To address this issue, I hypothesized that established notions of communicative and cognitive economy might predict the choice between word reuse and compounding for expressing an emerging concept. I tested this hypothesis via a computational analysis of English word meanings that emerged over the last century, and I found that strategy choice may be explained partly by a pressure for least effort.

Finally, I further examined the creation of non-literal compound labels by adopting a modelling-based approach and considering cognitive processes involved in compound interpretation. Specifically, I hypothesized that jointly considering analogy and composition provides a better account of non-literal labels. I formulated a computational model that integrates these cognitive processes to predict the compound label of a given concept, and I showed that the model more accurately reconstructs both literal and non-literal compound labels that are attested in English, Chinese, and German than simpler baseline models.

6.2 Future directions

The perspectives and computational approaches that I developed in this thesis provide a foundation for new lines of investigation into the creation of compound expressions and other aspects of the temporal evolution of the lexicon. I conclude by outlining several directions for future work.

6.2.1 Additional speaker-level processes

In Chapter 5, I have examined how speakers create compounds to express specific concepts by modelling analogy and composition in computational terms. However, there is a broader space of processes that may be employed by speakers to create new compounds. A natural direction for future work is to explore how the current model may be integrated with these processes to provide a better account of the data.

While I have modelled composition as a form of averaging, existing studies in linguistics suggest that compounding involves more specific forms of composition. This is apparent in classes of compounds where the head-modifier relations tend to be systematic. For instance, in many cases of incorporation and synthetic compounds, the intended meaning can be derived by treating the modifier as an argument of the verb contained in the head (Mithun, 1984; Bauer, 2017). This is exemplified by *house-hunt* and *taxi driver*, which refer to "hunting for a house" and "someone that drives taxi", respectively. Similarly, the intended meaning of an appositional compound is the intersection of the constituents (Wälchli, 2005). One classic example of this is *singer-songwriter*, which refers to someone that is both a singer and a songwriter. Future work may explore how to integrate these processes into computational models of compounding; one possibility is to devise a general composition function that captures all of these cases, whereas an alternative solution is to model each case with a separate function.

Moreover, it is plausible that the formation of certain exocentric compounds involves cognitive processes in word meaning extension. Recall that in Section 2.2.1.1, we saw that the meanings of exocentric compounds may be derived from their constituents by interpreting the latter in non-literal ways (Benczes, 2006; Bauer, 2010; Jackendoff, 2016). Although some exocentric compounds may be better explained by analogy (e.g., *green-collar*), others may have been created by speakers that chose to extend the meaning of a word before combining it with a modifier. For example, *firedog* refers to "iron support for burning logs" (Benczes, 2006, p. 175); to create this compound, the speaker may have

first extended the meaning of *dog* such that it refers to a metal support. Future work may explore how to model word meaning extensions in compounding; one intuitive starting point is to adapt existing modelling work on the cognitive processes in extensions (e.g., Ramiro et al., 2018; Pinto Jr & Xu, 2021) to the current context.

6.2.2 Connections to lexical evolution

In this thesis, I have mainly viewed compounding as a strategy for lexicalizing emerging concepts and enriching the vocabulary of a linguistic community. Here, I describe two extensions to the connections between communicative efficiency and lexicalization strategies that were established in Chapters 3 and 4.

6.2.2.1 Efficiency and language change

In Chapter 3, I proposed that conventional compounds and reuse items are shaped by competing pressures for informativeness and ease of use. By showing that these lexical items reflect relatively efficient tradeoffs, this study adds to the growing line of work suggesting that language is shaped to support efficient communication (e.g., Regier et al., 2015; Kemp et al., 2018; Gibson et al., 2019; Futrell & Hahn, 2022). However, similar to earlier efficiency-based accounts, a limitation of this study is that it did not specify the mechanisms that gave rise to the efficiency that was observed.

As a starting point, it may be useful to consider the different ways in which speakers relate to linguistic innovations. While some speakers adopt innovations from others, other speakers, typically known as innovators, are responsible for introducing linguistic innovations to the community (Labov, 2011; Milroy & Milroy, 1985; Del Tredici & Fernández, 2018). These two types of speakers may also correspond to different mechanisms that shape the efficiency of conventional labels. On the one hand, innovators may be motivated to create labels that are efficient (e.g., by using processes in Chapter 5) to achieve communicative success. On the other hand, non-innovators may be less prone

to adopting inefficient labels. This is consistent with studies suggesting that transparent form-meaning pairs are easier to retain (Rodd et al., 2012; Brusnighan & Folk, 2012) and that language users often view compounds with redundant constituents (e.g., *wind flag*) as unplausible (Downing, 1977; Günther & Marelli, 2016).

An important direction for future work is to examine the mechanisms that underlie efficient tradeoffs in lexicalization. To this end, it may be useful to examine how expressions for emerging concepts achieve efficient tradeoffs in controlled experiments with artificial languages: existing experimental work has successfully addressed many related issues, including the mechanisms that underlie the efficiency of semantic categories (Carstensen, Xu, Smith, & Regier, 2015; Carr, Smith, Culbertson, & Kirby, 2020) and word lengths (Kanwal et al., 2017), as well as the conditions that drive speakers to create compositional forms (Kirby et al., 2015; Raviv et al., 2019a, 2019b) and reuse items (Karjus et al., 2021; K. Smith, Bowerman, & Smith, 2025).

6.2.2.2 Additional lexicalization strategies

In Chapters 3 and 4, I have examined conventional labels produced by reuse and compounding, which are among the most common lexicalization strategies. However, there are various other strategies that play a non-trivial role in enriching the lexicon.

For instance, let us consider additional strategies that produce expressions similar to compounds. In English and other European languages, neoclassical compounds are lexical items with constituents from the vocabulary of Greek or Latin (Bauer, 1998). These items are relatively common (e.g., field-of-study words ending in *-logy*), but they are distinct from compounds because the Greek or Latin constituents are usually unable to function as independent words. Other compound-like expressions involve connecting word pairs through an additional grammatical element (Finkbeiner & Schlücker, 2019; Pepper, 2022). These elements may involve the genitive case (e.g., Japanese *kumo no so*; lit. spider's nest, 'spiderweb'), prepositions (e.g., *part of speech*, *man-at-arms*), or suffixes

for deriving relational adjectives (e.g., Dutch *stalen zenuwen*; lit. made-of-steel nerves, 'nerves of steel'). These expressions are less common than compounds but nonetheless numerous, making up about 48% of a crosslinguistic sample of lexicalized expressions that contain two nouns and potentially additional elements (Pepper, 2022).

Although they produce structurally different labels, the diverse lexicalization strategies of human language share a common communicative function by enabling the expression of emerging concepts. From this perspective, it is natural to ask whether the view that language is shaped to support efficient communication has the potential to account for strategies beyond reuse and compounding. The studies in Chapters 3 and 4 provide a concrete starting point for addressing this question. For instance, it should be straightforward to assess the efficiency of lexicalized English N + P + N constructions (e.g., *sense of humour*), given that they are similar to compounds and are covered by the lexical resources used in Chapter 3 (e.g., WordNet; Fellbaum, 1998).

Appendix A

Supplementary information for Chapter 2

A.1 Exceptions to criteria for compounds

This section provides examples of languages which contain items that are commonly regarded as compounds but violate at least one of the typical compound-defining conditions described in Section 2.1.

Condition 1: Compounds and phrases have distinct stress positions.

- Exception 1: In Romance languages, there is no distinctive word stress that distinguishes compounds from phrases (e.g., Rainer & Varela, 1992; Arnaud, 2015).
- Exception 2: In Hebrew, a noun-noun compound has the same stress pattern as a noun-noun phrase (Borer, 2011).

Condition 2: Morphological inflection is carried by a single constituent.

- Exception 1: In Romance languages, it is typical to inflect both constituents in coordinative compounds (Finkbeiner & Schlücker, 2019). For example, consider

the pluralized French compound *auteurs-compositeurs* 'songwriters'.

- Exception 2: In Finnish, both constituents of certain compounds are modified when they are inflected (Hyvärinen, 2019). For example, the allative of the compound *nuoripari* 'newlyweds' is *nuorelleparille*, which is modified by the suffix *-lle*.

Condition 3: Neither constituent is allowed to interact with other words in a sentence as individual words that are independent of the compound.

- Exception 1: In Chinese, constituents in verb-noun compounds may be separately modified by other words (Huang, Wang, & Chen, 2022). For example, the first constituent of *lí-hūn* 'divorce' is modified in *lí liǎng cì hūn* 'divorce twice'.

- Exception 2: In Hungarian, auxiliaries may split up noun-verb compounds in sentences (Kiefer, 1990). For example, the second constituent of *levelet ír* 'letter-write' is modified in *Péter levelet akar írni* 'Peter wants to write letters'.

A.2 Defining non-literal heads

Although the hypernymy-based definition of semantic heads is well-established for endocentric compounds (Bloomfield, 1933; Bauer, 2017), to my knowledge there is no established notion of semantic heads for exocentric compounds. In this section, I develop a notion of semantic heads for exocentric compounds, and I refer to constituents that correspond to this notion as non-literal heads. Following the intuition that the head is the more important constituent (Zwicky, 1985; Scalise & Fábregas, 2010), the main desideratum is that non-literal heads should be intuitively more or equally important relative to the non-heads in terms of their contribution to compound meanings. The rest of this section defines non-literal heads for different types of exocentric compounds based on the typology proposed by Bauer (2010): Section A.2.1 focuses on exocentric

compounds that are subordinative, Section A.2.2 focuses on coordinative compounds, and Section A.2.3 provides a brief discussion of the current definition.

A.2.1 Subordinative compounds

Following Bauer (2017), subordinative compounds are combinations in which one constituent is in some sense subordinate to the other.[1] Assuming that the meaning of an exocentric compound can always be derived via a non-literal interpretation of the constituents, I show that one can use the non-literal interpretation to treat the word combination as an endocentric compound that contains a semantic head in the traditional sense. This is achieved through one of two strategies:

- **Strategy 1:** If the meaning is derived by interpreting constituent B in a non-literal way and modifying it with constituent A, then constituent B under its non-literal interpretation functions as the traditional semantic head of the compound AB.

- **Strategy 2:** Otherwise, if the meaning is derived by interpreting the entire compound AB in a literal way and extending its literal interpretation by a figure of speech, then AB under the literal interpretation is an endocentric compound that possesses a traditional semantic head.

For exocentric compounds that are subordinative, I define the non-literal head as the traditional semantic head that is identified through one of the above strategies. In the following, I discuss how these non-literal heads are identified with respect to the subclasses of exocentric compounds proposed by Bauer (2010), except for coordinative compounds which will be discussed in the next section.

[1] Based on Bauer's discussion, this definition seems to at least subsume the non-coordinate subclasses defined by Bisetto and Scalise (2005). For compounds in these subclasses, one constituent is subordinate in the sense that it is an argument of the other constituent or it modifies the other constituent.

A.2.1.1 Bahuvrihi

The compound *AB* is bahuvrihi or possessive if its intended meaning is "having *AB*" or "an entity that has *AB*". Because it originally refers to "having much rice" or "one who has much rice" in Sanskrit, the compound *bahuvrihi* (lit. much-rice) itself is an example of bahuvrihi compounds (Bauer, 2010). There are also various bahuvrihi compounds in English, such as *paleface*, *birdbrain*, and *white-collar*. These compounds typically consist of a noun that another entity possesses and a second constituent that modifies the noun, and they are exocentric because none of the constituents is a hypernym of the intended property or referent. Nonetheless, observe that the meaning of a bahuvrihi compound can be derived from some literal interpretation of both constituents and its metonymic extension, which would enable the identification of non-literal heads via the second strategy. For example, observe that the non-literal head of *bahuvrihi* is *vrihi* (lit. rice), and the non-literal heads of the preceding English bahuvrihi compounds are *face*, *brain*, and *collar*, respectively.

A.2.1.2 Synthetic

In general, synthetic compounds contain a constituent that is derived from a verb such that the compound meaning can be obtained by treating the other constituent as the argument of the verb. English synthetic compounds are often endocentric, such as *bus driver*, *city employee*, or *pan-fried*. In this case, the traditional semantic head contains both a verb and a morpheme that indicates the word class of the intended concept. In contrast, the verb-based constituent in exocentric synthetic compounds does not contain any overt marking of the intended word class. For example, the French compound *gratte-ciel* (lit. scratch-sky) refers to skyscrapers, and the Japanese compound *tsume-kiri* (lit. nail-cut) refers to nail clippers; both compounds are nouns but their overall word class is not indicated by the verb-based constituents. Nonetheless, observe that if we view these synthetic compounds as endocentric by extending the meaning of their

verb constituents (e.g., by treating *kiri* as something that cuts), then we would be able to identify their non-literal heads via the first strategy. For example, observe that the non-literal heads of *gratte-ciel* and *tsume-kiri* are *gratte* and *kiri*, respectively.

A.2.1.3 Transpositional

Bauer (2010) defines transpositional compounds as cases in which the word class of the compound is not overt in either of the constituents. Based on the examples he provides, a large portion of this subclass appears to be coordinative compounds where both constituents have the same class that deviates from the class of the compound. In his subordinative examples, it appears that the literal meaning of the compound tends to capture an attribute of the intended referent: the Swahili compound *ujauzito* (lit. come-heavy) refers to pregnancy but the literal meaning refers to the acquisition of the fetus's weight; the Mandarin compound *zhuǎn-yǎn* (lit. turn-eye) refers to instantly but the literal meaning describes the time taken to complete an action. Similar to bahuvrihis, the meanings of these transpositional compounds can be obtained from a literal interpretation of both constituents and a further metonymic extension, which would allow us to apply the second strategy for identifying non-literal heads (e.g., the head of *zhuǎn-yǎn* is *zhuǎn*). However, it is unclear if all transpositional compounds are instances of metonymy, and the generality of this pattern is left for future work.

A.2.1.4 Metaphorical

According to Bauer (2010) and later Bauer (2017), a compound is metaphorical if it contains "a metaphor in their non-modifying element" or if the whole compound has a metaphorical interpretation. Observe that we may identify the non-literal head in each of these two cases by appropriately applying the two strategies. For example, *firedog* refers to "iron support for burning logs", where *dog* may be seen as a metaphor for a thing that guards over the fire like a watchdog (Benczes, 2006, p. 175); the non-literal

head is *dog* if we apply the first strategy. In contrast, *dust bowl* refers to "an area with no vegetation" and may be seen as a metaphor in which the source domain is a literal bowl full of dust; the non-literal head is *bowl* if we apply the second strategy.

A.2.2 Coordinative compounds

Coordinative compounds constitute a subclass of word combinations in which both constituents "are on an equal footing, not one which is subordinate to the other" (Bauer, 2017, p. 83). According to Bauer (2010), most coordinative compounds are exocentric, and the main exception is a subclass called appositional compounds, in which the compound denotes an intersection of the constituents (e.g., *hunter-gatherer*, *singer-songwriter*). These compounds are endocentric because both constituents are hypernyms of the compound and may be seen as semantic heads in the traditional sense.[2]

The most prototypical coordinative compounds across languages express superordinate categories relative to their constituents (Wälchli, 2005). For example, the Georgian compound *da-dzma* (lit. sister-brother) denotes the union of the constituent categories, and the Tagalog compound *araw-gabi* (lit. day-night) denotes "day and night" or "all the time", which is a generalization of the constituents. There does not seem to exist an intuitive way to apply the standard, hypernymy-based notion of semantic heads to these prototypical compounds because their constituents are hyponyms, but both constituents nonetheless contribute to the meaning of the compound equally by definition. To facilitate the identification of heads, I define the non-literal head of a non-appositional coordinative compound as the constituent in the general head position of endocentric compounds from the same language.

[2]For simplicity, I assume that the head of an appositional compound is the constituent that shares the same position with the heads of most endocentric compounds. For example, this implies that *songwriter* is the head of *singer-songwriter* since English compounds are generally right-headed (Lieber, 2011).

A.2.3 Discussion

In this section, I defined the notion of non-literal heads for various types of exocentric compounds. Similar to the notion of semantic heads proposed by Scalise and Fábregas (2010), non-literal heads may be seen as an extension of traditional semantic heads, which are relatively well-defined for endocentric compounds. This is the case because the traditional notion is directly incorporated into the definition of non-literal heads for both subordinative and coordinative compounds.

A natural question to ask is whether non-literal heads are well-defined for every exocentric compound. This involves at least two considerations: 1) the head should be always identifiable; and 2) the head should be unique. The first consideration is in part addressed by the fact that the definition covers a wide range of exocentric compounds discussed in the literature, but it is unclear if these compound types are exhaustive. Moreover, since transpositional compounds were not as well specified as the other subclasses, future work is required to fully specify how non-literal heads are defined for these compounds. The second consideration is relevant when there exist different ways to derive the meaning of the same exocentric compound (e.g., different figurative interpretations), and a problem would occur if they lead to different assignments of head and modifier when applying the strategies in Section A.2.1. It is possible to resolve this problem by prioritizing certain interpretations over others during head identification, but an investigation of the extant of this problem and its potential solutions is left for future work.

Appendix B

Supplementary information for Chapter 3

B.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.

B.1.1 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 $(C_1, W_1), (C_2, W_2), ..., (C_n, W_n)$.

B.1.1.1 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:

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

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.

B.1.1.2 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 Regier et al. (2015), 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:

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

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 B.1. We thus define average information loss as the expectation of KL divergence (which reduces to expected surprisal).

B.1.1.3 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:

$$L_\beta[E^*|\mathcal{L}] = \mathbb{E}[D(M||\hat{M})|\mathcal{L}',\mathcal{L}] + \beta\mathbb{E}[l(W)|\mathcal{L}'] \quad \text{(B.3)}$$

$$= \sum_{(c,w)\in\mathcal{L}'} p(c,w|\mathcal{L}') \cdot (h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w)) \quad \text{(B.4)}$$

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

$$\propto \sum_{(c,w)\in E^*} p(c,w|\mathcal{L}') \cdot (h(\hat{m}_{w,\mathcal{L}}(c)) + \beta l(w)) \quad \text{(B.6)}$$

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.1.2 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., Regier et al., 2015; Zaslavsky et al., 2018) 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, Y. Xu et al. (2016) 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 attested combinations of existing words or combinations involving subwords in general. Our discussion will focus on how subwords relate to expected word length and informativeness under existing frameworks.

Word length. We consider the two-stage approach in Mollica et al. (2021) 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., Chou, Lookabaugh, & Gray, 1989). 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 (Cover & Thomas, 2006), 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 (Pimentel et al., 2021). 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 (Mollica et al., 2021; Pimentel et al., 2021). This implicitly assumes that word forms are uniquely decodable, which does not hold over the whole lexicon in general (Ferrer-i Cancho et al., 2022). 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; Marchand, 1969) but not others (e.g., French; Van Goethem & Amiot, 2019).

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 (Regier et al., 2015; Y. Xu et al., 2016; Zaslavsky et al., 2018; Y. Xu, Liu, & Regier, 2020; Zaslavsky et al., 2021; Mollica et al., 2021; Chen et al., 2023; Aggarwal,

Watson, Senapati, & Stevenson, 2024). 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 Kemp and Regier (2012), 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 (Zaslavsky et al., 2018). 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., Zaslavsky et al., 2018). Nonetheless, experimental studies on language evolution (e.g., Raviv et al., 2019b, 2019a) and language production (e.g., Downing, 1977; Pugacheva & Günther, 2024) 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.

B.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.

B.2.1 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 (Fellbaum, 1998); 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^*.

B.2.1.1 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 (Bird, Klein, & Loper, 2009).

B.2.1.2 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; Davies, 2002). 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 (Bevilacqua & Navigli,

2020). In Appendix, Section B.2, 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; Michel et al., 2011). 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.

B.2.1.3 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; Gagné et al., 2019) or if it is etymologically formed via compounding in an extraction of Wiktionary (W. Wu & Yarowsky, 2020); 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; Kay et al., 2017) 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.

B.2.1.4 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 (Cook et al., 2014). Each sense definition is processed by removing non-alphabetical characters, converting to lowercase, and lemmatizing its tokens using the WordNet lemmatizer from NLTK (Bird et al., 2009). 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 Cook et al. (2014) focused on senses emerging after year 2000, we selected a new set of keywords that are culturally salient during the past century. We selected

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 B.1: Set of cultural keywords

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 Ryskina et al. (2020) 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 B.1.

Finally, we used the first citations of English compounds to help supplement our dataset of English reuse items. For each form-sense pair in our dataset of attested English 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

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 B.2: Number of existing forms and senses, and novel items across intervals for English

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.

B.2.1.5 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 B.2. Note that the smaller sample sizes of the final interval are caused by the sparsity of post-1980 entries in the HTE.

B.2.2 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.2.2.1 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 French or Finnish lemma forms of every concept in the English existing 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; Michel et al., 2011) for French and the Newspaper and Periodical Corpus of the National Library of Finland (FNC; National Library of Finland, 2014) for Finnish. The descriptive statistics for each existing lexicon are summarized in Tables B.3 and B.4. Note that fluctuations in the number of existing words in Table B.4 are due to sparsity of the Finnish corpus for certain historical intervals.

B.2.2.2 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 (W. Wu & Yarowsky, 2020), 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

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 B.3: 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 B.4: Number of existing forms and senses, and novel items across intervals for Finnish

the combination-based encoding E^* if their constituents do not contain stopwords. The descriptive statistics for each attested encoding are summarized in Tables B.3 and B.4. Note that the smaller number of novel items in the last interval is due to sparsity in the English data.

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.

B.2.3 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 (Bevilacqua & Navigli, 2020), 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 (Gulordava & Baroni, 2011). 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 (Cook et al., 2014; Hu et al., 2019) 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:

$$N(s_i) = \frac{p_m(s_i) + \alpha}{p_h(s_i) + \alpha} \tag{B.7}$$

Here α is a constant that prevents division by zero, which we set as 0.01 following Hu et al. (2019). 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 Gulordava and Baroni (2011), and we computed change scores for the lemmatized entries.

In Table B.5, 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) Frermann and Lapata (2016) used a collection of historical corpora (Davies, 2002; Diller, De Smet, & Tyrkkö, 2011; Popescu & Strapparava, 2015) and measured change following Cook et al. (2014) with a state-of-the-art dynamic Bayesian model; 2) Hu et al. (2019) used the same data and methods described above, except word senses were obtained from Oxford Dictionary[1] and disambiguation was based on a nearest-neighbour approach; and 3) Giulianelli, Del Tredici, and

[1] `https://www.lexico.com/`

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 B.5: Correlations between semantic change scores based on sense frequencies and human ratings of change

Fernández (2020) and Montariol, Martinc, and Pivovarova (2021) used COHA and word senses inferred from clustering BERT-based contextualized word embeddings (Devlin, Chang, Lee, & Toutanova, 2019). Our method yielded Pearson ($p < 0.001$, $n = 100$) and spearman ($p < 0.001$, $n = 100$) 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.

B.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.

B.3.1 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.

B.3.1.1 Datasets

We used three datasets of pairwise word similarity ratings: WordSim-353 (Finkelstein et al., 2002), MEN (Bruni et al., 2014), and SimLex-999 (Hill et al., 2015). 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 (Agirre et al., 2009). 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 (Fellbaum, 1998). 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 (Gagné et al., 2019) and the online version of Oxford Dictionary[2]. Second, we obtained $43,482$ COHA lemmas and their definitions that appear in archived webpages of Oxford Dictionary[3]. In total, this provided us with a set of

[2] `https://www.lexico.com/`
[3] `https://archive.org`

77,359 senses for 30,760 words.

B.3.1.2 Methods

In the main text, we mapped each word sense c to a vector by embedding the definition of c with Sentence-BERT (Reimers & Gurevych, 2019). 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 (Mikolov, Chen, et al., 2013; Mikolov et al., 2018).

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 (Mikolov, Chen, et al., 2013; Mikolov et al., 2018) 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

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 B.6: 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

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.

B.3.1.3 Results

We summarize our results for the similarity subset of WordSim-353 in Table B.6, for the relatedness subset of WordSim-353 in Table B.7, for MEN in Table B.8, and for SimLex-999 in Table B.9. 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.

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	**−0.699**	< 0.001	**−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)	**−0.509**	< 0.001	**−0.493**	< 0.001	–

Table B.7: Evaluating semantic spaces and word-level prototypes using word similarity ratings (WordSim-353 rel); abbreviations follow Table B.6

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	**−0.819**	< 0.001	**−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)	**−0.666**	< 0.001	**−0.683**	< 0.001	–

Table B.8: Evaluating semantic spaces and word-level prototypes using word similarity ratings (MEN); abbreviations follow Table B.6

Method	Pearson ρ	p-value	Spearman ρ	p-value	N
Word-level W2V	**−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)	**−0.438**	< 0.001	**−0.426**	< 0.001	–

Table B.9: Evaluating semantic spaces and word-level prototypes using word similarity ratings (SimLex-999); abbreviations follow Table B.6

B.3.2 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.3.2.1 Compound dataset

The Large Database of English Compounds (LADEC; Gagné et al., 2019) contains 8,957 English closed compounds compiled from the Brown, CELEX and COCA corpora and a dataset of phrases from Costello, Veale, and Dunne (2006). 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.3.2.2 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 (Mitchell & Lapata, 2010). 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:

$$q_{w,\mathcal{L}} = Aq_{x,\mathcal{L}} + Bq_{y,\mathcal{L}} \tag{B.8}$$

where $q_{w,\mathcal{L}}$ 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 B.8 where the matrices are replaced by scalar weights (Mitchell & Lapata, 2010). 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 B.8

by using a version of linear projection (Yazdani et al., 2015; Guevara, 2010; Marelli et al., 2017). 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 (Kingma & Ba, 2015) 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.3.2.3 Results

The correlations based on the best simple prototypes are shown in Table B.10. 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 B.11, 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 B.10: 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 B.11: Evaluating semantic spaces and simple prototypes using meaning predictability ratings; abbreviations follow Table B.6

B.3.3 Sensitivity parameter

In the main text, we combined sense and word-level representations with the similarity choice model (Luce, 1963; Nosofsky, 1986) 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 framework 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 3.5:

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

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 (Davies, 2002) and token frequencies from the Google Ngrams corpus (Michel et al., 2011) and the FNC (National Library of Finland, 2014), 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 B.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 B.1. 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 frequent senses, which are close to frequency-based prototypes by construction. 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

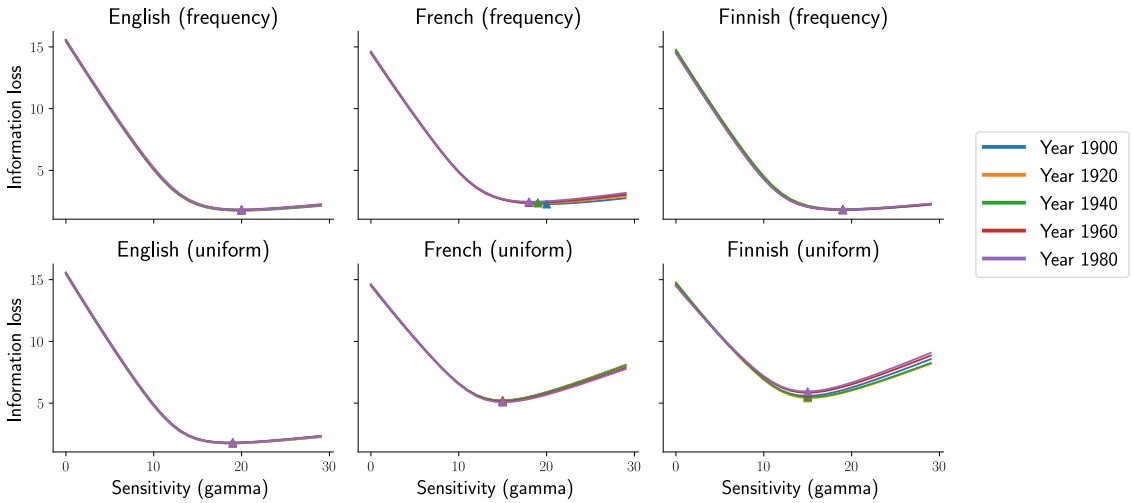

Figure B.1: Information loss incurred over communicating existing concepts. Triangles indicate the lowest information loss for a specific language and interval. Legend on the right indicates the starting year of the interval.

of monosemous words = 61.1%) and Finnish lexicons (average percentage of monosemous words = 50.8%).

B.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.

B.4.1 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

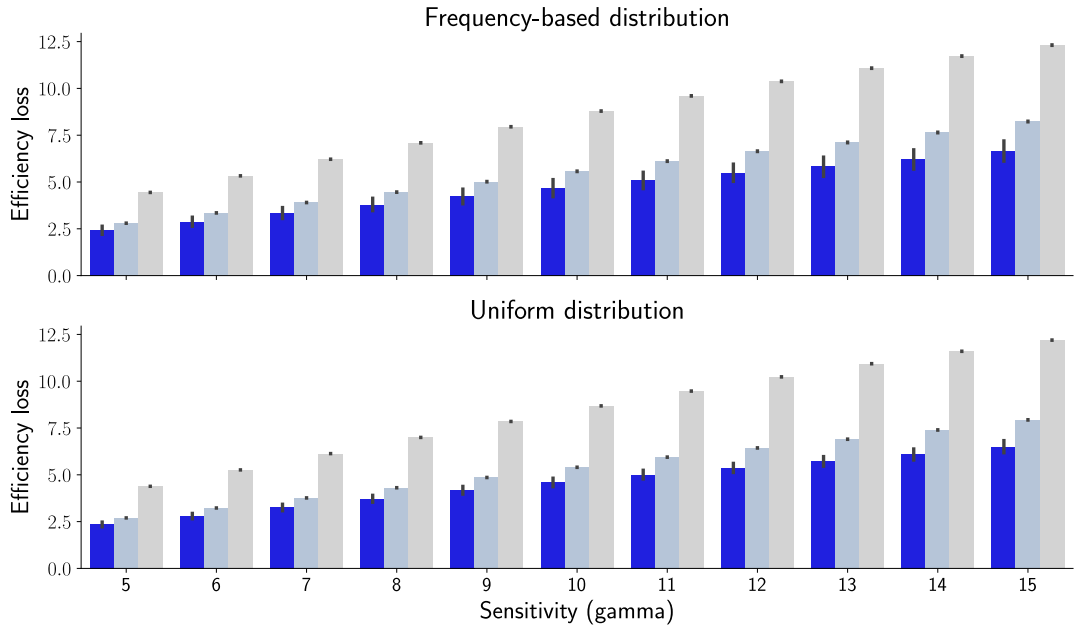

Figure B.2: Efficiency loss of attested reuse items and compounds relative to the average loss of baselines. As in the main text, blue bars correspond to attested items, light blue bars correspond to near-synonym baselines, and grey bars correspond to random baselines. Error bars show bootstrapped 95% confidence intervals.

over communicative need and speaker's 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 B.2 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.

B.4.2 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 (Aro, 2017).

B.4.2.1 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 (Lenzo, 2014) 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; New, Pallier, Brysbaert, & Ferrand, 2004). 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 (Labov, 2011). We leave more fine-grained phonological representations and their application to an efficiency-based analysis of reuse and combination for future work.

B.4.2.2 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 (Vaswani et al., 2017; S. Wu, Cotterell, & Hulden, 2021). 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 S. Wu et al. (2021) for the same dataset and the high French accuracy is likely due to a large number of inflected 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.4.2.3 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

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 B.12: 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 B.13: Number of existing forms and senses, and novel items for French after mapping orthographic forms to phonemic forms

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 B.12 and B.13.

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 B.3. We

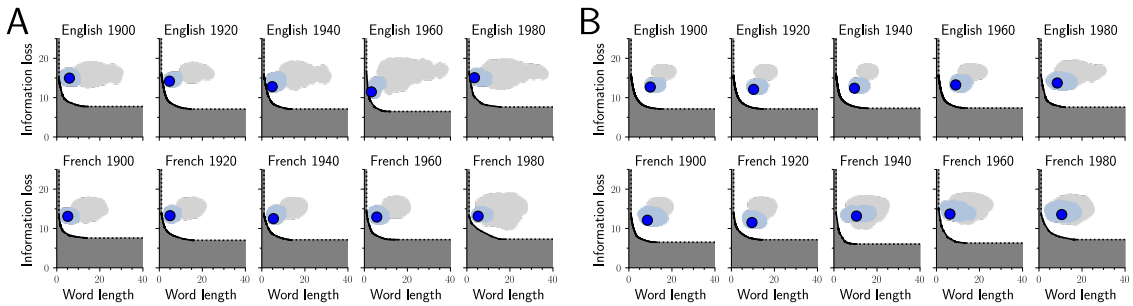

Figure B.3: Illustration comparing (A) attested reuse items and (B) attested compounds to the constructed baselines and the Pareto frontier using an implementation based on phonological representations of word forms.

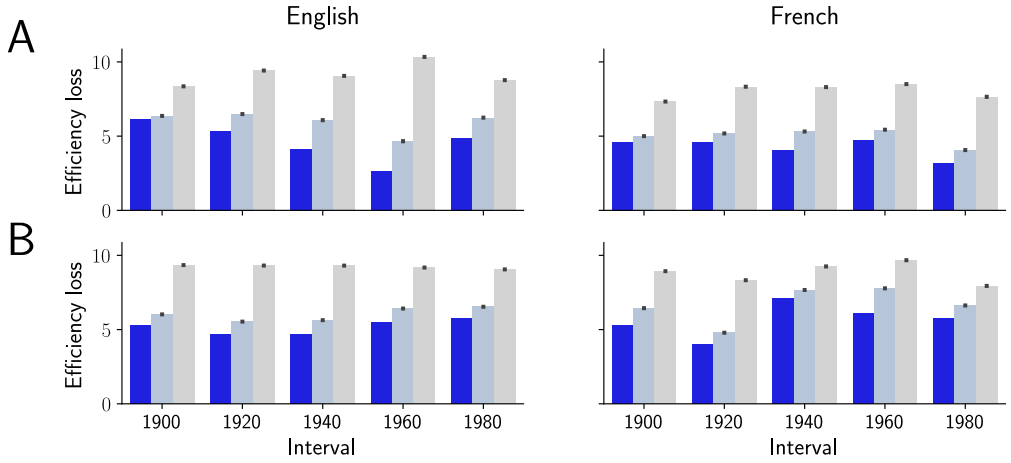

Figure B.4: Efficiency loss of (A) attested reuse items and (B) attested compounds based on phonological representations of word forms relative to the average loss of baselines. Error bars show bootstrapped 95% confidence intervals.

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 B.4 which confirms our qualitative observations. These results show that our tradeoff proposal is robust to different channels of communication.

B.4.3 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 (Reimers & Gurevych,

2019). Here we relax this assumption by measuring conceptual similarity with historical word embeddings (Mikolov, Chen, et al., 2013; Hamilton et al., 2016), and we analyze a subset of the English data used in the main text under this alternative implementation.

B.4.3.1 Methods

We used pre-trained English historical word embeddings provided by HistWords (Hamilton et al., 2016). Specifically, we used the version that was based on the Google Ngrams corpus (All English version 2; Michel et al., 2011), 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 (Mikolov, Chen, et al., 2013). 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 Appendix, Section B.3 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 Appendix, Section B.3. 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

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)	**−0.761**	< 0.001	**−0.738**	< 0.001	–

Table B.14: 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	**−0.604**	< 0.001	**−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)	**−0.489**	< 0.001	**−0.474**	< 0.001	–

Table B.15: Evaluating semantic spaces using word similarity ratings (WordSim-353 rel); abbreviations follow Table B.14

approaches: 1) sense representations based on sentence-BERT (Reimers & Gurevych, 2019), 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 B.14, for the relatedness subset of WordSim-353 in Table B.15, for MEN in Table B.16, for SimLex-999 in Table B.17, and for LADEC compound meaning predictability ratings in Table B.18. 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	**−0.685**	< 0.001	**−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)	**−0.671**	< 0.001	**−0.686**	< 0.001	–

Table B.16: Evaluating semantic spaces using word similarity ratings (MEN); abbreviations follow Table B.14

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)	**−0.441**	< 0.001	**−0.424**	< 0.001	–

Table B.17: Evaluating semantic spaces using word similarity ratings (SimLex-999); abbreviations follow Table B.14

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)	**−0.417**	< 0.001	**−0.415**	< 0.001	–

Table B.18: Evaluating semantic spaces using compound meaning predictability ratings; abbreviations follow Table B.14

B.4.3.2 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 every word sense in the existing lexicon $\mathcal{L}$ and attested encoding E^* if any non-stopword word in its definition does not exist in the vocabulary of the embeddings from t_0. The descriptive statistics of our final datasets are summarized in Table B.19.

As in the main text, we assessed the efficiency of an attested encoding by measuring

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 B.19: Number of existing forms and senses, and novel items across intervals after intersecting with HistWords

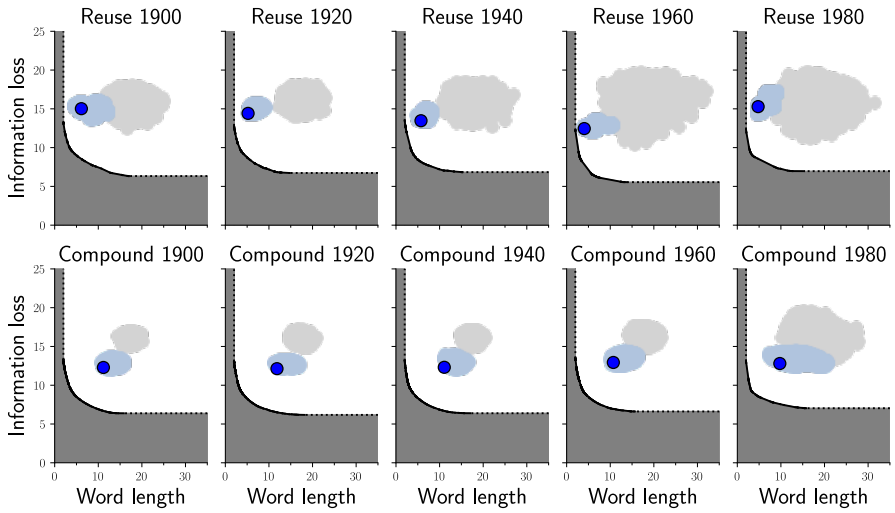

Figure B.5: Illustration comparing attested reuse items and compounds to constructed baselines and Pareto frontiers using an implementation based on HistWords.

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 B.5. 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 B.6 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 cannot be explained by potential systematic biases in using contemporary embeddings.

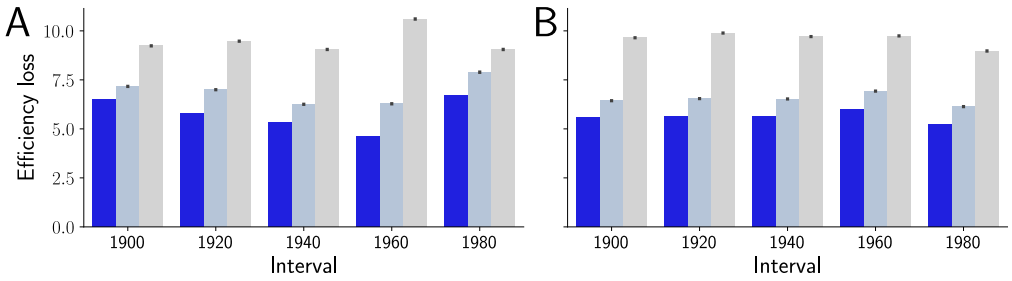

Figure B.6: Efficiency loss of (A) attested reuse items and (B) attested compounds relative to the average loss of baselines, under an implementation of our framework using HistWords. Error bars show bootstrapped 95% confidence intervals.

B.4.4 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 framework based on Oxford Dictionary. We then describe an implementation that assumes one-to-one correspondence between concept and form.

B.4.4.1 Oxford dictionary

In Appendix, Section B.3, 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 (Kay et al., 2017). To obtain novel reused items, we used the same keyword matching method (see Appendix, Section B.2) 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.

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 B.20: Number of existing forms and senses, and novel items across intervals for the Oxford Dictionary-based analysis

To obtain novel compounds, we focused on closed compounds since we only retrieved unigrams from OD and we identified these compounds using LADEC (Gagné et al., 2019) and Wiktionary (W. Wu & Yarowsky, 2020). 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 Hu et al. (2019) 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 B.20.

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 (Michel et al., 2011) and add-one smoothing as in the main text, but sense frequencies were estimated using the dataset from Hu et al. (2019). 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

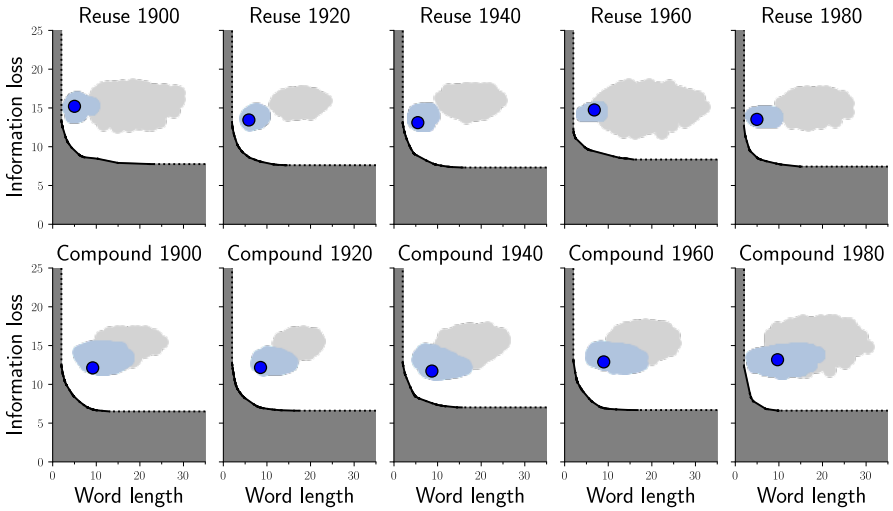

Figure B.7: Illustration comparing attested reuse items and compounds to constructed baselines and Pareto frontiers under an Oxford Dictionary-based implementation.

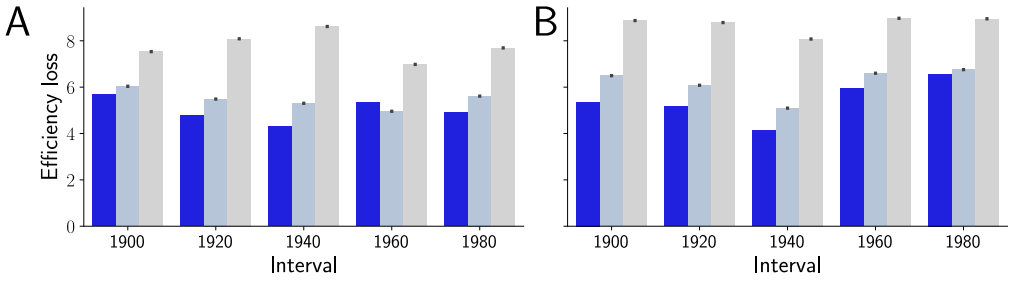

Figure B.8: Efficiency loss of (A) attested reuse items and (B) attested compounds relative to the average loss of baselines, under an implementation of our framework using the Oxford Dictionary. Error bars show bootstrapped 95% confidence intervals.

via our greedy method that scanned over $\beta = 0, 0.01, ..., 10$.

We compare these English reuse items and closed compounds to baselines in Figure B.7. 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 B.8 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.

B.4.4.2 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 analyses 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 (Kay et al., 2017) 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 (Gagné et al., 2019) and Wiktionary (W. Wu & Yarowsky, 2020) and checking their first citation in the HTE (Kay et al., 2017); 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., Mikolov, Chen, et al., 2013). Here, we used the same pre-trained word2vec embeddings (Mikolov, Chen, et al., 2013; Mikolov et al., 2018) that we validated against human ratings of English word similarity and compound meaning predictability in a pre-

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 B.21: Number of existing forms and novel compounds across intervals for the word-level analysis

vious 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 B.21.

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 (Michel et al., 2011) and add-one smoothing. We used the same settings for the listener distribution except we used a single pre-trained word2vec embedding (Mikolov et al., 2018) 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 Appendix, Section B.3. The baselines were created in the same way as in the main text. We estimated Pareto frontiers using the same method that scanned over $\beta = 0, 0.01, ..., 10$.

Figure B.9 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 B.10 shows the quantitative efficiency loss that intuitively corresponds to how far each encoding is away from the Pareto frontier. Overall, we observe that

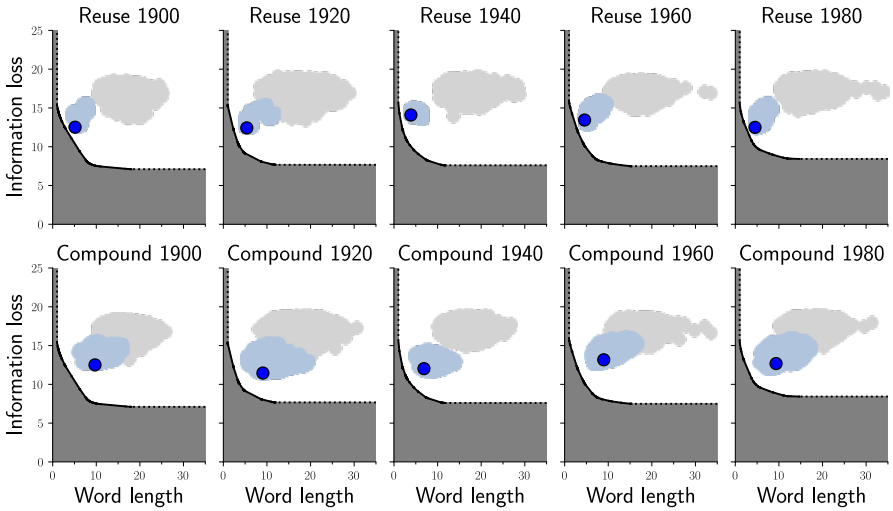

Figure B.9: Illustration comparing attested reuse items and compounds to constructed baselines and Pareto frontiers under the one-to-one assumption.

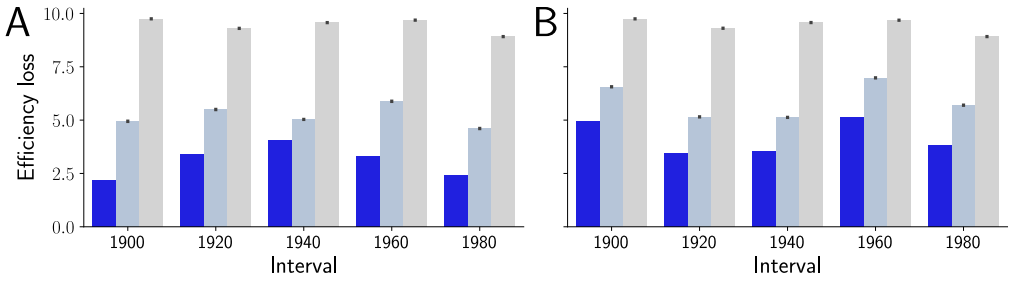

Figure B.10: Efficiency loss of (A) approximate reuse items based on head words and (B) attested compounds relative to the average loss of baselines, under an implementation of our framework using word embeddings. Error bars show bootstrapped 95% confidence intervals.

attested items are 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).

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 B.22: Correlations between orthography-based item-level efficiency and phoneme-based item-level efficiency

B.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.

B.5.1 Comparing orthographic and phonemic forms

Here we examine the item-level variation in efficiency based on phonemic forms. We reused our implementation in Appendix, Section B.4 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 B.22. Overall, we did not

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 B.23: Examples of literal items and their hypernyms in existing lexicons

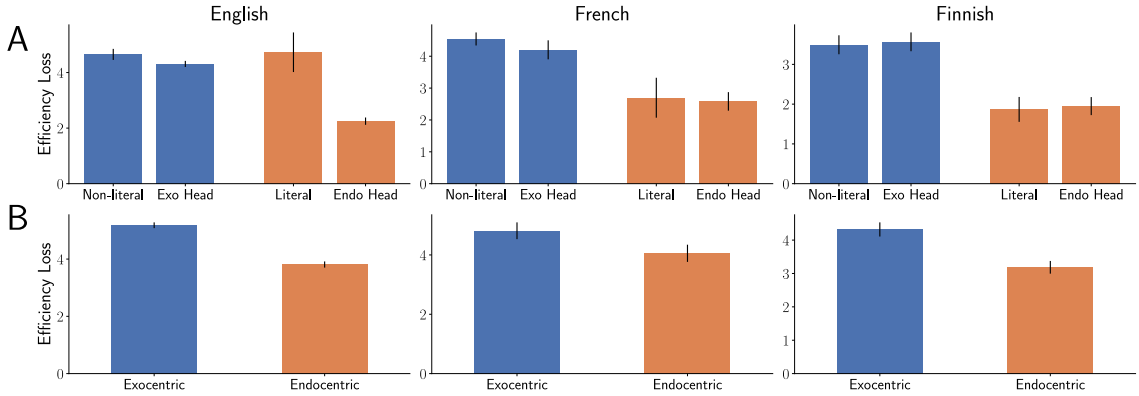

Figure B.11: Averages of item-level efficiency loss for literal and non-literal (A) reuse items and (B) compounds. For each plot in row (A), the left shows attested non-literal reuse items and additional non-literal reuse items based on exocentric compounds; the right shows attested literal reuse items and additional literal reuse items based on endocentric compounds. Error bars indicate bootstrapped 95% confidence intervals.

find significant differences in item-level variation between the representations, and thus we focused on analyzing item-level variation using orthographic forms.

B.5.2 Examples of literal items

Table B.23 shows examples of literal items from the main text.

B.5.3 Efficiency in literal and non-literal items

Figure B.11 illustrates the comparisons between literal and non-literal items presented in the main text.

B.5.4 Strategy comparison

We compare distributions of item-level costs between reuse and compounding, separately for the languages examined in the main text. Figure B.12 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 ($t(3346) = 13.046$, $p < 0.001$) and lower word length ($t(3346) = -34.908$, $p < 0.001$). Distributions of item-level costs for French and Finnish are summarized in Figures B.13 and B.14. 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 ($t(936) = -31.907$, $p < 0.001$). 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 ($t(1154) = 0.0202$, $p = 0.984$).

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.

B.5.5 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 be-

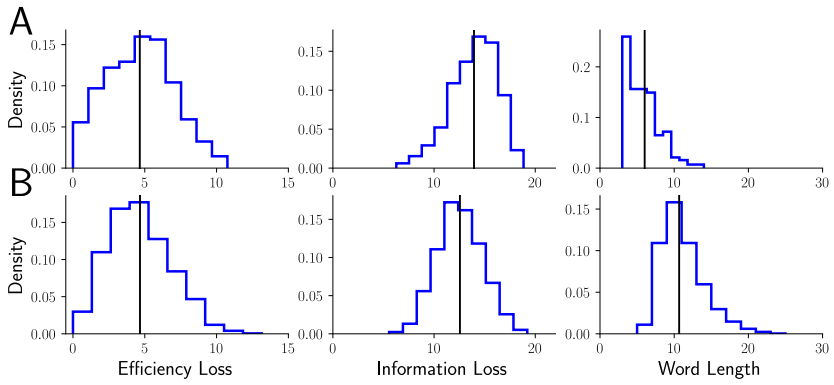

Figure B.12: Distributions of item-level efficiency loss, information loss, and word length for English (A) reuse items and (B) compounds. In each plot, the vertical line indicates the mean of the distribution.

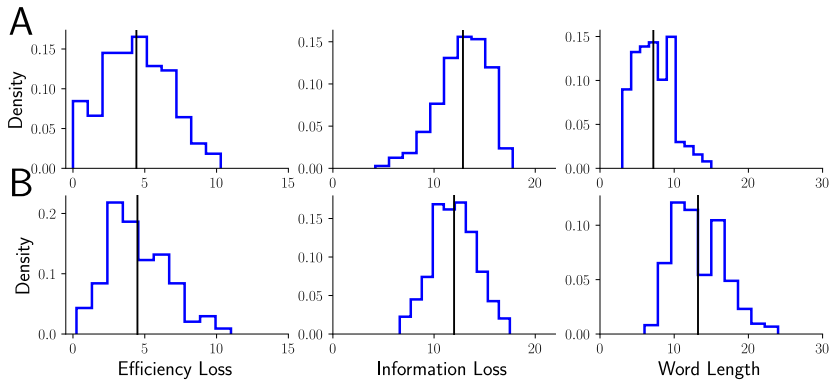

Figure B.13: Distributions of item-level efficiency loss, information loss, and word length for French (A) reuse items and (B) compounds. In each plot, the vertical line indicates the mean of the distribution.

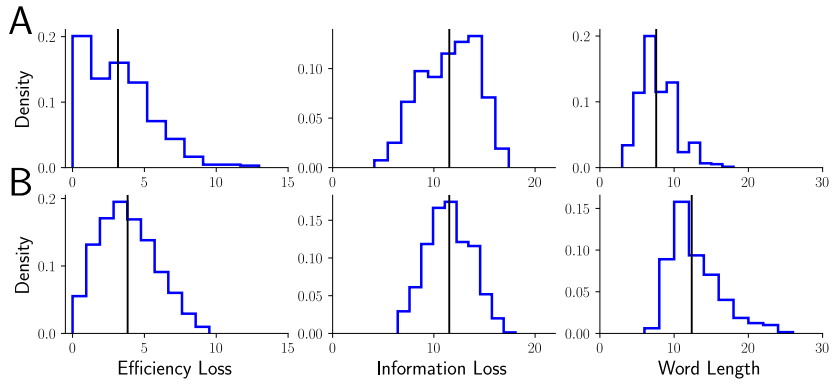

Figure B.14: Distributions of item-level efficiency loss, information loss, and word length for Finnish (A) reuse items and (B) compounds. In each plot, the vertical line indicates the mean of the distribution.

cause the original Princeton WordNet avoided linking a sense and its hyponyms to the same word (Miller, 1998), and 2) hyponymic relations can be taxonomic but also functional (Wierzbicka, 1984). 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 (Leacock et al., 1998) and Wu-Palmer similarity (Z. Wu & Palmer, 1994). 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:

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

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 B.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 B.24, and summarize analogous results based on Wu-Palmer similarity in Table B.25. In all cases, we observe

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 B.24: 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 B.25: Correlations between Wu-Palmer similarity and item-level efficiency loss

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.

B.5.6 Item frequency

Usage frequency is often correlated with the economy of use of a word form (e.g., Zipf, 1949; Mahowald et al., 2018). 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 B.26 and the results for compounds are summarized in Table B.27. Contrary to our prediction, we do not observe any significant correlation between item-

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 B.26: 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 B.27: Correlations between the relative frequency of compounds and item-level efficiency

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.

References

Aggarwal, J., Watson, J., Senapati, P., & Stevenson, S. (2024). The (in) efficiency of within-language variation in online communities. In L. Samuelson, S. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 46).

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009, June). A study on similarity and relatedness using distributional and WordNet-based approaches. In M. Ostendorf, M. Collins, S. Narayanan, D. W. Oard, & L. Vanderwende (Eds.), *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27). Boulder, Colorado: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N09-1003`

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

Algeo, J., & Algeo, A. S. (1991). *Fifty years among the new words: A dictionary of neologisms 1941-1991*. Cambridge University Press.

Allen, M. R. (1978). *Morphological investigations*. University of Connecticut.

Antoniová, V. K. (2020). An onomasiological approach to nominal compound semantics. *Word Structure*, *13*(3), 316–346.

Arnaud, P. J. (2015). Noun-noun compounds in French. In C. Iacobini, P. Müller, I. Ohnheiser, S. Olsen, & F. Rainer (Eds.), *Word-formation: An international*

handbook of the languages of Europe, vol. 1 (pp. 673–686). de Gruyter Mouton.

Arndt-Lappe, S. (2011). Towards an exemplar-based model of stress in English noun–noun compounds. *Journal of Linguistics*, *47*(3), 549–585.

Arndt-Lappe, S. (2014). Analogy in suffix rivalry: The case of English-ity and-ness. *English Language & Linguistics*, *18*(3), 497–548.

Arndt-Lappe, S. (2015). Word-formation and analogy. In C. Iacobini, P. Müller, I. Ohnheiser, S. Olsen, & F. Rainer (Eds.), *Word-formation : An international handbook of the languages of Europe* (Vol. 2, pp. 822–841). De Gruyter.

Arndt-Lappe, S. (2023). Different lexicons make different rivals. *Word Structure*, *16*(1), 24–48.

Aro, M. (2017). Learning to read Finnish. In L. Verhoeven & C. Perfetti (Eds.), *Learning to read across languages and writing systems* (pp. 416–436). Cambridge University Press.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database.* University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. *The crosslinguistic study of sentence processing*, *3*, 73–112.

Bauer, L. (1998). Is there a class of neoclassical compounds, and if so is it productive? *Linguistics*, *36*(3), 403–422.

Bauer, L. (2001). *Morphological productivity* (Vol. 95). Cambridge University Press.

Bauer, L. (2010). The typology of exocentric compounding. In *Cross-disciplinary issues in compounding* (pp. 167–176). John Benjamins Publishing Company.

Bauer, L. (2011). Typology of compounds. In R. Lieber & P. Štekauer (Eds.), *The Oxford handbook of compounding* (pp. 343–356). Oxford University Press.

Bauer, L. (2017). *Compounds and compounding* (Vol. 155). Cambridge University Press.

Bauer, L., & Tarasova, E. (2013). The meaning link in nominal compounds. *SKASE Journal of Theoretical Linguistics*, *10*(3).

Benczes, R. (2006). *Creative compounding in English: the semantics of metaphorical and metonymical noun-noun combinations*. John Benjamins Publishing Company.

Benczes, R. (2015). Are exocentric compounds really exocentric. *SKASE Journal of Theoretical Linguistics*, *12*(3), 54–73.

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

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

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bisetto, A., & Scalise, S. (2005). The classification of compounds. *Lingue e linguaggio*, *4*(2), 319–0.

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

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

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

Booij, G. (2010). Compound construction: Schemas or analogy? A construction morphology perspective. *Cross-disciplinary issues in compounding*, 93–108.

Booij, G. (2019). Compounds and multi-word expressions in Dutch. In B. Schlücker

(Ed.), *Complex lexical units* (pp. 95–126). De Gruyter.

Booij, G., & Hüning, M. (2014). Affixoids and constructional idioms. *Extending the scope of Construction Grammar*, 77–105.

Borer, H. (2011). Afro-Asiatic, Semitic: Hebrew. In *The Oxford handbook of compounding* (pp. 491–511). Oxford University Press. Retrieved from `https://doi.org/10.1093/oxfordhb/9780199695720.013.0027` doi: 10.1093/oxfordhb/9780199695720.013.0027

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

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

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

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of artificial intelligence research*, *49*, 1–47.

Brusnighan, S. M., & Folk, J. R. (2012). Combining contextual and morphemic cues is beneficial during incidental vocabulary acquisition: Semantic transparency in novel compound word processing. *Reading Research Quarterly*, *47*(2), 172–190.

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

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

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

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, *202*, 104289.

Carstensen, A., Xu, J., Smith, C. T., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. In R. Dale et al. (Eds.), *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 37).

Ceccagno, A., & Basciano, B. (2007). Compound headedness in Chinese: An analysis of neologisms. *Morphology*, *17*, 207–231.

Chapman, D., & Skousen, R. (2005). Analogical modeling and morphological change: The case of the adjectival negative prefix in English. *English Language & Linguistics*, *9*(2), 333–357.

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

Chou, P. A., Lookabaugh, T., & Gray, R. M. (1989). Entropy-constrained vector quantization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(1), 31–42.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509.

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

Cook, P., Lau, J. H., McCarthy, D., & Baldwin, T. (2014). Novel word-sense identification. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 1624–1635).

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. Retrieved from https://aclanthology.org/P06-2021

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience.

Daelemans, W., Berck, P., & Gillis, S. (1997). Data mining as a method for linguistic analysis: Dutch diminutives. *Folia Linguistica*, *31*, 57–75.

Daelemans, W., Gillis, S., & Durieux, G. (2013). Skousen's analogical modelling algorithm: a comparison with lazy learning. In *New methods in language processing* (pp. 3–15).

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2004). *Timbl: Tilburg memory based learner, version 5.1, reference guide* (Vol. 04-02). UVT Universiteit van Tilburg.

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

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

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M., & Baayen, R. H. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language*, *81*(1-3), 555–567.

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

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

of America.

de Saussure, F. (1916). *Course in general linguistics* (R. Harris, Trans.). London, UK: Bloomsbury Academic. (Translated edition published 2013)

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N19-1423` doi: 10.18653/v1/N19-1423

Diller, H.-J., De Smet, H., & Tyrkkö, J. (2011). A European database of descriptors of English electronic texts. *The European English Messenger*, *19*(2), 21–35.

Dima, C. (2015, September). Reverse-engineering language: A study on the semantic compositionality of German compounds. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1637–1642). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D15-1188/` doi: 10.18653/v1/D15-1188

Dima, C., de Kok, D., Witte, N., & Hinrichs, E. (2019). No word is an Island—A transformation weighting model for semantic composition. *Transactions of the Association for Computational Linguistics*, *7*, 437–451. Retrieved from `https://aclanthology.org/Q19-1025/` doi: 10.1162/tacl_a_00275

Dima, C., & Hinrichs, E. (2015, April). Automatic noun compound interpretation using deep neural networks and word embeddings. In M. Purver, M. Sadrzadeh, & M. Stone (Eds.), *Proceedings of the 11th international conference on computational semantics* (pp. 173–183). London, UK: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W15-0122/`

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

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

Eddington, D. (2002). A comparison of two analogical models: Tilburg memory-based learner versus analogical modeling. In *Analogical modeling: An exemplar-based approach to language* (pp. 141–155). John Benjamins Publishing Company.

Eddington, D. (2006). Look ma, no rules: Applying Skousen's analogical approach to Spanish nominals in-ión. In G. Wiebe, G. Libben, T. Priestly, R. Smyth, & H. Wang (Eds.), *Phonology, morphology, and the empirical imperative: Papers in honour of bruce l. derwing* (pp. 371–407). Crane Taipei.

Elzinga, D. (2006). English adjective comparison and analogy. *Lingua*, *116*(6), 757–770.

Fares, M., Oepen, S., & Velldal, E. (2018, October-November). Transfer and multi-task learning for noun–noun compound interpretation. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1488–1498). Brussels, Belgium: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D18-1178/` doi: 10.18653/v1/D18-1178

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

Ferrer-i Cancho, R., Bentz, C., & Seguin, C. (2022). Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics*, *29*(2), 165–194.

Finkbeiner, R., & Schlücker, B. (2019). Compounds and multi-word expressions in the languages of Europe. In B. Schlücker (Ed.), *Complex lexical units* (pp. 1–43). De Gruyter.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131.

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

Frermann, L., & Lapata, M. (2016). A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, *4*, 31–45.

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

Futrell, R., & Hahn, M. (2022). Information theory as a bridge between language function and language form. *Frontiers in Communication*, *7*. Retrieved from `https://www.frontiersin.org/articles/10.3389/fcomm.2022.657725/full` doi: 10.3389/fcomm.2022.657725

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 71.

Gagné, C. L., Spalding, T. L., & Schmidtke, D. (2019). LADEC: The large database of English compounds. *Behavior Research Methods*, *51*(5), 2152–2179.

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

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

Girju, R. (2007, June). Improving the interpretation of noun phrases with cross-linguistic information. In A. Zaenen & A. van den Bosch (Eds.), *Proceedings of the 45th annual meeting of the Association for Computational Linguistics* (pp. 568–575). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P07-1072/`

Giulianelli, M., Del Tredici, M., & Fernández, R. (2020, July). Analysing lexical semantic change with contextualised word representations. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 3960–3973). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.365 doi: 10.18653/v1/2020.acl-main.365

Grewal, K., & Xu, Y. (2021). Chaining algorithms and historical adjective extension. In *Computational approaches to semantic change* (Vol. 6, p. 189-218). Berlin: Language Science Press.

Grzega, J. (2011, 07). Compounding from an onomasiological perspective. In R. Lieber & P. Štekauer (Eds.), *The Oxford handbook of compounding.* Oxford University Press. Retrieved from https://doi.org/10.1093/oxfordhb/9780199695720.013.0011 doi: 10.1093/oxfordhb/9780199695720.013.0011

Guevara, E. (2010, July). A regression model of adjective-noun compositionality in distributional semantics. In R. Basili & M. Pennacchiotti (Eds.), *Proceedings of the 2010 workshop on GEometrical models of natural language semantics* (pp. 33–37). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W10-2805

Gulordava, K., & Baroni, M. (2011, July). A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In S. Pado & Y. Peirsman (Eds.), *Proceedings of the GEMS 2011 workshop on GEometrical models of natural language semantics* (pp. 67–71). Edinburgh, UK: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W11-2508

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

Günther, F., Marelli, M., & Bölte, J. (2020). Semantic transparency effects in German compounds: A large dataset and multiple-task investigation. *Behavior Research Methods*, *52*(3), 1208–1224.

Günther, F., Smolka, E., & Marelli, M. (2019). 'Understanding' differs between English and German: Capturing systematic language differences of complex words. *Cortex*, *116*, 168–175.

Hahn, M., Degen, J., & Futrell, R. (2021). Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal. *Psychological Review*, *128*(4), 726.

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

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

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016, August). Diachronic word embeddings reveal statistical laws of semantic change. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 1489–1501). Berlin, Germany: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P16-1141` doi: 10.18653/v1/P16-1141

Hampton, J. (1991). The combination of prototype concepts. In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 91–116). Psychology Press.

Haspelmath, M. (2025). Compound and incorporation constructions as combinations of unexpandable roots. *Zeitschrift für Wortbildung / Journal of Word Formation*,

9(1), 1–23.

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2023). From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review*, *130*(4), 977.

Hendrickx, I., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Szpakowicz, S., & Veale, T. (2013, June). SemEval-2013 task 4: Free paraphrases of noun compounds. In S. Manandhar & D. Yuret (Eds.), *Second joint conference on lexical and computational semantics (*SEM), volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)* (pp. 138–143). Atlanta, Georgia, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S13-2025/

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hofmann, V., Weissweiler, L., Mortensen, D. R., Schütze, H., & Pierrehumbert, J. B. (2025). Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences*, *122*(19), e2423232122.

Hsieh, C.-Y., Marelli, M., & Rastle, K. (2025). Compositional processing in the recognition of Chinese compounds: Behavioural and computational studies. *Psychonomic Bulletin & Review*, 1–12.

Hu, R., Li, S., & Liang, S. (2019, July). Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 3899–3908). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1379 doi: 10.18653/v1/P19-1379

Huang, C.-R., Wang, H., & Chen, I.-H. (2022). Characters as basic lexical units and monosyllabicity in Chinese. In C.-R. Huang, Y.-H. Lin, I.-H. Chen, & Y.-Y. Hsu (Eds.), *The Cambridge handbook of Chinese linguistics* (pp. 74–96). Cambridge University Press.

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

Jackendoff, R. (2010). The ecology of English noun-noun compounds. In *Meaning and the lexicon* (pp. 413–451). Oxford University Press Oxford.

Jackendoff, R. (2016). English noun-noun compounds in conceptual semantics. In P. ten Hacken (Ed.), *The semantics of compounding* (pp. 15–37). Cambridge University Press.

Jespersen, O. (1942). *A modern English grammar on historical principles Part VI, morphology*. Routledge.

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

Karjus, A., Blythe, R. A., Kirby, S., Wang, T., & Smith, K. (2021). Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, *45*(9), e13035.

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., Gaby, A., & Regier, T. (2019). Season naming and the local environment. In C. Freksa, A. Goel, & C. Seifert (Eds.), *Proceedings of the 41st annual meeting of the Cognitive Science Society* (pp. 539–545).

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

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

Kiefer, F. (1990). Noun incorporation in Hungarian. *Acta Linguistica Hungarica*, *40*(1/2), 149–177.

Kim, S. N., & Baldwin, T. (2005). Automatic interpretation of noun compounds using WordNet similarity. In *Second international joint conference on natural language processing: Full papers.* Retrieved from `https://aclanthology.org/I05-1082/` doi: 10.1007/11562214_82

Kingma, D., & Ba, J. (2015, 12). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *International conference on learning representations (iclr).* San Diega, CA, USA.

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

Klégr, A., & Čermák, J. (2010). Neologisms of the 'on-the-pattern-of' type: Analogy as a word formation process. *The Prague School and Theories of Structure*, 229–241.

Koch, P. (2008). Cognitive onomasiology and lexical change: Around the eye. In *From polysemy to semantic change: Towards a typology of lexical semantic associations* (pp. 107–137). John Benjamins Publishing Company.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*(1), 113–114.

Krott, A., Baayen, R. H., & Schreuder, R. (2001). Analogy in morphology: modeling the choice of linking morphemes in Dutch. *Linguistics*, *39*(1), 51–93.

Krott, A., Schreuder, R., Baayen, R. H., & Dressler, W. U. (2007). Analogical effects on linking elements in German compound words. *Language and Cognitive Processes*, *22*(1), 25–57.

Labov, W. (2011). *Principles of linguistic change, volume 3: Cognitive and cultural*

factors (Vol. 3). John Wiley & Sons.

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

Lang, S. (2014). *Toki Pona: The language of good*. Tawhid.

Lauer, M. (1995). Designing statistical language learners: Experiments on noun compounds. *PhD Thesis, Department of Computing Macquarie University*.

Lazaridou, A., Vecchi, E. M., & Baroni, M. (2013, October). Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1908–1913). Seattle, Washington, USA: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D13-1196/`

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

Le-Khac, P. H., Healy, G., & Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. *IEEE Access*, *8*, 193907–193934.

Lenzo, K. (2014). *The CMU pronouncing dictionary (version 0.7 b)*. Carnegie Mellon University. (Avaiable at www.speech.cs.cmu.edu/cgi-bin/cmudict/. Accessed May 26, 2024.)

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

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

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

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical reasoning on

Chinese morphological and semantic relations. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 2: Short papers)* (pp. 138–143). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/P18-2023`

Lieber, R. (1983). Argument linking and compounds in English. *Linguistic Inquiry*, *14*(2), 251–285.

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

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

Lieber, R. (2016). Compounding in the lexical semantic framework. In P. ten Hacken (Ed.), *The semantics of compounding* (p. 38–53). Cambridge University Press.

Lieber, R., & Štekauer, P. (2011a). Introduction: Status and definition of compounding. In R. Lieber & P. Štekauer (Eds.), *The Oxford handbook of compounding* (pp. 3–18). Oxford University Press.

Lieber, R., & Štekauer, P. (2011b). *The Oxford handbook of compounding*. Oxford University Press.

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

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

Liu, J., Liu, J., Chen, L., Liang, J., Xiao, Y., Xu, H., ... Xie, R. (2022). Noun compound interpretation with relation classification and paraphrasing. *IEEE Transactions on Knowledge and Data Engineering*, *35*(9), 8757–8769.

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

Maguire, P., Maguire, R., & Cater, A. W. (2010). The influence of interactional semantic patterns on the interpretation of noun–noun compounds. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, *36*(2), 288.

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

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

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

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

Marelli, M., Gagné, C. L., & Spalding, T. L. (2017). Compounding as abstract operation in semantic space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, *166*, 207–224.

Martín, F. M. d. P., Bertram, R., Häikiö, T., Schreuder, R., & Baayen, R. H. (2004). Morphological family size in a morphologically rich language: the case of Finnish compared with Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1271.

Mattiello, E., & Dressler, W. U. (2018). The morphosemantic transparency/opacity of novel English analogical compounds and compound families. *Studia Anglica Posnaniensia*, *53*(1), 67–114.

Mattiello, E., & Dressler, W. U. (2022). Dualism and superposition in the analysis of English synthetic compounds ending in-er. *Linguistics*, *60*(2), 395–461.

Meylan, S. C., & Griffiths, T. L. (2021). The challenges of large-scale, web-based language datasets: Word length and predictability revisited. *Cognitive Science*, *45*(6), e12983.

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

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

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018, May). Advances in pre-training distributed word representations. In N. Calzolari et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from `https://aclanthology.org/L18-1008`

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

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

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

Mitchell, J., & Lapata, M. (2008, June). Vector-based models of semantic composition. In J. D. Moore, S. Teufel, J. Allan, & S. Furui (Eds.), *Proceedings of ACL-08: HLT* (pp. 236–244). Columbus, Ohio: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P08-1028`

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

Mithun, M. (1984). The evolution of noun incorporation. *Language*, *60*(4), 847–894.

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

Monaghan, P., & Roberts, S. G. (2019). Cognitive influences in language evolution: Psycholinguistic predictors of loan word borrowing. *Cognition*, *186*, 147–158.

Montariol, S., Martinc, M., & Pivovarova, L. (2021, June). Scalable and interpretable semantic change detection. In K. Toutanova et al. (Eds.), *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 4642–4652). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.naacl-main.369` doi: 10.18653/v1/2021.naacl-main.369

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

Nakov, P., & Hearst, M. (2006). Using verbs to characterize noun-noun relations. In *Artificial intelligence: Methodology, systems, and applications: 12th international conference, AIMSA 2006, Varna, Bulgaria, September 12-15, 2006. proceedings 12* (pp. 233–244).

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

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524.

Norcliffe, E., & Majid, A. (2024). Word formation patterns in the perception domain: a typological study of cross-modal semantic associations. *Linguistic Typology*, *28*(3), 419–459.

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

Ó Séaghdha, D., & Copestake, A. (2009, March). Using lexical and relational similarity

to classify semantic relations. In A. Lascarides, C. Gardent, & J. Nivre (Eds.), *Proceedings of the 12th conference of the European Chapter of the ACL (EACL 2009)* (pp. 621–629). Athens, Greece: Association for Computational Linguistics. Retrieved from https://aclanthology.org/E09-1071/

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

Oxford University Press. (2025). *software (n.), sense 2.a.* Retrieved from https://www.oed.com/dictionary/software_n?tab=meaning_and_use (Accessed on 2025-03-01)

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

Pepper, S. (2022). Defining and typologizing binominal lexemes. In S. Pepper, F. Masini, & S. Mattiola (Eds.), *Binominal lexemes in cross-linguistic perspective: Towards a typology of complex lexemes* (pp. 23–72). De Gruyter Mouton.

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.

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

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

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and*

Brain Sciences, *13*(4), 707–727.

Pinto Jr, R. F., & Xu, Y. (2021). A computational theory of child overextension. *Cognition*, *206*, 104472.

Plag, I. (2006). Morphology in pidgins and creoles. In K. Brown (Ed.), *Encyclopedia of language & linguistics (second edition)* (Second Edition ed., p. 304-308). Oxford: Elsevier.

Plag, I., Kawaletz, L., Arndt-Lappe, S., & Lieber, R. (2023). Analogical modeling of derivational semantics: Two case studies. In S. Kotowski & I. Plag (Eds.), *The semantics of derivational morphology.* (pp. 103–141). De Gruyter.

Ponkiya, G., Patel, K., Bhattacharyya, P., & Palshikar, G. (2018, August). Treat us like the sequences we are: Prepositional paraphrasing of noun compounds using LSTM. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 1827–1836). Santa Fe, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/C18-1155/

Popescu, O., & Strapparava, C. (2015, June). SemEval 2015, task 7: Diachronic text evaluation. In P. Nakov, T. Zesch, D. Cer, & D. Jurgens (Eds.), *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 870–878). Denver, Colorado: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S15-2147 doi: 10.18653/v1/S15-2147

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

Rainer, F., & Varela, S. (1992). Compounding in Spanish. *Rivista di Linguistica*, *4*(1), 117–142.

Ralli, A. (2013). Compounding and its locus of realization: Evidence from Greek and Turkish. *Word Structure*, *6*(2), 181–200.

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the histor-

ical emergence of word senses. *Proceedings of the National Academy of Sciences*, *115*(10), 2323–2328.

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

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

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. Retrieved from `https://aclanthology.org/I11-1024`

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

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

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

Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., & Davis, M. H. (2012). Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition*, *40*, 1095–1108.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experi-*

mental Psychology: General, *104*(3), 192.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1–28.

Ryder, M. E. (1994). *Ordered chaos: The interpretation of English noun-noun compounds*. University of California Press.

Ryskina, M., Rabinovich, E., Berg-Kirkpatrick, T., Mortensen, D., & Tsvetkov, Y. (2020, January). Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In A. Ettinger, G. Jarosz, & J. Pater (Eds.), *Proceedings of the Society for Computation in Linguistics 2020* (pp. 367–376). New York, New York: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2020.scil-1.43`

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

Scalise, S., & Fábregas, A. (2010). The head in compounding. In *Cross-disciplinary issues in compounding* (pp. 109–126). John Benjamins Publishing Company.

Scalise, S., Fábregas, A., & Forza, F. (2009). Exocentricity in compounding. *Gengo Kenkyu (Journal of the Linguistic Society of Japan)*, *135*, 49–84.

Schlücker, B. (Ed.). (2019). *Complex lexical units: Compounds and multi-word expressions*. Berlin, Boston: De Gruyter.

Schlücker, B. (2019). Compounds and multi-word expressions in German. In B. Schlücker (Ed.), *Complex lexical units* (pp. 69–94). De Gruyter.

Schlücker, B. (2016). Adjective-noun compounding in parallel architecture. In P. ten Hacken (Ed.), *The semantics of compounding* (p. 178–191). Cambridge University Press.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in science conference* (Vol. 57, p. 61).

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shwartz, V., & Dagan, I. (2018, July). Paraphrase to explicate: Revealing implicit noun-compound relations. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 1200–1211). Melbourne, Australia: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P18-1111/` doi: 10.18653/v1/P18-1111

Shwartz, V., & Waterson, C. (2018, June). Olive oil is made *of* olives, baby oil is made *for* babies: Interpreting noun compounds using paraphrases in a neural model. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, volume 2 (short papers)* (pp. 218–224). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N18-2035/` doi: 10.18653/v1/N18-2035

Skousen, R. (1989). *Analogical modeling of language*. Kluwer.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science*, *12*(4), 485–527.

Smith, K. (2022). How language learning and language use create linguistic structure. *Current Directions in Psychological Science*, *31*(2), 177–186.

Smith, K., Bowerman, J., & Smith, A. D. (2025). Semantic extension in a novel communication system is facilitated by salient shared associations. *Cognition*, *261*, 106129.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012, July). Semantic compo-

sitionality through recursive matrix-vector spaces. In J. Tsujii, J. Henderson, & M. Paşca (Eds.), *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1201–1211). Jeju Island, Korea: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D12-1110/`

Spencer, A. (1991). *Morphological theory: An introduction to word structure in generative grammar*. Oxford: Blackwell.

Srinivasan, M., Al-Mughairy, S., Foushee, R., & Barner, D. (2017). Learning language from within: Children use semantic generalizations to infer word meanings. *Cognition*, *159*, 11–24.

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

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

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

Štekauer, P. (2016). Compounding from an onomasiological perspective. In P. ten Hacken (Ed.), *The semantics of compounding* (pp. 54–68). Cambridge University Press Cambridge.

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

Sun, Z., Zemel, R., & Xu, Y. (2021). A computational framework for slang generation. *Transactions of the Association for Computational Linguistics*, *9*, 462–478. Retrieved from `https://aclanthology.org/2021.tacl-1.28/` doi: 10.1162/tacl_a_00378

Tagliavini, C. (1949). Di alcune denominazioni della "pupilla" (studio di onomasiologia, con speciale riguardo alle lingue camito-semitiche e negro-africane). In *Scritti*

minori (p. 529-568). Bologna.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description iii: Grammatical categories and the lexicon* (pp. 57–149). Cambridge University Press.

ten Hacken, P. (2016). *The semantics of compounding*. Cambridge University Press.

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

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Tjuka, A., & List, J.-M. (2024). Partial colexifications reveal directional tendencies in object naming. *Yearbook of the German Cognitive Linguistics Association*, *12*(1), 95–112.

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

Tratz, S., & Hovy, E. (2010b, July). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In J. Hajič, S. Carberry, S. Clark, & J. Nivre (Eds.), *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 678–687). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P10-1070/`

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

Ullmann, S. (1953). Descriptive semantics and linguistic typology. *Word*, *9*(3), 225–240.

Urban, M. (2012). *Analyzability and semantic associations in referring expressions: A study in comparative lexicology* (Unpublished doctoral dissertation). Leiden Uni-

versity.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 31, pp. 6000 – 6010). Curran Associates Inc. Retrieved from `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`

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

Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2016, August). Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 1671–1682). Berlin, Germany: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P16-1158/` doi: 10.18653/v1/P16-1158

Wälchli, B. (2005). *Co-compounds and natural coordination.* Oxford University Press, USA.

Wang, S., & Bond, F. (2013, October). Building the Chinese open Wordnet (COW): Starting from core synsets. In P. Bhattacharyya & K.-S. Choi (Eds.), *Proceedings of the 11th workshop on Asian language resources* (pp. 10–18). Nagoya, Japan: Asian Federation of Natural Language Processing. Retrieved from `https://aclanthology.org/W13-4302`

Warren, B. (1978). *Semantic patterns of noun-noun compounds.* Acta Universitatis Gothoburgensis.

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

Westbury, C., & Hollis, G. (2019). Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behavior Research Methods*, *51*(3), 1371–1398.

Wierzbicka, A. (1984). Apples are not a "kind of fruit": The semantics of human categorization. *American Ethnologist*, *11*(2), 313–328.

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

Wisniewski, E. J., & Love, B. C. (1998). Relations versus properties in conceptual combination. *Journal of Memory and Language*, *38*(2), 177–202.

Wu, S., Cotterell, R., & Hulden, M. (2021, April). Applying the transformer to character-level transduction. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: Main volume* (pp. 1901–1907). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.eacl-main.163` doi: 10.18653/v1/2021.eacl-main.163

Wu, W., & Yarowsky, D. (2020, May). Computational etymology and word emergence. In N. Calzolari et al. (Eds.), *Proceedings of the 12th language resources and evaluation conference.* Marseille, France: European Language Resources Association. Retrieved from `https://www.aclweb.org/anthology/2020.lrec-1.397`

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

Xu, A., Kemp, C., Frermann, L., & Xu, Y. (2022). Word formation supports efficient communication: The case of compounds. In J. Culbertson, A. Perfors, H. Rabagliati,

& V. Ramenzoni (Eds.), *Proceedings of the 44th annual meeting of the Cognitive Science Society.*

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

Xu, A., Kemp, C., Frermann, L., & Xu, Y. (2024). Word reuse and combination support efficient communication of emerging concepts. *Proceedings of the National Academy of Sciences*, *121*(46), e2406971121.

Xu, A., Kemp, C., Frermann, L., & Xu, Y. (2025). Modelling compounding across languages with analogy and composition. In D. Barner, N. Bramley, A. Ruggeri, & C. Walker (Eds.), *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 47).

Xu, J., & Li, X. (2014). Structural and semantic non-correspondences between Chinese splittable compounds and their English translations: A Chinese-English parallel corpus-based study. *Corpus Linguistics and Linguistic Theory*, *10*(1), 79–101.

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

Xu, Y., & Kemp, C. (2015). A computational evaluation of two laws of semantic change. In *Proceedings of the 37th annual meeting of the Cognitive Science Society.*

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

Xu, Y., Malt, B. C., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive Psychology*, *96*, 41–53.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient

communication: The case of container names. *Cognitive Science*, *40*(8), 2081–2094.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., & Matsumoto, Y. (2020). Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 23–30). Association for Computational Linguistics.

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

Yu, L., & Xu, Y. (2021, November). Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 920–931). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.emnlp-main.71` doi: 10.18653/v1/2021.emnlp-main.71

Yu, L., & Xu, Y. (2023, July). Word sense extension. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 3281–3294). Toronto, Canada: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2023.acl-long.184/` doi: 10.18653/v1/2023.acl-long.184

Yu, L., & Xu, Y. (2025). Infinite mixture chaining: An efficiency-based framework for the dynamic construction of word meaning. *Open Mind: Discoveries in Cognitive*

Science, *9*, 1–24.

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

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

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

Zwicky, A. M. (1985). Heads. *Journal of Linguistics*, *21*(1), 1–29.