Language Sciences 83 (2021) 101326

Contents lists available at ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

Hyphenation as a compounding technique in English

Kun Sun*, R. Harald Baayen

Department of Linguistics, University of Tübingen, Germany

ARTICLE INFO

Article history:
Received 11 August 2019
Received in revised form 16 August 2020
Accepted 5 September 2020
Available online xxx

Keywords:
Hyphenation
Right-headedness
Correct parsing
The S-Curve model
Discriminative learning

ABSTRACT

Hyphenated compounds have largely been neglected in the studies of compounding, which have seldom analysed compounds in context. In this study, we argue that the hyphen use in compounds is strongly motivated. Hyphenation is used when words form a unit, which reduces the possibility of parsing them into separate units or other forms. The current study adopts a new perspective on contextual factors, namely, which part of speech (PoS) the compound as a whole belongs to and how people correctly parse a compound into a unit. This process can be observed and analysed by considering examples. This study therefore holds that hyphenation might have gradually become a compounding technique that differs from general compounding principles. To better understand hyphenated compounds and the motivation for using hyphenation, we conduct a quantitative investigation into their distribution frequency to explore how English hyphenated compounds have been used in over the last 200 years. Diachronic change in the frequency of the distribution for compounds has seldom been considered. This question is explored by using frequency data obtained from the three databases that contain hyphenated compounds. Diachronic analysis shows that the frequencies of tokens and types in hyphenated compounds have been increasing, and changes in both frequencies follow the S-curve model. Historical evidence shows that hyphenation in compounds, as an orthographic form, does not seem to disappear easily. Familiarity and economy, as suggested in the cognitive studies of compounding, cannot adequately explain this phenomenon. The three databases that we used provide cross-verification that suggests that hyphenation has evolved into a compounding technique. Language users probably unconsciously take advantage of the discriminative learning model to remind themselves that these combinations should be parsed differently. Thus the hyphenation compounding technique facilitates communication efficiency. Overall, this study significantly enhances our understanding of the nature of compounding, the motivations for using hyphenation, and its cognitive processing.

1. Introduction

Compounds are found in many languages around the world, and more importantly, compounding is one of the most widespread morphological techniques (Dressler, 2007). Compounding is also “the most frequently used way of making new lexemes in many languages” (Booij, 2012: 75). Moreover, compounding plays an important role in English word-formation,

* Corresponding author.
E-mail address: kun.sun@uni-tuebingen.de (K. Sun).

https://doi.org/10.1016/j.langsci.2020.101326

and is one of the most frequently used methods to create new words in English. Due to its importance, compounding has been widely and extensively explored from the phonological (Plag, 2018), morphological (Masini, 2009; Scalise and Vogel, 2010), syntactic (Peter and Neeleman, 2010), cognitive (Jarema, 2006; Fehringer, 2012; Kuperman and Bertram, 2013) and computational (Pirrelli et al., 2010) perspectives.

When two (or more) words join to form a new word, the result is a compound, which usually gives rise to a new meaning. The meaning of this compound could be similar to or different from the meaning of its components in isolation. Generally, there are three types of compounds with respect to their orthographic forms, namely, open compounds (there is a space between the words, such as “firing squad”), hyphenated compounds (such as “long-term”), and closed compounds (a solid form, such as “blackboard”). There are also criteria based on the part of speech (PoS) that PoS the entire compound belongs to. Two-, three- and four-word compounds can be found if the number of components in a compound is used as the criterion.

A hyphenated compound can be understood simply as a compound consisting of two (or more) words with hyphenation. However, it is rarely recognized that hyphenated compounds are quite distinct from their counterparts. For example, the use of hyphenation makes it possible for two or more consecutive words within a sentence to be linked to form a new unit with new lexical, syntactic and semantic functions. These hyphenated combinations can express temporary concepts or a whole range of concepts and things by turning a phrase into a word, such as “state-of-the-art” and “top-of-the-line”. In these cases, hyphenation can be used to signal a modifier prior to a noun or a noun phrase of a phrasal compound. Sanchez-Stockhammer (2018:187f) showed that compounds in attributive positions tend to be hyphenated. The preference for hyphenation in the attributive position is due to the need to parse correctly or avoid ambiguity. More importantly, the compositional mechanism for forming hyphenated compounds is quite distinct from the compositional mechanism for other types of compounds.

Nonetheless, hyphenated compounds have largely been ignored in the academic literature. Likely because of the common belief that using hyphenation to link words (such as punctuation practices) tends to be conventional (i.e., just a matter of spelling), only a few studies have explored hyphenated compounds, such as Mondorf (2009) and Kuperman and Bertram (2013). It also seems reasonable to assume that hyphenation acts as an intermediate form between open and closed compounds and that the hyphenated form will likely disappear from the orthographic form of a compound for reasons of economy (Sanchez-Stockhammer, 2018: 348). Furthermore, it is assumed that the possible arbitrariness of hyphenation practices is associated with the randomness of hyphenated compounds because hyphenation is sometimes used to combine two or more words for various reasons. People tend to infer from the random use of hyphenation in combining two or more words that the use of hyphenated compounds itself is arbitrary. The other possible reason for the lack of attention to hyphenated compounds is that in the past studies mentioned above, compounds are usually treated in isolation without context. Specifically, compounds are analysed by considering factors such as semantic transparency, the relation between two (or more) components, and the PoS of the components, whereas compounds are not analysed within sentences. The following contextual factors have not been considered in past studies: to which PoS the compound as a whole belongs, how people correctly parse a compound into a unit, what the syntactic order is of the components, etc. By considering these factors, we aim to identify the specific characteristics of hyphenated compounds that are distinct from the other types of compounds. Furthermore, although compounds are studied in morphology, hyphens are viewed merely as one graphemic/spelling device. However, morphology can be interfaced with spelling (Galani et al., 2011). The current study holds that the hyphen as a graphemic symbol has been successfully adopted to help produce compounds. To some extent, this aspect of spelling has become a compounding technique.

Since hyphenation in compounds is assumed to be a matter of (habitual) spelling, this assumption might indicate that hyphenation is redundant and tends to disappear over time because of considerations of economy. However, hyphenation in compounds has not disappeared at all. Instead, this study hypothesizes that hyphenation in compounds is strongly motivated rather than used as a practice that is only habitually or randomly used. We also propose that hyphenation has gradually evolved as a compounding technique in English. However, its function has been seriously underestimated. The motivation for using hyphenation in compounds can be clearly explained by theoretical analysis. A historical perspective not only can provide convincing evidence of and insight into this phenomenon but also can aid in better understanding the development of orthographic forms in compounds. The availability of large-scale corpora makes it possible to conduct quantitative investigations in this area. Past relevant studies have seldom adopted a diachronic perspective to examine the change in the frequency of compounds over time. This study therefore employs a diachronic perspective in investigating the changes that hyphenated compounds have undergone over the last 200 years.

Our main concerns are about whether hyphenation has declined in frequency, and whether hyphenation in compounds is friendly to English readers. To clarify the concerns, the structure of this study is arranged as follows. The theoretical analysis inspires us to hypothesize that hyphenation in compounding is strongly motivated and that hyphenated compounds are quite distinct from common compounds. However, we need to use diachronic data to testify or refute it. If hyphenated compounds are unmotivated or no longer motivated, they will decline in frequency. Based on our collection and analysis of different types of data, we might conclude that if hyphenated compounds have not declined in frequency, then there is good reason to believe that they are motivated. Furthermore, the possible motivations for hyphenation can be inferred. For instance, the discriminative approach to language comprehension and production is efficient, and the discriminative learning model (Baayen et al., 2011) can explain the function of hyphenation in compounding. Hyphenation helps to discriminate different types of compounds, which helps to reduce uncertainty and confusion with the other types of compounds. Hyphenation can be used to convey messages successfully with minimal effort, which leads to communication efficiency.

2. Background and hypothesis

2.1. Previous studies: core characteristics of compounds

Some core characteristics of compounds and hyphenated compounds in English must be considered before we pursue an in-depth quantitative and diachronic analysis of this phenomenon. According to past studies, most English compound words have three core characteristics specifically: *right-headedness*, idiom-likeness and a syntactic order.

Plag (2018: 135) holds that the vast majority of compounds in English can be interpreted such that the left-hand member modifies the right-hand member in some way, e.g., a "black tea" is a type of tea that is more oxidized than oolong, green, and white teas. In this compound, "black" is a modifier and "tea" is a head. Such a compound exhibits what is called a modifier–head structure. i.e., a compound with a binary structure (Dressler, 2007) (or endocentric structure).

Compound heads usually occur on the right-hand side (Williams, 1981), and the entire compound acquires most of its semantic and syntactic information from its head. This principle is called *right headedness*. For instance, if the head is a verb, the compound will be a verb (e.g. "baby-sit"); if the head is a count noun, the compound will be a count noun (e.g. "waterfall"). Therefore, *right headedness* involves two factors. When we speak of breaking the rule of *right-headedness*, it indicates a violation of either of these two factors.

Compounds are traditionally distinguished from phrases by examining their fixedness. It is impossible to insert any other words, such as adjectives, between two nouns in compounds. For instance, it is unacceptable to say a "post an office". In this respect, compounds are similar to idioms; therefore, this criterion can be called *idiom-likeness*(Dressler, 2007).

Compounding is usually treated as an important morphological strategy. However, an important principle for creating compounds is the combination of words that follow the syntactic order (the syntactic order of functional elements that constitute a linguistic unit). This principle is often neglected, and it includes, for example, the following combinations: Subject(doer) + Verb; Verb + Noun; Verb + Prep./Adv.; Adj. + Noun; and Noun + Noun. These combinations can be treated as the canonical syntactic order that occurs with high frequency (Hunston and Francis, 2000: 51-59; Jurafsky, 1996). Although the syntactic order takes effect at the syntactic level, it still plays a morphologically important role in compounding. Some of the main types of syntactic order can be exemplified in compounding as follow: Subject(doer) + Verb: "daybreak, headache, rainfall", etc.; Verb + Noun: "playboy, firing squad, wading bird, watchdog", etc.; Adj. + Noun: "longboat, blackboard, greenhouse", etc.; and Noun (modifier) + Noun: "cigar smoker, cable car, power plant, silkworm, piano keys, dragonfly", etc.

Bauer and Renouf (2001) used a large corpus of British newspapers to investigate different types of compounds in English, and they specified some cases that violated textitright-headedness; these cases were mostly hyphenated forms, although this feature was not explicitly pointed out in their study.

2.2. Hyphenated compounds

The majority of studies of compounding have analysed compounds without context and have seldom considered the PoS perspective. The previous studies mentioned above analysed only the PoS of each component in a compound while ignoring the function and use of the compound in its context. Since a compound is still a word that works within a sentence, the PoS should be an important characteristic within it. This study is mainly concerned with the PoS of hyphenated compounds as a whole and examines the core features of these compounds according to the PoS to which the entire compound belongs (rather than according to the PoS of its components). We focus on hyphenated compound adjectives, hyphenated compound nouns and hyphenated compound verbs.[1]

Concerning the PoS of hyphenated compounds as a whole, we can further explore the characteristics of these compounds by considering the contextual factors. The use of hyphenation in compounds is prone to a) violating rightheadedness or frequently used syntactic order and b) avoiding ambiguity, which can in turn influence the use of hyphenation. Most hyphenated compounds violate the principles of right-headedness, syntactic order, or both. We also find that using hyphenation in these compounds helps readers parse two or more words into one unified combination. Without hyphenation, there is a high risk of ambiguity; that is, two or more words might be parsed into the components of other syntactic elements within a sentence. For this reason, we emphasize the influence of context on compounds. Once such a unit including a hyphen is parsed into a compound, its composition mechanism usually does not abide by the principles of right-headedness, syntactic order, or both. In the following, we discuss the three types of hyphenated compounds according to their PoS and explain their characteristics on the basis of the aforementioned factors.

Hyphenated-compound adjectives Hyphenated-compound adjectives can be classified into three types according to their respective components: a) ADJ + NOUN/ADJ/VERB: "long-term, high-tech, high-quality, low-income; red-hot; two-faced, good-looking, better-placed, sure-fire", etc.; b) NOUN + ADJ/VERB/NOUN: "ice-cold, sky-high, coal-black, oil-rich,

[1] We do not use the terms nominal, adjectival or verbal compounds (Plag, 2018: 144-154) to refer to the PoS to which the compound as a whole belongs because a nominal compound is not necessarily a noun, and an adjectival compound is not necessarily an adjective. Nominal compounds refer to compounds with nouns as heads (Plag, 2018: 144), but a compound with a noun head is not necessarily a noun, particularly in the case of hyphenated compounds.

snow-blind, life-preserving, labour-saving, time-consuming, record-breaking, part-time”, etc.; and c) Phrasal compound: “black-and-white, state-of-the-art, wellto-do”, etc.[2]

From the PoS perspective, the hyphenated compound types of ADJ + NOUN/VERB or NOUN + VERB cannot be treated as ADJECTIVEs according to the principle of *right-headedness*, because their right heads are not adjectives. Thus, if a “*head*” in a compound is NOUN/VERB, then the entire compound should be a NOUN/VERB according to *right-headedness*. For instance, the compounds “long-term, high-tech, high-quality, low-income, and good-looking” cannot be treated as adjectives because their right heads are nouns/verbs (“term, tech, quality, income, and looking”). However, with the use of hyphenation, the entire compound can work as an adjective. As mentioned above, violating either principle (the head is on the right, and a whole acquires its semantic and syntactic information from its head) of *right-headedness* means breaking the rule of *right-headedness*. Hyphenated compounds thus violate the principle of *right-headedness*.

It is noted that the types of ADJ + NOUN/VERB or NOUN + VERB somewhat tend to take on specified forms such as ADJ + V-ed, ADJ + N-ed, NOUN+ V-ed or NOUN+ V-ing, which means that the second component in such a compound is likely to possess the structure of *ed* or *ing*. We can observe numerous examples such as “good-natured” (ADJ + N-ed), “black-eyed”(ADJ + Ned), “rosy-cheeked” (NOUN + N-ed), “snow-covered” (NOUN + V-ed [particle]), “short-lived” (NOUN + V-ed [particle]), “life-giving” (NOUN+V-ing) etc. The form of *N-ed* is rarely used alone and usually occurs in a hyphenated compound. It is called an *-ed* denominal adjective which differs from a typical adjective (Ljung, 1976; Takehisa, 2017). These *N-ed* adjectives are derived from nominals. The meaning associated with denominal -ed adjectives originates from the suffix's denotation that requires a relation. In this sense, since its right head *N-ed* is not really an adjective, the type of combination *X* + *Ned* breaks the principle of *right-headedness*. As for the type of *X* + *V-ed/ing*, the order of the first component and V-ed/ing (the second component) is not frequent with respect to the frequently used syntactic order mentioned above.

For instance, “snow-covered” and “short-lived” could have the normal order of “covered [with/by] snow” and “lived [for] short [time]”, respectively. The normal order of “life-giving” should be “giving [sb, a] life”. In this sense, most *X* + *V-ed/ing* components violate the normal syntactic order.[3]

The type NOUN + ADJ follows the principle of *right-headedness*, but its order of combination is also not normal with respect to the syntactic rule. For instance, adjectives are unlikely to be combined with (action or behavior) verbs. Nouns usually do not precede adjectives, but one such form is “ice-cold” (*cold* modifier + *ice*-head). This study treats compounds that consist of three (or more) words as phrasal compounds because such a compound might be a phrase that originally lacked hyphenation. For example, “well”, which is located on the left, is the focus of “well-to-do”; “black-and-white” is formed with a coordination between “black” and “white”, and “state”, the head, is located on the left in the phrasal compound “state-of-the-art”. These phrasal compounds therefore violate the principle of *right-headedness*, but the use of a hyphen makes it possible to transform a phrase into a compound.

Many high-frequency compound adjectives are normally used as modifiers (or attributives) that precede nouns (such as “short-term, large-scale, high-profile, would-be”). This pattern of ADJ + NOUN is easily seen as a compound noun because the right head is not an adjective (“term, scale, profile, be”, etc.). Thus, they might acquire a different syntactic function (when treated as adjectives). Accordingly, they do not follow the principle of *right-headedness* (a compound that acquires syntactic information from its right head). The following examples illustrate why hyphens are used in compounding when the context is provided.

There are some compound words with “style (noun)”, such as “1920s-style dancing, Chicago-style hyphenation, according to Chicago style, and headlinestyle capitalization, using headline style”. Among these examples, if a compound with “style” is used as an adjective modifying a noun, a hyphen becomes useful; otherwise, hyphenation does not occur, e.g., “according to Chicago style”, and “using headline style”. We accept a “full-length mirror”, but “the mirror is full length”. “Full length” in “the mirror is full length” can also be treated as an adjective here (which works as a predicate), whereas “full length” is hyphenated when it acts as an adjective (modifier of a noun) in “full-length mirror”. If this usage is not the case, readers find it difficult to judge and parse “full length” as a unit because the presence of both nouns “length” and “mirror” can lead to confusion (Gordon et al., 2002, 2006; Van Dyke and McElree, 2011). Without hyphenation, there is a high risk of ambiguity. For this reason, hyphens become crucial for ADJ + NOUN compounds that function as adjectives. There is nothing on the right side for “full length” in “the mirror is full length”; therefore it is posited that readers do not need a mark or sign to help judge and parse the syntactic features of this combination. The possible purpose for using hyphenation in these compounds is to help readers understand that the combination will be treated as one unit to avoid the risk of ambiguity.

[2] The PoSs of these hyphenated compounds can be found in some authoritative English dictionaries (such as Oxford, Longman, and Macmillan). Additionally, like other common adjective, these compounds have comparative and superlative degrees. For example, we can state, “more large-scale” and “the most large-scale”. Some high-frequency compounds that modify nouns have been treated as adjectives by these dictionaries. Similarly, the PoSs of hyphenated-compound nouns or verbs also have features similar to the features of common nouns or verbs and can be found in the dictionaries. Although some newly-created hyphened compounds are not in these dictionaries, their PoS can be contextually identified. Generally, they should be not be treated as attributives (Berg, 2011).

[3] Synthetic compounds are also called *synthetic compound nouns*. Such compounds usually have two forms, namely, Noun + Verb-ing (e.g., city-planning, decision-making) and Noun + Verb-er (e.g., dishwasher, taxi driver). The structure of NOUN + V-ing in our database is likely to be hyphenated. In our database, there are 33 compounds which can be treated as synthetic compounds. The reason for this is that the order of two components violates the canonical syntactic order. The other form of synthetic compounds Noun + Verb-er rarely occurs in our database (There are two examples, “story-teller, fund-raiser”). The less use of hyphenation in such construction is because of no violation of right-headedness or syntactic order. For example, “fund-raiser” is a noun but its right head is also noun, so the hyphenation seems unnecessary.

Hyphenated-compound nouns Hyphenated-compound nouns can be simply classified into two types as follows: a) NOUN + NOUN/ADJ/VERB: "actor-manager, major-general, decision-making, president-elect"; and b) VERB + ADV/PREP: "take-off, sell-out, wrap-up, sit-in".

The relationship between "actor" and "manager" is coordinated such that there is no head for this compound. "Decision-making" and "president-elect" could have the normal order of "making [a] decision" and "elected president", respectively. They violate the rule of syntactic order. Hyphenation helps readers to analyse two words as one unit, which reduces the possibility of parsing them into two separate units or other forms in the context. For instance, "decision making" without hyphenation could be interpreted as "the decision makes something happen". Other compound nouns with ungrammatical combinations require hyphens, such as the ADJ + VERB type, e.g.,"double-tune, free-associate" (Bauer, 1983: 208). Furthermore, we often see a "PREFIX + NOUN" type, e.g., "T-shirt, X-ray, ex-wife". This type is not considered in this study.

Note that one type V + ADV/PREP combined by hyphenation works as either noun or adjective. When a hyphen is not inserted into such a verb phrase, the combination of the phrase can hardly be treated as noun. As a result, the PoS of type V + ADV/PREP does not derive from *right-headedness* because their right heads are not nouns ("-off, -out, -up" are not nouns). For instance, "drive-in" is originally a verb phrase, but it acts as a noun after a hyphen is inserted. The combination violates the principle of *right-headedness* because it has no noun at all.

Hyphenated-compound verbs These compounds mainly include: NOUN/VERB/ADJ + VERB, such as "lip-read, bottle-feed; window-shop; stir-fry, freeze-dry; dry-clean".

For some compound verbs in this category, the normal order could be VERB + (or other elements) + NOUN, such as "shop (at the) window, read (by watching) lips, feed (with the) bottle". When the order is reversed, hyphenation becomes useful because it can help readers treat the combination as a new unit. Following traditional compounding methods, this combination would not be analysed as a compound. When the two verbs "stir" and "fry" are combined, they are coordinated, which means that there is no head in this compound. In this sense, it becomes fairly important to use hyphenation in a coordinated structure if they need to be combined as a compound. When "dry" is an adjective, its direct combination with the verb "clean" is rare. In contrast, the right head in "underestimate" is the verb "estimate"; thus, the compound abides by the principle of *right-headedness* and a hyphen seems to have little use in such a compound. Generally, these aforementioned combinations would be unlikely to be parsed as compounds without the use of hyphenation.[4]

Additionally, in using hyphenation, reduplication-style compounds and balancestyle (coordinated or tautological) compounds combine similar constituents (Wälchli, 2005), i.e. there is no head (or there would be two heads) in this type of compound, which indicates that it also breaks the principle of right-headedness (Benczes, 2014). The reduplication of compound words requires hyphenation, such as "goody–goody, wiggle-waggle, and uh-huh" etc. There are many examples of balance-style hyphenated compounds, such as "day-to-day, and face-to-face". As a matter of fact, we can treat these compounds as a special type—the coordinate type although they actually break the principle of *right-headedness*. Overall, we classify all of the hyphenated compounds into three types according to their composition characteristics, namely, violating *right-headedness*, violating the syntactic order, and the coordinate type.[5]

We have provided a list of the main types of hyphenated compounds in accordance with the syntactic properties of their constitutive elements instead of providing an exhaustive list of all types. Checking the various types of hyphenated compounds mentioned in some books on English morphology (Bauer et al., 2013: 201-213), we find that the majority of hyphenated compounds violate the principles of *right-headedness*, syntactic order, or both. Despite the differences between hyphenated compounds and other types of compounds, they still share great similarity, that is, two (or more) components in a hyphenated compound should be treated as one unified word, just like an idiom instead of being separated, which is the same as the other types of compounds. According to the above characteristics, we follow two steps to roughly identify hyphenated compounds. First, two components in a hyphenated compound should be unified to form a

[4] The context usually allows for disambiguating between a compound reading and other potentially available interpretations. A quick search for the examples mentioned above in the COCA/BNC also clearly shows that they are sometimes used without hyphens. However, these cases require the specific context. For example, if "window-shop" (verb) is followed by nouns, the hyphen is needed. The form "window shop" also exists, but it cannot be followed by a noun; rather, it can be followed by a preposition or a comma. When "window shop" is followed by a noun, there is a substantial risk of ambiguity. In addition, "window shop" does not exist in the Corpus of Historical American English (COHA). In this sense, "window-shop" is not necessarily equal to "window shop", or 'window shop" is an alternative of "windowshop". In this sense, hyphenated compound verbs usually maintain their hyphenation when followed by nouns. This example is quite similar to the example of "full-length" in subsection 2.1. Hyphenated compound adjectives also usually maintain their hyphenation when followed by nouns. However, hyphenation might not be needed, When the context provides clues that are necessary to correctly analyse these compounds.

[5] The third reviewer points that the difference between endocentric vs exocentric compounds is helpful in understanding compounding composition. Exocentric compounds usually refer to those compounds without headedness, such as "scarecrow, pickpocket". "A scarecrow" is an object designed to scare not only crows but all birds. Bauer (2008) believes that these compounds, a Romance type, are indeed exocentric. However, with regard to *verb phrase* compounds, Bauer (2008) holds that the process of composition is probably endocentric. For instance, "drive-in" might be deemed to be a type of driving, "clean-up" is a form of cleaning. According to Bauer (2008), the other types ("red cap", "birdbrain") are endocentric in structure but happen to be interpreted figuratively. Benczes (2004, 2005, 2015) hold the similar view of exoncetric compounds with Bauer (2008). Overall, the Roman type is really exocentric but the other types are endocentric in structure. The three types of hyphenated compounds in our database do not contain the Roman type of exocentric compounds. There are 48 *verb-phrase* hyphenated compounds in the DEHC (e.g. take-off, stand-up). However, as discussed above, *verb-phrase* compounds can be analyzed in structure (i.e. the head should be a verb), so *verb-phrase* hyphenated compounds should not be treated as exocentric. In short, the exocentric concept seems less helpful in analyzing the structure of hyphenated compounds.

different meaning. Second, hyphenation should not be removed; otherwise, the syntactic or semantic function of this compound will change.

2.3. The hypothesis and research questions

The use of hyphenation in compounds might vary slightly according to the context, as noted above. However, the primary purpose of hyphenation might be to help readers to correctly parse a combination as a new unit. Otherwise, the combination would be unlikely to be treated as a compound without hyphenation, and there is a risk of ambiguity. When hyphenation joins two or more words in a new combination, the combination itself acquires new lexical, semantic and syntactic functions. That is, when two words are put together, it does not necessarily mean that these two words are always treated as a compound. Without hyphenation, the combination is unlikely to be treated as a compound.

We find that the practice of hyphenation in compounding leads to at least two results. First, this hyphenated combination forms a new compound, and second, the new compound cannot be interpreted and analysed by using the traditional methods (morphological, syntactic or semantic) of parsing common compounds. However, hyphenated compounds violate the general compounding principles of (*right-headedness* and syntactic order). This violation can be observed and analysed through specific examples of hyphenated compounds. Our observation is supported by Bauer (2017: 4) who holds that the lack of orthographic unity in open compounds may create ambiguity with syntactic phrases, while hyphenated or closed compounds can disambiguate this distinction.

It can be assumed that the primary purpose of using hyphenation in compounds and the results of violating the principles of *right-headedness* and syntactic orders are reciprocally related. When two or more words are joined in a unit with a hyphen, the hyphenated compound usually has a different structure than common compounds. When this outcome is not the case, hyphenation becomes unnecessary; conversely, when the structure of such a combination violates the principles of compounding, it can yield a new meaning and a new syntactic function (two aspects in one event). If hyphenation in compounds is strongly motivated, hyphens should not disappear so easily over time. Notably, hyphens are used at random in some expressions (not compounds). If people did not need to use hyphenation as a signal in parsing a combination as a compound (i.e., if the hyphenation were useless in causing these expressions to produce new meanings or syntactic functions), the use of such hyphens would likely disappear over time.

To test this hypothesis, we examine hyphenated compounds from a diachronic perspective. Token frequency from a diachronic perspective is usually used to examine the motivation for using a language structure or for the emergence of a language structure (Diessel, 2007; Hilpert, 2015). If the frequency decreases, we might conclude that the use of hyphens in compounds is not motivated. Our specific research questions are as follows: a) Has hyphenation declined in frequency? and b) Is hyphenation in compounds useful for English readers or writers? The diachronic change in the frequency of hyphenated compounds is of great help in evaluating the motivation for using hyphenation in compounds. To this end, we collect data from three databases on frequency changes, as shown in the following section.

3. Materials and Methods

3.1. Materials

We can derive accurate and comprehensive data on the historical changes in the frequency of hyphenated compounds from large-scale corpora using natural language processing (NLP) technology. BYU Corpora and the Google N-gram corpus are used to obtain the data in this study. The main corpora include the COHA (the Corpus of Historical American English) (Davies, 2010) and the Google N-gram Corpus (Books, 2016).

The COHA is the largest structured corpus of historical American English. It contains more than 400 million words of text from the 1810s to the 2000s, and it is classified by genre and by decade. The N-Gram corpus of Google Books contains digitalized text sources printed between 1500 and 2008 in eight languages. "It is also currently the world's largest corpus and the only corpus that enables resolution at a fine temporal scale (yearly) over a long period of time. " (Michel et al., 2011). A number of hyphenated compounds were chosen to examine the tendencies of their frequency changes using the data derived from the Google Books corpus. This corpus contains American and British English and its range is more comprehensive than that of COHA. However, the COHA, which consists of multiple genres, might represent the English language more accurately than the N-gram corpus of Google books, which comprises only digitalized books.

3.2. Methods

The DEHC In this study, we gather data on the occurrences of two-/three/four-word hyphenated expressions (sometimes constrained by the PoS) by searching the COHA with the following search queries (SQ): SQ1: *-*(j*—nn*—v*)[6]; SQ2:*-*-*; and SQ3: *-*-*-*. This capacity is not available with the wildcard when searching with "*" in the Google N-gram corpus, but

[6] The actual operation is *-*j*, *-*nn*, and *-*v*. The three search queries can be used to find adjectival two-word compounds, nominal two-word compounds, and verbal two-word compounds.

the N-gram Viewer (Google N-gram corpus) supports some specific compound searches. Therefore, the COHA is used to obtain general data on some patterns of hyphenated compounds, and the Google N-gram corpus is used to search for the trends of changes in specific compounds.

The search method is used to obtain a massive number of hyphenated expressions mixed with hyphenated compounds. Some of these create *noise* in the actual data; therefore, *hyphenated expressions* (which contain hyphens, but some are not truly compounds) must be manually excluded. Additionally, the PoS tags for the hyphenated compounds retrieved from the COHA are not very precise, which means that we had to also check them manually. For instance, some hyphenated numerical expressions are not compounds at all although their frequency is quite high in the COHA.[7]

To reduce the negative impact from the *noise*, we built a **database of English hyphenated compounds (DEHC)** to ensure that the data collected are valid. After choosing 880 different high-frequency hyphenated compounds retrieved from the COHA according to the raw frequency and the frequency of one million words, we were able to use the data from this hyphenated compound corpus that has been manually filtered to conduct quantitative analyses. This process is helpful for achieving a better understanding of the diachronic changes.

The construction of this database (the DEHC) must be discussed in greater detail. After the 1000 most frequent two-word hyphenated compounds are retrieved from the COHA by raw frequency and one million words (the upper limit for retrieval is 1000), these non-hyphenated compounds are manually excluded, including combination of numbers (such as “one-fifth, year-old”, etc.), prefix-type hyphenated words (such as “ex-wife, self-defense, and mother-in-law”), and meaningless hyphenated symbols (such as, “and-, – ” etc.). After the top 1000 three-word hyphenated forms are retrieved, the same manual evaluation method is also applied. In addition to this process, words with a raw frequency less than 20 are not taken into account because most of them are not the type of hyphenated compounds that we discuss in this study. When four-word hyphenated forms are retrieved, those with a frequency of less than 10 are excluded. Actually, the frequency per one million for the most four-word expressions is quite small, to the point of being almost negligible (it is difficult to treat these expressions as compounds). Therefore, five-word hyphenated expressions are not considered because of their low frequency, e.g., the highest-frequency five-word hyphenated compound (“take-it-or-leave-it”) has frequencies of 0.08 (per one million words) and 32 (raw frequency). Our standard frequency is based on one million words; thus these words with low frequency can be neglected.[8]

The DEHC contains 880 different hyphenated compounds, specifically, 675 two-word compounds, 148 three-word compounds and 57 four-word compounds. Each hyphenated compound in the DEHC contains the following information and data: *lemma*; *the PoS*; *the number of components* (two-, three- or four-word types); *composition types(which violate right-headedness, the syntactic order, and the coordinate type)*; *the PoS of each component* and *the standardized frequency per decade*. The data on frequency was obtained from the COHA which represents the English language more precisely than the Google N-gram corpus, and frequency here is the standardized frequency calculated by one million words per decade from 1810 to 2000. The diversified types of hyphenated compounds discussed in 2.2 can be found in this database. The overall information on the DEHC is shown in Table 1.[9] In rare cases (eight compounds in our database), we find hyphenated compounds with several parts of speech (PoS). Specifically, these hyphenated compounds are in a cooperative type, e.g. “fifty–fifty, avant-garde, stir-fry, shilly-shally”. There is no immediate method to determine the exact number of hyphenated compounds in contemporary English, but this database includes the most frequently seen hyphenated compounds. In this sense, our method can guarantee an approximation of the real frequency distribution of hyphenated compounds in the COHA. In addition to using this fairly large database, the data are also obtained by extracting some high-frequency hyphenated compounds from the DEHC to enable comparisons with hyphenated expressions that are not compounds, as will be discussed in Section 4. The methods for obtaining the second and third types of databases are introduced in the corresponding sections.

Table 1
The DEHC composition (n = 880).

Feature	Item1	Item2	Item3	Item4
Number of component	two word (675, 77%)	three word (147, 17%)	four word (57, 6%)	NA
Composition types	violating right-headedness (RH) (528, 60%)	violating syntactic order (SO) (270, 26%)	coordinate type (CO) (111, 12%)	no violation (18, 2%)
PoS(as a whole)	adj (730, 83%)	noun (122, 14%)	verb (16, 2%)	adv (8, 1%); inter (3)

[7] More specific examples are given here. The first example concerns the PoS. The combination “ice-cream” is tagged as a noun in the COHA. Prior to the 1900s, this combination was equal to “ice cream”. However, after the 1900s, “ice-cream” became an adjective to modify a noun. The other examples, “three-year” and “two-year”, are tagged as adjectives with high frequency, but they are not compounds because there are unlimited possibilities of stating the combination of ”number-year”.

[8] The first reviewer pointed out that there will be a potential confound if we consider only the most frequent words over two centuries, not the most frequent words of each decade. Given that the later decades of the COHA contain more data than the earlier decades, this procedure might induce a certain bias towards types that are more frequent, or become more frequent, in the later periods. We have to be cautious when assessing frequency changes in diachronic corpora whose time slices differ in size, especially when we limit the scope of the investigation to the *n* most frequent items in the full corpus.

[9] The DEHC can be downloaded from https://osf.io/9g728/.

Quantitative and statistical methods The present study examines the diachronic change in the frequency of hyphenated compounds. As discussed in the subsection concerning the hypothesis, if the frequency decreases from a diachronic perspective, we might conclude that the use of hyphens in compounds is not motivated. However, there is good reason to believe that this use is in fact motivated. Frequency is a very important standard to measure the motivation for using hyphenation. The present study uses the frequency based on one million, which can be set in the interface of the BYU corpora and is calculated through "the raw frequency/the sub-corpus size * 1 million".

After obtaining the frequency of hyphenated compounds, a frequency distribution is performed because it can tell how frequencies are distributed, i.e. how frequencies changes over the course of decades. We also examine the frequency distribution for different types of hyphenated compounds (e.g., the number of components, composition type etc.). As discussed in the Background section, hyphenated compounds can be classified into three types according to its composition type; therefore we feel that it is necessary to conduct a quantitative analysis for the three types. In addition, we can take advantage of the data from the DEHC to explore the relationship among the variables and further understand how the variable *decade* has influenced the frequency changes in *token*/*type* and how the other variables have influenced the changes of frequencies. For instance, *token* or *type* can be treated as a response variable, and the other variables, such as *decade*, *type*, and *token*. are considered predictive variables. Some given statistical methods can be applied to test whether predictive variables exert effects on the response variable. The test can be performed with *Generalized Additive Mixed Models (GAMM) (*Wood, 2017) which is effective in fitting (non)linear models. GAMM allows the estimation of random effects for cases, treatment, trend, and other covariates. GAMM analysis can be executed by an R package "mgcv" (Wood, 2015).

The ideal here would be to compare hyphenated compounds and non-hyphenated compounds quantitatively. However, a quantitative and comprehensive comparison is almost impossible because non-hyphenated compounds are not annotated with tags and cannot be retrieved from current large-scale English corpora. Despite this disability, some available databases concerning some given types of non-hyphenated compounds can be used to perform comparisons to some extent.[10]. We can obtain data on the frequency of hyphenated compounds from corpora due to their special tags—a hyphen can thus be retrieved. Notably, a typically hyphenated compound without a hyphen might not be a compound at all in most cases. Additionally, the considerable number of non-hyphenated compounds (including numerous newly created compounds that have not been accepted by authoritative dictionaries) renders it impossible to collect reliable data on their frequency. Instead, we compare some typical cases among the three different forms (open, hyphenated, and closed forms) by using the data retrieved from the Google N-gram corpus.

To examine the proposed hypothesis in this study, we collected data from three databases, namely, the DEHC (large), the database on high-frequency hyphenated compounds (intermediate) and the database on the specific compounds from the Google N-gram corpus (small but specific). The methods for obtaining the second and third databases are elaborated on in the following sections.

4. The historical change of frequency

4.1. Token and type frequencies

This section mainly concerns changes in the frequency of different types of hyphenated compounds from a historical perspective to determine the regularities in these changes. The diachronic perspective is very helpful not only in gaining insight into why hyphenation in compounds might be preserved historically but also in testing the hypothesis that using hyphens in compounds is strongly motivated.

The DEHC data show that hyphenated adjectives and nouns (tokens) constitute 83% and 14% of the overall set, respectively. The majority of hyphenated compounds (98%) violate either *right-headedness* or the syntactic order or both. Among them, hyphenated adjectives (99%) follow neither *right-headedness* nor the syntactic order. Conversely, hyphenated nouns[11] follow *right-headedness* and syntactic principles and constitute almost 7% of all hyphenated compound nouns. Some of them have corresponding nonhyphenated forms with the same meaning.

It might be useful to compare the historical changes in the frequency of both types (the number of distinct words in a text, corpus, the composition types etc.) and tokens (the total number of words in a text, corpus, etc., regardless of how often they are repeated) for English hyphenated compounds using the DEHC that we created. The frequency of tokens for the different types of hyphenated compounds (e.g. according to the number of their components, such as two or three words) over the last

[10] A recent collection of close compounds is available (Gagné et al., 2019) However, there are no databases for open compounds or other types of compounds. Our database concerns the diachronic information on hyphenated compounds, whereas the database of Gagné et al. (2019) does not contain any diachronic data. Nonetheless, it would be interesting to investigate how the frequencies of different types of compounds developed over history and compare them using the available compound databases. Such a comparison is presented in the supplementary material. The data on the diachronic frequency of closed compounds is available at https://osf.io/9g728/.

[11] Some hyphenated compounds have not been accepted by dictionaries, but their PoS can be identified through their use in context. In most cases, these hyphenated compound nouns, which follow the syntactic order and *right-headedness* are ones that have corresponding non-hyphenated forms; in such instances, both forms coexist, such as "room-mate vs. room mate, living-room vs. living room, camp-fire vs. camp fire, bank-notes vs. bank notes, and easy-chair vs. easy chair". A very interesting example, "dinner-table", demonstrates that it was a noun at the beginning of the 19th century, but later, it gradually came to be used as an adjective. In the following section, we investigate how the frequency of these coexisting forms changed over time.

two hundred years can be obtained using the DEHC (as shown in Table 1 in the supplementary material). The total frequency has increased over the entire two-hundred-year period, and this increase is very obvious in the first forty years. It became relatively moderate before again increasing dramatically after the 1980s. We present the frequency of these types in our data (the total number of types is 880). The historical changes in the frequency of types share a similarity with the changes in tokens (as shown in Table 2 in the supplementary material). The total frequency has increased over the entire two-hundred-year period. The increase in total frequency for these types was also very obvious in the first forty years. Low-frequency hyphenated compounds that were mostly created after the 1990s have been excluded from our sample, given the small number of types in recent years. In fact, the frequency of types should have increased markedly in recent years. In addition, it is necessary to investigate the changes of frequency in the three types of hyphenated compounds according to the composition types (violating right-headedness, violating the syntactic order and coordinate), which has been discussed in the Methods subsection.

The process of a change in frequency in many phenomena of morphological, semantic and syntactic changes can be described by the Piotrowski law (Piotrowski and Piotrowski, 1974; Altmann, 1992; Wu et al., 2016), which “represents the development (increase and/or decrease) of the portion of new units or forms over time” (Turenne, 2010). Growth processes in language, such as vocabulary growth, the dispersion of foreign or loan words, and changes in the inflectional system, abide by the Piotrowski law, and this adherence corresponds to growth models in other scientific disciplines. The Piotrowski law is actually a special case of the so-called *logistic model*. When a logistic model is plotted, its shape looks like an S-curve; therefore, it is called an S-curve model, as shown in the following equation and the left panel of Fig. 1.

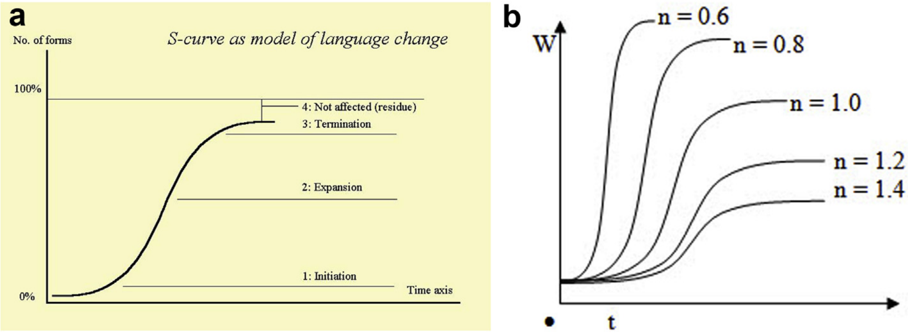

Fig. 1. Logistic functions and their curves.

The Piotrowski law is described by the following equation:$y = \frac{1}{1+a * \exp(-b * x)}$

In this equation, *a* and *b* are the parameters of the model; x stands for time. However, the generalized logistic function (or the Richards curve) is a widely used growth model that will fit a wide range of S-shaped growth curves (Richards, 1959; Tjørve and Tjørve, 2010). The logistic curve is symmetrical about the point of inflection of the curve. However, the growth curve in the generalized logistic function is asymmetrical. The generalized logistic function can be described by the following equation: $W = A\,(1 - exp\,(-k * t))^n$. In this equation, *W* is the value of the growth measurement (mass or length) at time *t*, *A* means maximum growth, and *k* and *n* are model parameters. If n = 1, the Richard's growth model reduces to the logistic model (or the Piotrowski law). Normally, the value of *n* will be around one, which is shown as the right panel in Fig. 1. The mean relative growth rate is given by R (rate) = k / (n+1).

The S-curve in the figure means that changes will be slow at the beginning, will then speed up, and will then again slow down at the end (Denison, 2003; Blythe and Croft, 2012; Stadler et al., 2016). The S-shaped curve is typically used to model the diffusion of linguistic changes across time. If speakers at any point in time are aware of the variants that are preferred and the variants that are being increasingly neglected, then the language can move in a definite direction, as was the case with the drift from synthetic to analytic in the history of English (Nevalainen, 2015). The propagation of change would seem to follow a pattern found in spheres other than language. In essence, an S-curve describes a change that starts slowly, gathers speed and proceeds rapidly but then stops—or at least slows considerably—before it reaches completion.

A logistic S-shaped curve is commonly seen in quantitative studies, but it is not the only pattern that emerges from systematic analyses. In addition to the possibility of stable variation, there are other options. For instance, Richard's S-curve growth has been found in the growth of multiple phenomena (McArdle et al., 2002; Katsanevakis, 2006; Koya and Goshu, 2013). The data on token/type frequencies are plotted in Fig. 2 and demonstrate what an S-shaped growth curve looks like. However, a little difference from an S-curve model is that the growth of hyphenated compounds from the 1800s seems very rapid. As we know, an S-curve model has slow growth at the initial stage and then undergoes a rapid increase. The reason for the slightly quick growth that starts from the 1800s is that the 1800s was likely to be an intermediate stage rather than the starting point. The earliest data in the COHA dates from the 1800s. However, the emergence of hyphenated compounds should have been seen before the 1800s, and the frequency of some hyphenated compounds should have increased slowly at

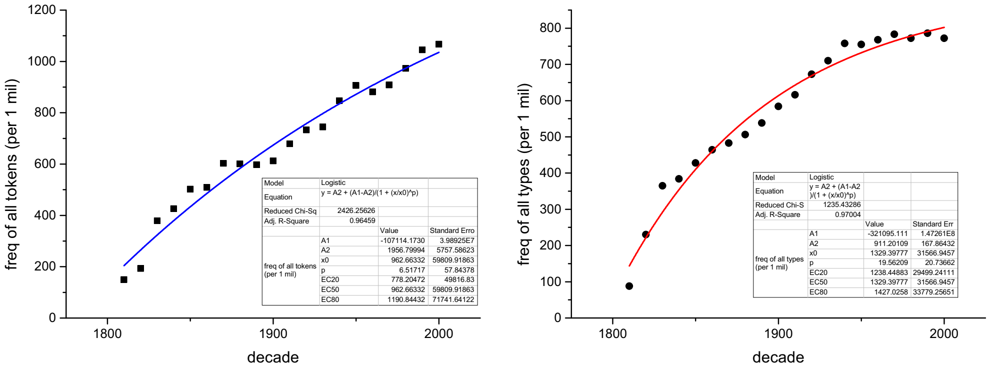

Fig. 2. Fitting the diachronic distribution of the frequency of *all tokens* and *all types* for hyphenated compounds in the DEHC (Here, the frequency is a standard frequency measured per one million words rather than the raw frequency). The R^2 value for the tokens and types are 0.96 and 0.97, respectively. The *p*-value in the two graphs is <0.01. A higher value of R^2 indicates better fitting.

the initial stage (prior to the 1800s). Fig. 2 is not the whole picture because of the unavailability of data that occurred before the 1800s. We have discussed in the Materials and Methods section why the COHA was chosen to obtain the data of word frequency in the DEHC. Despite this, in this case, the change in the token/type frequency of hyphenated compounds still presents an S-curve. Given the missing data that occurred before the 1800s, the whole picture of the frequency distribution might look more like a typical S-curve. The S-curve model of the changes in the frequency of hyphenation compounds might indicate that these compounds took a similar route to the general growth in vocabulary. As a compounding technique, hyphenation was supposed to be slowly accepted by language users at the initial stage. Afterwards, the expansion of hyphenation accelerated.

The data for all tokens and all types can form an S-curve model regarding the change in total frequency. The increases in the type and token frequencies for hyphenated compounds over the last two hundred years are quite significant. Hyphenated compounding was slowly accepted by language users and was initially not used very frequently. Afterwards, the expansion of hyphenated compounds occurred more quickly than previously. The increase in the type frequency for hyphenated compounds began slowly after the 1980s. In contrast, the token frequency of hyphenated compounds seems to continue to increase more quickly than the type frequency after the 1980s. Similarly, the token/type frequency of the three types (according to the composition type) also forms an S-curve model without the availability of the data that occurred prior to the 1800s, which is shown in Fig. 3. The token frequency of the three types also seems to continue to increase more quickly than type frequency after the 1980s. Note that even the 2000s is not the final stage of the growth of hyphenated compounds. In this sense, more time might be needed to see what a larger picture of an S-curve looks like. Additionally, the S-curve shape of the token frequency for coordinate hyphenated compounds is likely to be a Richard's curve (n = 1.4), which is shown in the right panel of Fig. 1 rather than a typical logistic curve. Overall, all of this suggests that the frequency distribution of three types is consistent with the frequency distribution of the entire data, and thus enhances the argument of an S-curve model followed by a frequency distribution in hyphenated compounds. We therefore offer a preliminary conclusion that hyphenation in compounds is motivated.[12]

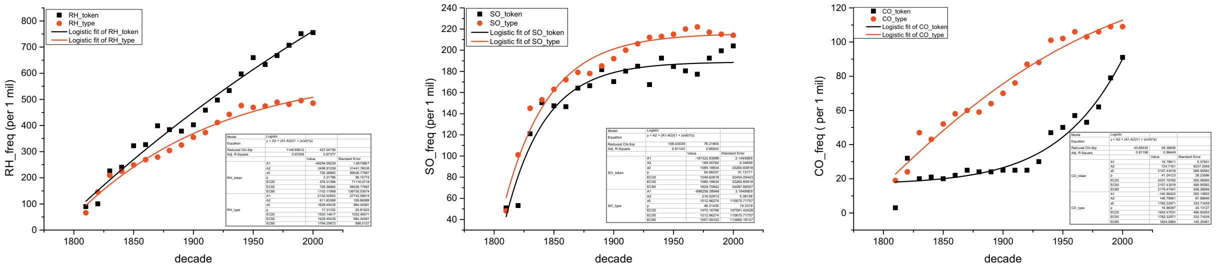

Fig. 3. Fitting the diachronic distribution of the frequency of the three types of hyphenated compounds (per one million words). The three types of hyphenated compounds share great similarity in the distribution of the frequency of *all tokens* and *all types* as shown in Fig. 2. The left panel is RH, the centre is SO and the right is CO. All cases fit a logistic distribution very well. The R^2 value of the three types of *token* are 0.97, 0.91 and 0.91 respectively; The R^2 value of the three types of *type* are 0.97, 0.96 and 0.96 respectively. The *p*-value in all cases is <0.01. A higher value of R^2 indicates better fitting. RH=violating right-headedness; SO = violating syntactic order; CO = coordinate.

[12] Considering that many low-frequency hyphenated compounds have been created over the last thirty years, the question arises regarding what caused the slight deviation from the inverse logistic regression in the distribution of token frequency. In addition, the S-curved model is also seen in the changes in both the type and token frequencies for the close compounds derived from Gagné et al. (2019), as indicated in the supplementary material.

The GAMM fitting was used to examine the effects among the variables. If the tokens (or types) in the DEHC are treated as a response variable, then the other variables, such as *decade* and *types* (or *tokens*) are treated as predictive variables. The GAMM fitting was of great help in analysing whether the effects of these predictive variables occurred, and why the frequency distribution forms an S-curve. The results are shown in the following Fig. 4.

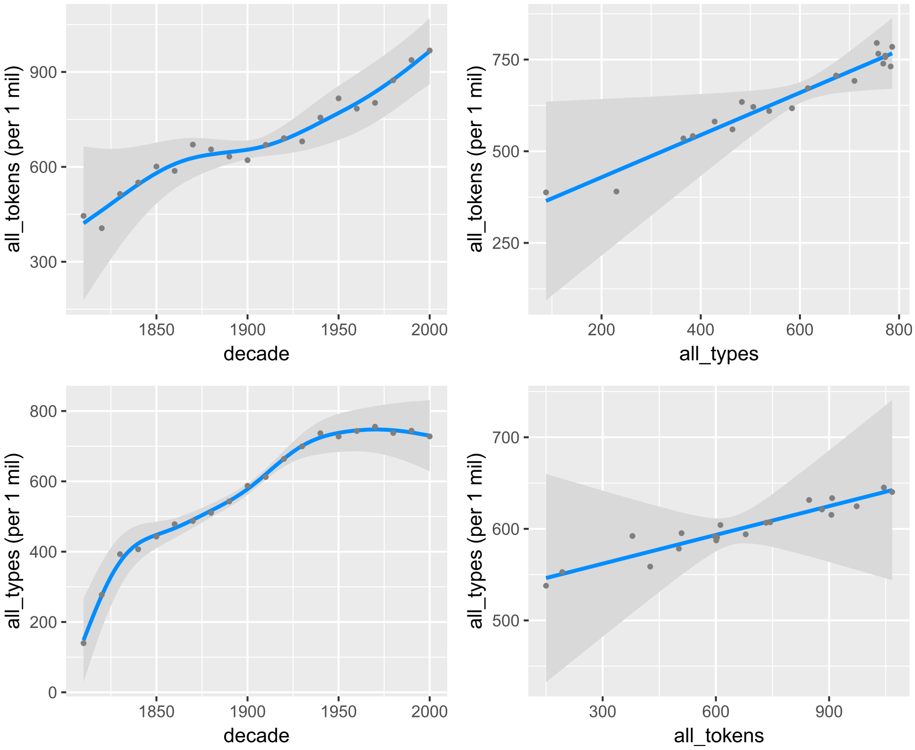

Fig. 4. The effects of the predictive variables on changes of tokens/types in the DEHC.

As shown in Figs. 2 and 3, *decade* has a strongly positive effect on the change in the token/type frequencies. However, Fig. 4 reveals that the *smooth* function of the variable *decade* seems to have a significantly positive effect on either tokens or types, but the smooth effect forms a non-linear curve. The effects of *decade* on *token/type* in Fig. 4 (the left two panels) are consistent with the distribution fitting model of *decade* and *token/type* in Fig. 2 which is nonlinear (i.e. logistic) rather than linear. In contrast, *all tokens* or *all types* have a markedly positive effect on one another for each decade in the DEHC.

4.2. A case study: high-frequency hyphenated compounds

For the second database mentioned above, some high-frequency hyphenated compounds are chosen to examine their changes over the last two hundred years. We chose hyphenated compounds with high frequencies in the same category. Three groups are chosen to enable comparisons.

Compounds in the **first** group share the similarity that these hyphenated adjectives as modifiers of nouns have the structure "* – noun". These hyphenated compounds violate *right-headedness*. The **second** group meets the criterion that the two elements in a compound violate the syntactic combinational order. The **third** group consists of hyphenated expressions that are not compounds, including the prefix and number types, and a few hyphenated compounds that have corresponding non-hyphenated forms.

In addition to high frequency, different types are considered in the same category to enable comprehensive representations. For example, all of the hyphenated types in the second group violate the syntactic order, and we chose different types of combinations according to the PoS of their elements (avoiding repeated types), such as Noun + Adj., Noun + Verb, Noun + V-ed/ing., and Adj. + V-ed/ing. The third group contains several types. Some hyphenated compounds have corresponding non-hyphenated forms. The other types include Suffix + Word and Adj. + Noun as a noun; *Numerical hyphenated expressions* are not compounds although they adhere to the syntactic order and the principle of *right-headedness*. The hyphenated compounds in Groups 1 and 2 violate the *right-headedness* and syntactic order respectively. However, the hyphenated expressions in the third group are not real compounds, and hyphenation has the function of enabling correct parsing.

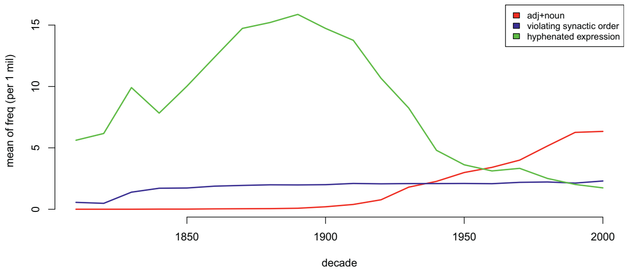

Fig. 5. Historical frequency of the three types of high-frequency hyphenated compounds (the frequency by per million). The difference between *adj+n* and *violating syntactic order* is marginally significant based on an analysis of variance (ANOVA) (p-value = 0.0236 < 0.05); the difference between *adj+n* and *hyphenated expression* is highly significant (*p*-value = 3.66e-05 < 0.001); and the difference between *violating syntactic order* and *hyphenated expression* is slightly significant (*p*-value = 0.029 < 0.05). The Chi-square test result is: *X*-squared = 98.596, df = 38, *p*-value = 2.785e-07 < 0.001. This means that the three groups might be different.

Thirty words for each group (Groups 1 and 2) are chosen from the DEHC according to the ascending rank of their total raw frequency and diversified types (which is shown in Tables 3–5 of the supplementary material). In addition, thirty of the hyphenated expressions from Group 3 are added. They were chosen because of their relatively high frequency in the COHA. We obtained the frequency per million for each word and then calculated the mean of all of the frequencies in each decade for this group (thirty words). We plotted in Fig. 5 the three curves that represent the frequency change of the three groups over twenty decades.

The data in Fig. 5 show that the majority of hyphenated compounds (*adj + noun* curve and *violating syntactic order* curve) occurred after the 1800s. As time passed, the frequency of some compounds increased. The tendency of the two curves also basically follows the general tendency of all hyphenated compounds, as shown in Fig. 2. By contrast, the expressions with inserted hyphens were not real compounds, but were used much earlier than real hyphenated compounds. The three groups clearly exhibit two different trends in their frequency distribution from the perspective of historical changes. In the first and second groups, these hyphenated compounds were produced late but used frequently after the 1900s and have recently become much more frequent. This trend is especially true of the compounds in the first group. In contrast to this tendency, hyphenated expressions in the third group were frequently used before the 1900s, but their frequency has decreased since the 1890s and was overtaken by the first two groups in the 1960s and 1990s respectively.

The ninety hyphenated compounds and expressions in the three groups can reveal a general tendency helping us to attain a better view of the historical changes in the different types of hyphenated compounds previously discussed. The rise in the use of compounds mainly occurred in the 19th century, with an increase and enrichment of the vocabulary and expressions of the English language. The data in Groups 1 and 2 show that hyphens have not disappeared from compounds. Instead, the use of hyphens is currently being reinforced. Conversely it is possible that hyphenation was randomly used for insertion between two (or more) words for various reasons over the course of history. Hyphenation in the group of hyphenated expressions is simply a matter of spelling, and it does not play a role in parsing or reducing ambiguity in Groups 1 and 2. Examples and discussion concerning this argument can be found in the text around Fig. 6, which is why hyphenated expressions in the third group experienced a decrease in frequency since the 1890s. This outcome confirms the hypothesis that if hyphenation is not useful in helping to parse a combination as a compound, then hyphens also tend to disappear. That is, if these hyphenated expressions have a structure that fundamentally resembles the structure of common compounds, hyphens are redundant because they are not economical. This topic is further discussed in the following section.

4.3. *An evolutionary development of orthographic forms?*

This section addresses the question of why hyphenation in compounds can be preserved and further explores whether other orthographic forms will disappear in compounding. This exploration is based on the Google N-gram corpus.

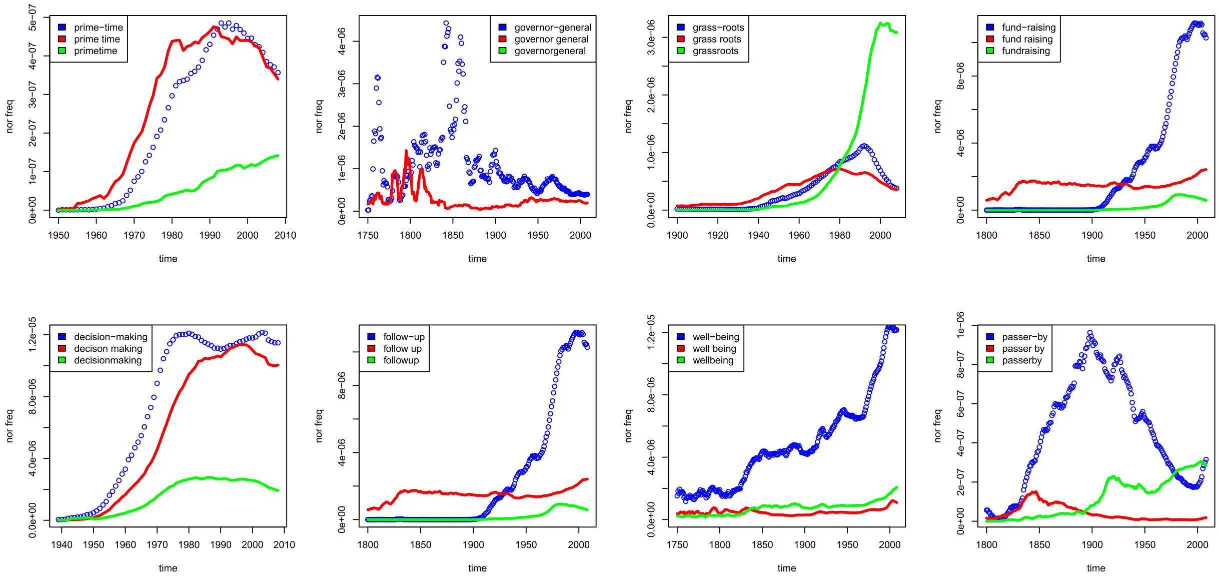

Fig. 6. Change in the frequency of compound nouns in the N-gram corpus.

Compounds have three orthographic forms. However, there is a widespread view that “most compounds prefer one spelling” (Bauer and Renouf, 2001: 450). This view on compound forms is based on the data from three “noun + noun” compound corpora. Obviously, the NOUN + NOUN type of compound, like a small proportion of all compounds overall, cannot provide an adequate basis for substantiating the claim that there is also a preference for one spelling over other types of compounds. Therefore, Bauer and Renouf (2001: 450-451) also noted that the spelling of compounds with an adjective or verbal head seems to exhibit a different type of variability, and hyphenation seems to be preferred in adjectival compounds (compounds that end with an adjective, such as “sugarfree”, in which the PoS of the whole compound is not necessarily adjectival) or verbal compounds (compounds that end with a verb, such as “finger-catch”, in which the PoS of the whole compound is not necessarily a verb). They also suggest that highly frequent forms are likely to become lexicalized, which causes hyphens to disappear. r.

Some research (Liège, 2017) has even claimed that when expressions become more popular or take on special meanings, compounds follow a gradual evolution from two or more separate words (open compounds) or form hyphenated words to single words (closed form, i.e., solid form without a space), such as the following:

audio visual	audio-visual	audiovisual
wild life	wild-life	wildlife

Hyphenation in compounds is believed to be an intermediate form among the three types of compounding forms (Sanchez-Stockhammer, 2018: 348). This assumption seems to be both reasonable and fair, but this popular view ignores the strong motivations for using hyphens in compounds. To further understand the changes in these hyphenated compounds and to determine whether a thesis of this type of evolutionary process is justified, we used the Google N-gram corpus to collect data and examine changes in the frequency of some typical compounds and expressions. Searches were performed simultaneously for the three forms (spaced, hyphenated, and closed forms). For example, after inputting “primetime NOUN, prime time NOUN, primetime NOUN” (a comma is a divider in the Google N-gram Viewer, and “NOUN” searches for the item as a noun), we used these data to generate a graph that shows the three frequency change curves for the three forms with different colors. However, these graphs, which are directly exported from the Google N-gram Viewer, are not high resolution. Therefore, we first ran a Python script to capture the data on a group of words that we wanted to compare. Based on these data, we plotted graphs by using an R script. To perform convenient comparisons, we classified these compounds into several groups according to their features.

The Google N-gram corpus (“English 2009” in the N-gram Viewer) was chosen to examine the historical changes in frequency for these specific combinations because the data that occurred before the 1800s can be traced in the Google N-gram corpus. Although the Google N-gram corpus does not allow wildcard searches with “*”, it supports some specific compound searches. Accordingly, the COHA was used to obtain general data on the DEHC, while the Google N-gram corpus was used to search for the tendencies of changes in specific compounds.

To gain insight into the historical changes in orthographic forms for different types of hyphenated compounds, we examine them one by one in accordance with their PoS or characteristics. We selected a number of compounds just because they are frequent in the COHA, and we also selected an equal number of compounds for each of five categories. In addition, to

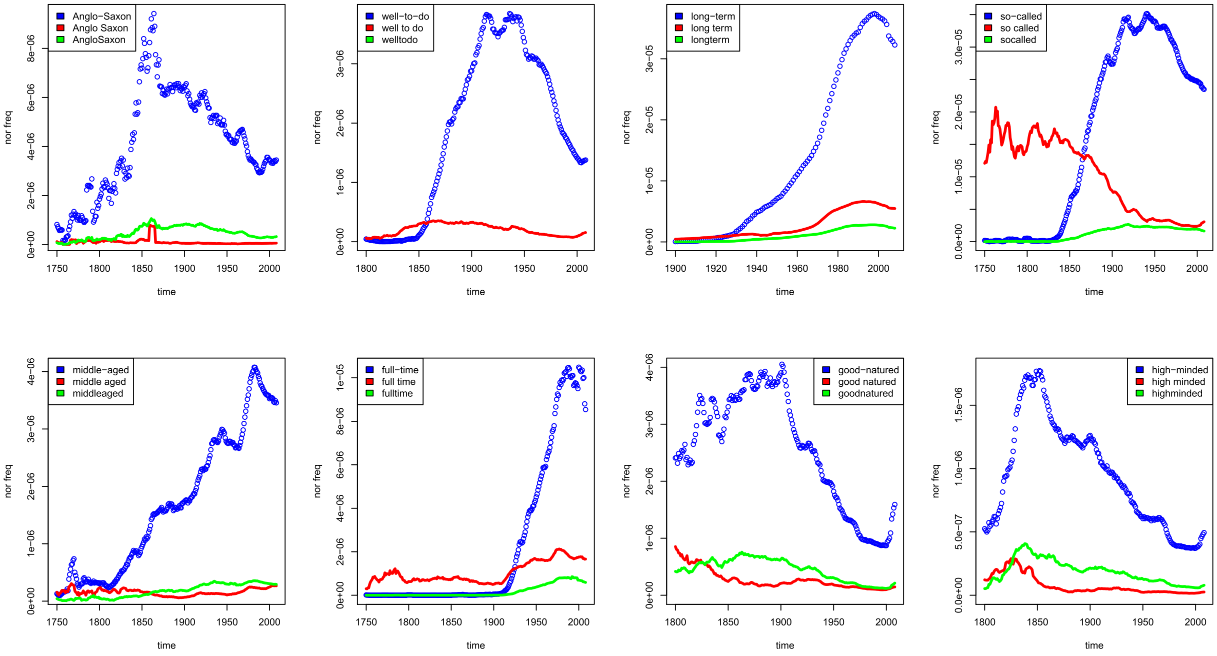

Fig. 7. Change in the frequency of compound adjectives.

test whether strong lexicalization eventually leads to a fixed spelling or not, we contrasted highly frequent and well-entrenched compounds with rarer and less-entrenched compounds. For this purpose, we chose a dozen hyphenated compounds and their counterparts and classified them into five groups, which represent the third type of database in this study, to enable an examination and comparisons.

First, we examined the selected compound nouns. According to Fig. 6, some common hyphenated nouns have decreased in frequency and are likely to decline in number. They will likely be replaced by their corresponding nonhyphenated forms in the future. Regarding these closed compounds, they might have undergone a gradual evolutionary process from open forms to hyphenated forms (or forms without hyphens). It seems reasonable to conclude that open or closed forms will replace hyphenated forms to become their final form, such as “grassroots” vs. “grassroots”. Nonetheless, it cannot be claimed with certainty that a closed form will be the final form.

In contrast, hyphenated compound nouns that violate right-headedness or the syntactic order do not seem to lose their hyphens as easily. For an example with respect to frequency, “fund-raising” has been gaining precedence over “fund raising” over the course of time (“fund raising” can be parsed differently depending on the context; it thus has various usages and is not truly equivalent to “fund-raising”). Such hyphenated compounds are not treated as nouns when hyphens are removed from the verb-phrase-like types such as “take-off, sell-out, wrap-up, and sit-in”. When the hyphen is removed from a coordinate-style hyphenated compound, such as “actor-manager”, the compound is likely to be treated as two words rather than as a compound.

Hyphenated nouns that follow the syntactic order or *right-headedness* (e.g. “prime-time”, “grass-roots”) tended to lose their hyphens more easily than hyphenated compounds that violated the two principles. Most of these hyphenated compound nouns that lost their hyphens easily have corresponding nonhyphenated forms. Moreover, with regard to frequency, these hyphenated compound nouns exhibit no priority over non-hyphenated forms as time passes. If the removal of hyphenation has no negative impact on the correct parsing of these expressions by language users, readers might prefer simplified orthographic forms for reasons of economy. Next, let us examine the compound adjectives.

Fig. 7 shows that hyphenated compound adjectives (which act as modifiers of nouns) that break the principle of *right-headedness* successfully compete with other forms and are therefore unlikely to be replaced by their corresponding forms without hyphens in the future. For example, with its structure violating the principle of *right-headedness*, the two words in “Anglo-Saxon” are in coordination, and this compound enjoys an absolute advantage over its nonhyphenated forms, “Anglo Saxon” and “AngloSaxon”. Hyphens are not always easily lost in phrasal hyphenated compounds, as the example of “well-to-do” shows. Their structure violates the principle of *right-headedness*, and common compound modifiers (very likely to be used

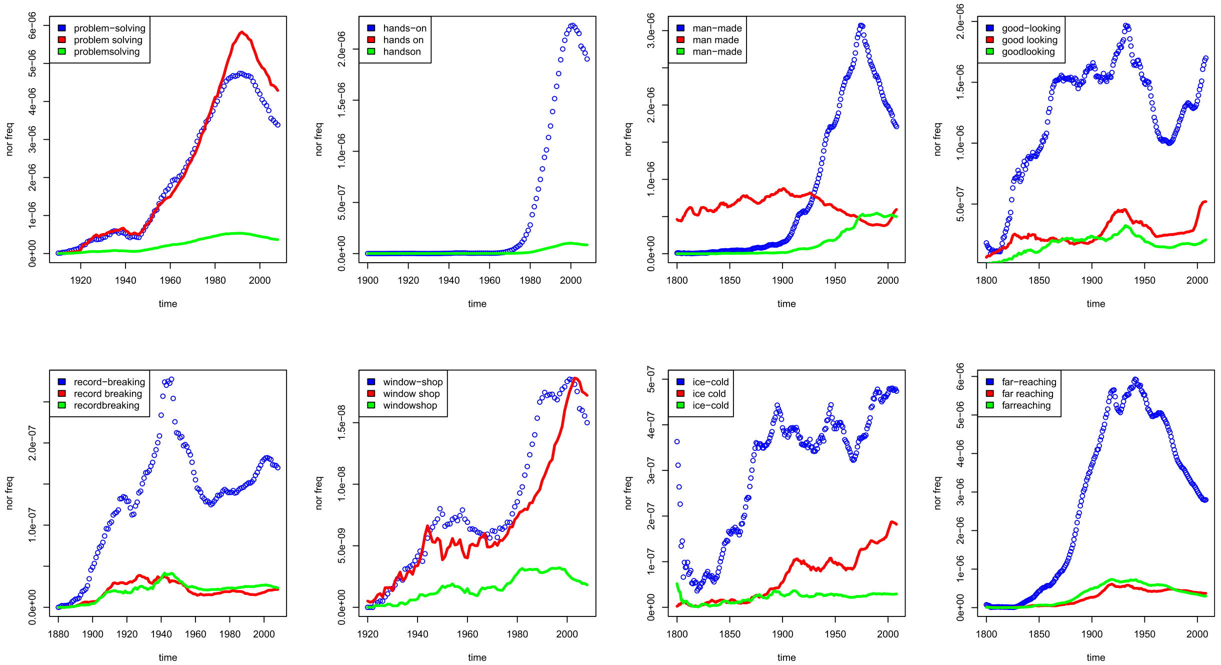

Fig. 8. Change in the frequency of compounds that violate the syntactic order.

attributively), such as "long-term, so-called, and old-fashioned", have a strong tendency towards the use of hyphenated forms. They are highly unlikely to use open or closed forms, as shown in Fig. 7. In this regard, hyphenated compound adjectives and phrasal hyphenated compounds do not tend to follow a gradual evolutionary process. When the hyphens are removed from these compounds, the result can be substantial confusion for language users. The following paragraph adopts another perspective in examining the compounds that violate the syntactic order.

The hyphenated compounds in Fig. 8 that violate the syntactic order show a steady increase in frequency regarding the retention of the hyphenated form when acting as an adjective or an adverb, such as "problem-solving (adj.), hands-on (adv.), good-looking (adj.), window-shop (v.), and man-made (adj.)". When hyphenated compounds act as nouns or verbs, the forms without hyphens are likely to replace the hyphenated forms in the long term, as in "fundraising (noun) vs. fund-raising and windowshop (verb) vs. window-shop".[13]

Nevertheless, the increase and decrease in the frequency of hyphenated nouns and verbs does not necessarily follow the shift from hyphenated to nonhyphenated forms, because in some compounds, the forms without hyphens were created earlier and have dominated consistently throughout the entire period. At the least, the frequency changes in these words cannot directly support the hypothesis of an evolutionary process from the hyphenated form to the open order. or closed form. In the following, we examine some old and new hyphenated expressions.

As shown in Fig. 9, when people became familiar with some of the hyphenated expressions, the hyphens gradually fell out of use, probably for reasons of economy and convenience. For example, the form "to-day" was quite popular in the 19th century, but in the present day, most native readers would accept only "today". "Camp-fire" surpassed "camp fire" or "campfire" before the 1920s, but the last 100 years have witnessed the decline of this hyphenated spelling. Hyphenation in these expressions seems to be slightly redundant because people do not need it to parse these expressions as compounds. Additionally, regarding the expressions in Fig. 9, there are no forms or intermediate stage, such as "vicepresident, to day, North East", at all.

It is possible that people included hyphens in some expressions based on their personal styles or at random. Some hyphenated compound nouns with corresponding non-hyphenated forms can clearly provide direct evidence for this possibility (e.g., "dining-room vs. dining room"). Hyphenated forms are likely to gradually become less favored, as shown in Supplementary Table 5 and Fig. 9. When hyphenation does not provide information for parsing these expressions, people will abandon it gradually for reasons of economy. It is noteworthy that most of these old hyphenated expressions that easily lost their hyphens are nouns.

[13] The verb "window-shop" is closely related to the noun "window-shopping". The point then is that it is likely that highly related forms such as these are likely to use or drop the hyphen together, under analogical pressure. We have therefore examined how many such potential cases are present in our database. It turned out to be 0.34% (three compounds), which can be negligible.

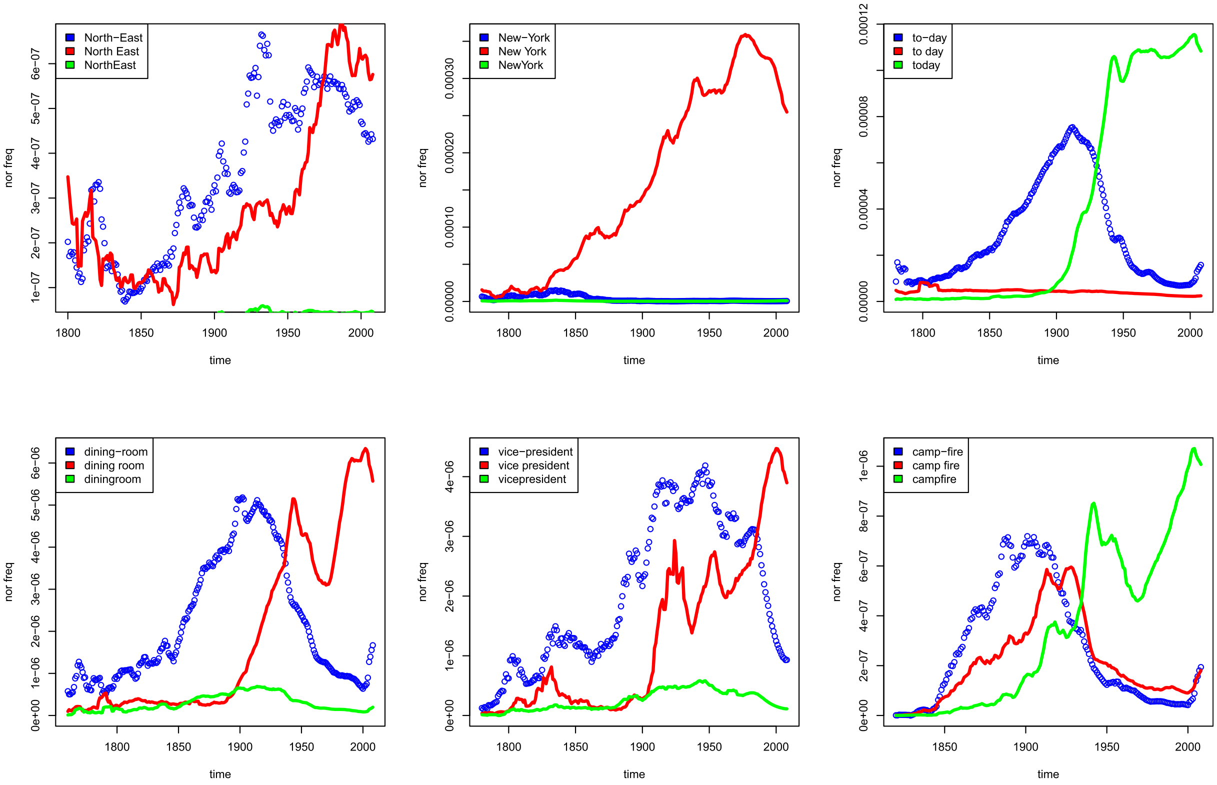

Fig. 9. Change in the frequency of old hyphenated expressions (occurring before 1800s).

These trends mentioned in Figs. 6–9 can also be confirmed by the development of some new compound words that were selected from the word list update of the online Oxford English Dictionary (2018), which is shown in Fig. 10. For example, “full-scale”, a hyphenated compound adjective, is frequently used currently, but people seldom use its non-hyphenated form “full scale/fullscale”. Although “hacky-sack” is a rather new compound noun, the decline in the hyphenated form is clearly seen in recent data since it has been gradually replaced by the form “hacky sack”. Hyphenation in the two compounds plays distinct roles. “Hacky-sack” follows compounding principles, while “full-scale” works as an adjective and violates *right-headedness*. Even without a hyphen in “hacky-sack”, people are still able to parse it correctly. In contrast, the combination of “full-scale” is very likely to be parsed incorrectly in the context without hyphenation.

Therefore, hyphenated compound adjectives that modify noun and phrasal compounds do not easily lose their hyphens in their orthographic form, as revealed from a historical perspective regarding this phenomenon. In contrast, hyphenated compound nouns are more likely to be replaced by their non-hyphenated forms over the course of history. Hyphenated nouns that violate *right-headedness* or the syntactic order do not seem to lose their hyphens as easily. This might be implicitly consistent with the fact that hyphenated nouns that follow *right-headedness* and syntactic principles constitute almost 7% of all hyphenated compound nouns. All of these outcomes can be ascertained by using the data from the three corpora mentioned by Bauer et al. (2013). However, it is difficult to judge whether close orthographic forms for compound nouns are privileged in their competition with open forms.

Accordingly, the data from the three types of databases support the hypothesis that the use of hyphenation is strongly motivated in hyphenated compounds. The majority of hyphenated compounds used frequently in English have been filtered and selected over the course of history. Most of the hyphenated compounds (98% in the DEHC) used frequently in English violate *right-headedness* or the syntactic order. Moreover, it is interesting to note that when hyphenation is used at random in some expressions or compounds, people tend to gradually abandon the habit of using hyphens for reasons of economy, as discussed above. The diachronic analysis confirms that the long-term use of hyphens in hyphenated compounds is strongly motivated. We further assume that hyphenation might provide special information to readers to confirm that they are compounds–a point further discussed in the following section. Generally, all the evidence supports the hypothesis that hyphenation has gradually evolved into an effective compounding technique.

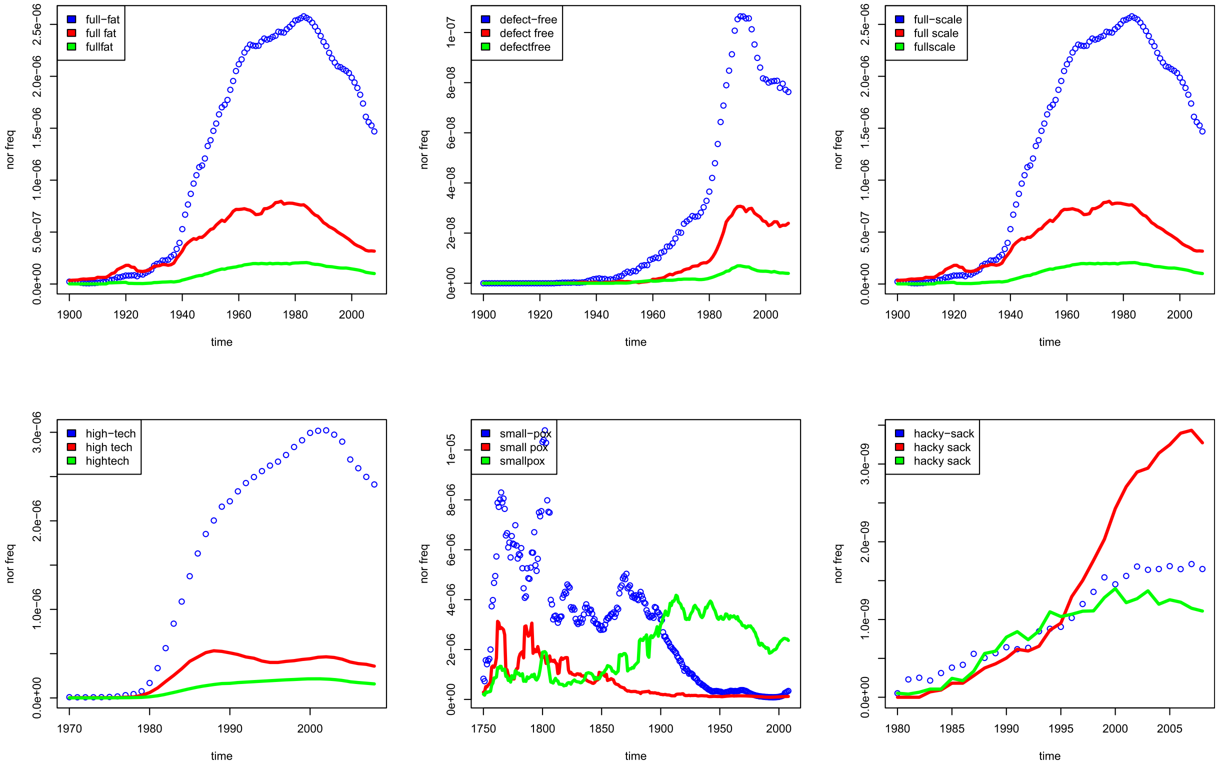

Fig. 10. Change in the frequency of new compounds.

5. Discussion: language users and discriminative learning

Given the finding that hyphenated compounds do not lose their hyphens easily, although it seems plausible that hyphens aid in decoding a compound, it is an open question regarding whether hyphenation is also reader-friendly or is used instead as a writer-friendly device.

Kuperman and Bertram (2013) held that economy appears to underlie the preference for certain spelling formats because they best fit the strategies that readers employ when processing complex written words. The processing disadvantage in recognizing a concatenated compound that is more likely to occur in another spelling is on par with the effects of word frequency and length. Therefore, Kuperman and Bertram (2013) posited that language structure is shaped through language use, i.e., the degree of familiarity with compounds determines their orthographic forms. Precisely when users become more familiar with compounds, they are more likely to discard the hyphenation. The fading of hyphenation in some compounds supports the view that users will abandon a hyphen within a compound for reasons of economy as they become familiar with the compounding forms.

However, the diachronic investigation in this study reveals that many hyphenated compounds have not lost their hyphens even when people have become familiar with them over a period of two hundred years. Reasons of familiarity and economy might not adequately explain this finding.

Considering the different characteristics of hyphenated compounds, we therefore provide a different perspective to adequately account for hyphenation, which can be treated as a discriminative signal to distinguish from the other types of compounds. People might acquire the rules of hyphenation compounds by learning this discriminative signal. The state-of-the-art theory of the discriminative learning model might explain how humans process hyphenation in compounds.

Modern learning theory begins with Ivan Pavlov and his famous observations about bells and dog food (Pavlov, 1927). Pavlov's initial results led to the formation of a straightforward theory of (animal) learning: if a cue is present and an outcome follows, then an animal notices the co-occurrence and subsequently learns to associate the two. Based on Pavlov's discovery, Rescorla and Wagner (1972) developed learning equations that perfectly capture and formalize this finding. Their learning theory (Rescorla and Wagner 1972) not only predicts a substantial body of findings in the animal literature but also has recently been found to predict aspects of first language acquisition and implicit learning in adults (Ramscar et al., 2010, 2013). Based on these learning equations, Baayen et al. (2011) proposed a discriminative learning model that posited that humans learn words and process lexical items based on the deviation between an observed event and the predicted event.

A discriminative learning model, as a computational model for processing words, has a simple structure with only two layers (which avoids the hidden layers in neural networks). Discriminative learning for lexical processing typically uses large numbers of simple cues, such as letter pairs or letter triplets, but cues can also be words, acoustic features, or constructional properties. Similarly, outcomes can range from lexical and grammatical features to idioms and constructions. With this approach, the morphology and syntax are implicit in the distribution of cues and outcomes that are collected and serve as outcomes. These results show that it might indeed be possible to proceed in an end-to-end way from sublexical features (letter pairs or triplets) to a conceptual interpretation without mediation by morphemes and word forms. Inspired by information theory (Shannon, 1948), discriminative learning views the signal as a code that has evolved to discriminate between the messages that the sender and the receiver exchange. The discrimination of messages proceeds based on the signal itself, without mediation by the layers of “hidden” units of phonemes, morphemes, and word forms.

The discriminative perspective indicates that language production can likely be viewed as the process of using this system to construct a message that best represents a language user's intended meaning. Language comprehension is simply understood as using a simpler signal to predict the intended meaning. The primary function of a *cue* is to reduce uncertainty in language comprehension and obtain production at the lowest cost. The discriminative approach to language learning, comprehension and production is efficient (Ellis, 2006; Hsu et al., 2011; Baayen et al., 2019), and it can explain the function of hyphenation in compounding. A hyphen is a cue, the outcome is the combination of the compound acquisition of new semantic and syntactic functions. The hyphen cue is closely associated with the function of the whole compound. The other function of hyphenation is to discriminate this type from other types of compounds, which helps to reduce uncertainty and confusion with other types of compounds.

The discriminative learning model is actually one model that might be helpful in interpreting how humans communicate efficiently. An efficient use of language might enable a speaker to transmit many different messages successfully with minimal effort (Gibson et al., 2019). In the discriminative learning model, a cue, which usually encodes a simpler code (form, string, content, etc.), can be associated with a relatively complex outcome (information or content), i.e., a simpler form can predict the full information about the associated item. This model, which accords with the principle of communication efficiency, shows that this form of language enables successful communication while minimizing effort. For this reason, the discriminative learning model can explain many phenomena concerning language comprehension, learning and production.

From the perspective of the discriminative model, hyphenation might be useful in making English language users unconsciously aware that hyphenated compounds are quite unique and in reminding readers not to interpret and analyse them with the methods for parsing common compounds. The removal of hyphenation indicates the elimination of this discriminative function and could easily lead to misunderstanding through readers' internalized knowledge of compounding rules. For example, “high-quality” as a compound adjective is easily parsed into two separated parts without hyphenation; according to the principle of *right-headedness*, the compound can be analysed as a noun fairly easily. However, hyphenation, as a discriminative cue, is very helpful in reminding users of the unconventional nature of its constituent structure. As a cue, the existence of the hyphen in compounds actually aids in distinguishing the outcome and in showing that these compounds were created according to rules that are distinct from general compounding principles. The interpretation of the discriminative model can be supported by another perspective of *markedness*. Markedness in language refers to an extra sign to indicate unusual language constructions, which allows an easy distinction and processing in cognition (Andrews, 1990; Herbert, 2011). In linguistics, *markedness* can apply to, among others, phonological, grammatical, and semantic oppositions, such as “lion/lioness, child/children, old/young”, etc. Similarly, an extra mark (hyphen) is needed to distinguish hyphenated compounds (marked) and the other types of compounds (less marked) such that English readers use them more conveniently and efficiently. Although the two perspectives are consistent, the discriminative model is prone to cognitive and computational processes compared with the markedness theory that emphases linguistic interpretation.

Even when language users are very familiar with these compounds, hyphenation has not disappeared because its role in providing a discriminative cue is quite crucial in signalling the unconventional nature of such compounds, that is, being contrary to the language user's internalized knowledge of compounding principles. The other reason is that these unusual compounding operations can boost productivity and be easily used. This technique of hyphenation has successfully resulted in communication efficiency and has been mastered by language users. Likely in some cases, using quotation marks around a phrase (or underlining a phrase) can cause the phrase to become a temporary word. This use of other graphemic symbols shares a function similar to that of hyphenation. When no graphemic symbol is used, it is difficult to ensure that readers will treat a phrase as a word. In this sense, language users have great flexibility and freedom in creating new compounds by using this technique and can cause a phrase to become a new (probably temporary) word simply by adding hyphens to enable a correct parsing of the word.[14]

Generally, discriminative learning actually explains how a binary system works for the same phenomenon of language. With respect to compounding, both hyphenation and general compounding work effectively. However, an additional mark/symbol is needed to distinguish them to help English readers use them more conveniently and efficiently. Accordingly, hyphenation functions as a useful technique in compounding. Hyphenation is reader-friendly and can also be used as a writer-friendly device. The frequency of hyphenated compounds continues to increase, but it is unclear what will ultimately occur when the hyphenation compounding technique competes with general compounding rules. A longer period of historical change might be needed after all.

[14] Newly created compounds with hyphens are easy to see in magazines or newspaper articles for a discerning reader. Although some of these hyphenated compounds express temporary meanings, most of them accord with the results of the analysis in this study.

6. Conclusion

This study adopts the following two new perspectives that have not been considered in the academic literature: contextual factors (the PoS to which a compound as a whole belongs and how people correctly parse a compound into a unit) and the diachronic frequency of hyphenated compounds. Through quantitative and diachronic investigations of hyphenated compounds in English from three databases, we verified the hypothesis proposed in this study. Hyphenation has gradually become a compounding technique from a diachronic perspective. Hyphenated compounding, as a useful morphological technique, has been widely applied by readers. This fact has unfortunately been ignored and underestimated in past studies. We found that the change in the frequency of hyphenated compounds basically follows the S-curve model of language changes. Furthermore, hyphenation in compounds is likely to be preserved in language. However, hyphenated expressions tend to lose their hyphens when language users become familiar with them. Since the use of hyphens in compounds is strongly motivated rather than being used randomly, hyphenation in compounds has become prevalent over the course of history, or in some cases, hyphenation in hyphenated expressions has been abandoned for reasons of economy. We also developed an adequate account of why hyphenation is useful for language users. From the perspective of discriminative learning, hyphenation works as a discriminative cue and thus helps people unconsciously know that these hyphenated combinations are quite different from common compounds. Because it helps in achieving communication efficiency to some extent, hyphenation in compounding can increase productivity and convenience.

In the future, more quantitative investigations will be conducted to compare hyphenated compounds and the other types of compounds while considering additional parameters, such as the lexical family size, semantic similarity between components, and semantic neighborhood. Such studies might provide direct evidence that demonstrates the difference between hyphenated compounds and other compounds and that reveals how native speakers process hyphenation in compounds differently.

Acknowledgments

The earlier version was presented in Conference of Mental Lexicon (2018, Edmonton, Canada). We express our sincere gratitude to those who gave us comments to improve this manuscript in this conference. We would like to thank the three reviewers (particularly the first reviewer) for their insightful and constructive comments on the paper, as these comments led us to an improvement of the work. The second author received the ERC (European Research Council) advanced grant (grant No. 742545).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.langsci.2020.101326.

References

Altmann, Gabriel, 1992. What is language synergetics? Series B: humaniora, 16., chapter Piotrowski's Law of language change. Oulu: Acta Univ. Ouluensis, 34–35.

Andrews, Edna, 1990. Markedness Theory. Duke University Press, Durham, NC.

Baayen, Harald R., Milin, Petar, Durdevic̆, Filipovic̆, Hendrix, Peter, Marelli, Marco, 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. Psychol. Rev. 118 (3), 438–481. https://doi.org/10.1037/a0023851.

Baayen, Harald R., Chuang, Yu-Ying, Shafaei-Bajestan, Elnaz, Blevins, James, P., 2019. The discriminative lexicon: a unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. Complexity 2019. https://doi.org/10.1155/2019/4895891.

Bauer, Laurie, 1983. English Word-formation. Cambridge University Press, Cambridge.

Bauer, Laurie, 2008. English exocentric compounds. In: Amiot, Dany (Ed.), La Composition Dans Uneperspective Typologique. Artois Presses Université, Arras, pp. 35–47.

Bauer, Laurie, 2017. Compounds and Compounding. Cambridge University Press, Cambridge.

Bauer, Laurie, Renouf, Antoinette, 2001. A corpus-based study of compounding in English. J. Engl. Ling. 29 (2), 101–123. https://doi.org/10.1177/00754240122005251.

Bauer, Laurie, Lieber, Rochelle, Plag, Ingo, 2013. The Oxford Reference Guide to English Morphology. Oxford University Press, Oxford.

Benczes, Réka, 2004. On the analysability of English exocentric compounds. Jezikoslovlje 5 (1,2), 1–21.

Benczes, Réka, 2005. Metaphor-and metonymy-based compounds in English: a cognitive linguistic approach. *A*. Ling. Hung. 52 (2–3), 173–198. https://doi.org/10.1556/aling.52.2005.2-3.3.

Benczes, Réka, 2014. Repetitions which are not repetitions: the non-redundant nature of tautological compounds. Engl. Lang. Ling. 18 (3), 431–447. https://doi.org/10.1017/S1360674314000112.

Benczes, Réka, 2015. Are exocentric compounds really exocentric? SKASE J. of Theor. Linguist 12 (3), 54–74.

Berg, Thomas, 2011. The modification of compounds by attributive adjectives. Lang. Sci. 33 (5), 725–737. https://doi.org/10.1016/j.langsci.2011.05.001.

Blythe, Blythe R., Croft, William, 2012. S-curves and the mechanisms of propagation in language change. Language 88 (2), 269–304.

Booij, Geert, 2012. The Grammar of Words: An Introduction to Linguistic Morphology. Oxford University Press, Oxford.

Books, Google, 2016. Google Books Ngram Data. Available online at. https://books.google.com/ngrams.

Davies, Mark, 2010. The Corpus of Historical American English (COHA). Available online at: https://corpus.byu.edu/coha/.

Denison, David, 2003. Log (ist) ic and simplistic s-curves. In: ickey, Raymond (Ed.), Motives for Language Change. Cambridge University Press, Cambridge, pp. 54–70.

Dictionary, Oxford English, 2018. New Words in Update. Available online at. http://public.oed.com/the-oed-today/recent-updates-to-the-oed/december2016-update/new-words-list-december-2016.

Diessel, Holger, 2007. Frequency effects in language acquisition, language use, and diachronic change. New Ideas Psychol. 25 (2), 108–127. https://doi.org/10.1016/j.newideapsych.2007.02.002.

Dressler, Wolfgang, 2007. Compound types. In: Libben, Gary, Jarema, Gonia (Eds.), The Representation and Processing of Compound Words. Oxford University Press, Oxford, pp. 23–44.

Ellis, Nick C., 2006. Selective attention and transfer phenomena in l2 acquisition: contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. Appl. Linguist. 27 (2), 164–194. https://doi.org/10.1093/applin/aml015.

Fehringer, Carol, 2012. The lexical representation of compound words in English: evidence from aphasia. Lang. Sci. 34 (1), 65–75. https://doi.org/10.1016/j.langsci.2011.06.001.

Gagné, Christina L., Spalding, Thomas, Schmidtke, Daniel, 2019. LADEC: the large database of English compounds. Behav. Res. Methods 51 (5), 2152–2179. https://doi.org/10.3758/s13428-019-01282-6.

Galani, Alexandra, Hicks, Glyn, Tsoulas, George, 2011. Morphology and its Interfaces. John Benjamins, Amsterdam.

Gibson, Edward, Futrell, Richard, Piandadosi, T. Setevn, Dautriche, Isabelle, Mahowald, Kyle, Bergen, Leon, Levy, Roger, 2019. How efficiency shapes human language. Trends Cognit. Sci. 23 (5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003.

Gordon, Peter C., Hendrick, Randall, Levine, William H., 2002. Memory-load interference in syntactic processing. Psychol. Sci. 13 (5), 425–430. https://doi.org/10.1111/1467-9280.00475.

Gordon, Peter C., Hendrick, Randall, Johnson, Marcus, Lee, Yoonhyoung, 2006. Similarity-based interference during language comprehension: evidence from eye tracking during reading. J. Exp. Psychol. Learn. Mem. Cogn. 32 (6), 1304. https://doi.org/10.1037/0278-7393.32.6.1304.

Herbert, Robert K., 2011. Language Universals, Markedness Theory, and Natural Phonetic Processes. Walter de Gruyter, Berlin.

Hilpert, Martin, 2015. From hand-carved to computer-based: noun-participle compounding and the upward strengthening hypothesis. Cognit. Ling. 26 (1), 113–147. https://doi.org/10.1515/cog-2014-0001.

Hsu, Anne S., Chater, Nick, Vitányi, Paul M.B., 2011. The probabilistic analysis of language acquisition: theoretical, computational, and experimental analysis. Cognition 120 (3), 380–390. https://doi.org/10.1016/j.cognition.2011.02.013.

Hunston, Susan, Francis, Gill, 2000. Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English. John Benjamins, Amsterdam.

Jarema, Gonia, 2006. Compound representation and processing: a cross-language perspective. In: Libben, Gary, Jarema, Gonia (Eds.), The Representation and Processing of Compound Words. Oxford University Press, Oxford, pp. 45–70.

Jurafsky, Daniel, 1996. A probabilistic model of lexical and syntactic access and disambiguation. Cognit. Sci. 20 (2), 137–194. https://doi.org/10.1207/s15516709cog2002_1.

Katsanevakis, Stelios, 2006. Modelling fish growth: model selection, multi-model inference and model selection uncertainty. Fish. Res. 81 (2–3), 229–235. https://doi.org/10.1016/j.fishres.2006.07.002.

Koya, Purnachandra Rao, Goshu, Ayele Taye, 2013. Generalized mathematical model for biological growths. Open J. Model. Simulat. 2013 (1), 42–53. https://doi.org/10.4236/ojmsi.2013.14008.

Kuperman, Victor, Bertram, Raymond, 2013. Moving spaces: spelling alternation in English noun-noun compounds. Lang. Cognit. Process. 28 (7), 939–966. https://doi.org/10.1080/01690965.2012.701757.

Mondorf, Britta, 2009. How lexicalization reflected in hyphenation affects variation and word formation. In: Dufter, Andreas, Fleischer, Jürg, Seiler, Guido (Eds.), Describing and Modeling Variation in Grammar. Walter de Gruyter, Berlin, pp. 361–388.

Li`ege, Uni, 2017. Compound Word. Available online at: http://promethee.philo.ulg.ac.be/engdep1/download/defiswitt/doc/compounds.pdf.

Ljung, Ljung, 1976. -ed adjectives revisited. J. Linguist. 12 (1), 159–168.

Masini, Francesca, 2009. Phrasal lexemes, compounds and phrases: a constructionist perspective. Word Struct. 2 (2), 254–271. https://doi.org/10.3366/e1750124509000440.

McArdle, John J., Ferrer-Caja, Emilio, Hamagami, Fumiaki, Woodcock, Richard W., 2002. Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. Dev. Psychol. 38 (1), 115–142. https://doi.org/10.1037/0012-1649.38.1.115.

Michel, Jean-Baptiste, Shen, Yuan Kui, Aiden, Aviva Presser, Veres, Adrian, Gray, Matthew K., The Google Books Team, et al., 2011. Quantitative analysis of culture using millions of digitized books. Science 331 (6014), 176–182. https://doi.org/10.1126/science.1199644.

Nevalainen, T., 2015. Descriptive adequacy of the s-curve model in diachronic studies of language change. In: Sanchez-Stockhammer, Christina (Ed.), Can We Predict Linguistic Change? Varieng, Helsinki. Available online at: http://www.helsinki.fi/varieng/series/volumes/16/nevalainen/.

Pavlov, Ivan P., 1927. Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex. Oxford University Press, Oxford.

Peter, Ackema, Neeleman, Ad, 2010. The role of syntax and morphology in compounding. In: Scalise, Sergio, Vogel, Irene (Eds.), Cross-Disciplinary Issues in Compounding. John Benjamins, Amsterdam, pp. 21–36.

Piotrowski, Anna A., Piotrowsk, Rajmond G., 1974. Matematiceskie modeli diachronii i tekstoobrazovanija. Statistica Reči I Avtomatičeskij Analiz Teksta. Nauka, Leningrad, pp. 361–400.

Pirrelli, Vito, Guevara, Emiliano, Baroni, Marco, 2010. Computational issues in compound processing. In: Scalise, Sergio, Vogel, Irene (Eds.), Cross-Disciplinary Issues in Compounding. John Benjamins, Amsterdam, pp. 271–286.

Plag, Ingo, 2018. Word-formation in English. Cambridge University Press, Cambridge.

Ramscar, Michael, Yarlett, Daniel, Dye, Melody, Denny, Katie, Thorpe, Kirsten, 2010. The effects of feature-label-order and their implications for symbolic learning. Cognit. Sci. 34 (6), 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x.

Ramscar, Michael, Dye, Melody, McCauley, Stewart M., 2013. Error and expectation in language learning: the curious absence of “mouses” in adult speech. Language 89 (4), 760–793. https://doi.org/10.1353/lan.2013.0068.

Rescorla, R.A., Wagner, A.R., et al., 1972. A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, Abraham H., Prokasy, William Frederick (Eds.), Classical Conditioning II: Current Research and Theory. Appleton-Century-Crofts, pp. 64–99.

Richards, F.G., 1959. A flexible growth function for empirical use. J. Exp. Bot. 10 (2), 290–301. https://doi.org/10.1093/jxb/10.2.290.

Sanchez-Stockhammer, Christina, 2018. English Compounds and Their Spelling. Cambridge University Press, Cambridge.

Scalise, Sergio, Vogel, Irene, 2010. Cross-Disciplinary Issues in Compounding. John Benjamins, Amsterdam.

Shannon, ClaudeE., 1948. A mathematical theory of communication. Bell System Tech. J. 27 (3), 379–423.

Stadler, Kevin, Blythe, Richard A., Smith, Kenny, Kirby, Simon, 2016. Momentum in language change: a model of self-actuating s-shaped curves. Lang. Dynam. Change 6 (2), 171–198. https://doi.org/10.1163/22105832-00602005.

Takehisa, Tomokazu, 2017. Remarks on denominal-ed adjectives. In: Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation. The National University (Phillippines), pp. 196–205.

Tjørve, Even, Tjørve, Kathleen M., 2010. A unified approach to the Richards-model family for use in growth analyses: why we need only two model forms. J. Theor. Biol. 267 (3), 417–425. https://doi.org/10.1016/j.jtbi.2010.09.008.

Turenne, Nicolas, 2010. Modelling noun-phrase dynamics in specialized text collections. J. Quant. Ling. 17 (3), 212–228. https://doi.org/10.1080/09296174.2010.485447.

Van Dyke, Julie A., McElree, Brian, 2011. Cue-dependent interference in comprehension. J. Mem. Lang. 65 (3), 247–263. https://doi.org/10.1016/j.jml.2011.05.002.

Wälchli, Bernhard, 2005. Co-compounds and Natural Coordination. Oxford University Press, Oxford.

Williams, Edwin, 1981. Argument structure and morphology. Ling. Rev. 1 (1), 81–114. https://doi.org/10.1515/tlir.1981.1.1.81.

Wood, Simon, 2015. Package ‘mgcv’. R Package Version 1, 29.

Wood, Simon, 2017. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC, London.

Wu, Junhui, He, Qingshuan, Feng, Guangwu, 2016. Rethinking the grammaticalization of future be going to: a corpus-based approach. J. Quant. Ling. 23 (4), 317–341. https://doi.org/10.1080/09296174.2016.1226427.