English Language and Linguistics, 26.2: 413–442.
doi:10.1017/S1360674321000344

An exemplar-based approach to composite predicates in the history of American English

JOHN D. SUNDQUIST
Purdue University
(Received 15 April 2020; revised 1 August 2021)

This study examines composite predicates (CPs) in the history of American English and uses an exemplar-based model to explain changes in the frequency of verb–noun pairings over time. Two different types of verb-nominal CPs are considered, including those like *take a look*, in which a light verb occurs with an abstract nominal object, and others like *lose sight*, with a more lexically specific verb. Using a corpus of texts written between 1820 and 2009, I track the frequency of different CPs and analyze several families of semantically related nouns that occur with the same verb (e.g. *take a look*, *peak*, etc.). Representative families are analyzed to determine the presence of highly frequent verb–noun pairings, or exemplars, that separate themselves over time. The success of exemplars is evaluated according to several factors that may shape patterns of use, including the relative size of noun families, the frequency band of tokens of each family and the distribution of tokens across types within a family. Results indicate that the two types of CPs differ with respect to the evolution and success of exemplary verb–noun pairings and indicate that frequency bands play a role while the size of the noun family and their distributional patterns do not.

Keywords: frequency, productivity, composite predicates, light verbs, exemplars

1 Introduction

This study is a corpus-based analysis of multi-word verbal expressions known as composite predicates (CPs) and their historical development in American English from the nineteenth century to the early twenty-first century. One of the most common types of CPs contains a polysemous transitive verb or light verb (e.g. *take* or *make*), which often combines with an abstract NP complement that carries the semantic weight of a verb (e.g. *take a look* and *make an assumption*). However, verbo-nominal CPs come in many shapes and sizes, including those with 'heavier', more lexically specific verbs that are paired with a variety of NP complements (e.g. *bear witness* or *lose sight*). CPs with light verbs and those with more lexically specific verbs generally share a number of characteristics but differ in several important ways, including varying degrees of idiomaticity and syntactic flexibility (Brinton 2008: 34).

Following the assumptions laid out for an exemplar-based Construction Grammar model in Bybee & Eddington (2006), Taylor (2012) and Bybee (2006, 2013) in this

study, I evaluate the extent to which some verb + complement pairings function as exemplars in the formation of other similar pairings with the same verb. Exemplars are central, highly frequent constructions, words or phrases that serve as a basis for grouping similarities together. Thus, a CP like *take a look* serves as an exemplar for pairings of this verb with other semantically similar nouns (e.g. *glance*, *peek*, *gander,* etc.). As a point of departure in the discussion on exemplars, I focus on the differences between types of verbs in CPs, comparing those with light verbs like *make* or *take* with those that contain more lexically specific verbs like *bear* and *lose*. The main research question of the study is: are light verbs in CPs more likely to generate exemplars than more lexically specific verbs?

Focusing on the differences between types of CPs, I aim to identify factors that test which verbs in CPs are more or less likely to generate exemplars. These factors include the relative size of each family of semantically similar nouns that occur in complements with each verb, the token frequency bands of these families, and the overall distributional pattern of frequencies within families of similar nouns. Nominal pairings with two of the most common light verbs are considered, namely, *take* and *make*, and compared with those that occur with two typical examples of lexically specific verbs, *bear* and *lose.*

The article is organized as follows: in section 2, I provide an overview of terminology and classification of different CPs before discussing how they relate to exemplars within a diachronic, usage-based Construction Grammar model. In section 3, I describe the corpus-based methodology and data collection procedures, including an overview of ways to identify semantically related nouns in NP complements to verbs within CPs. After presenting the empirical results in section 4, I discuss these findings in section 5 in terms of their relevance to exemplars and the ways in which the factors outlined above affect entrenchment, productivity and coverage. Conclusions will address the diachronic implications of these findings for our understanding of how various CP types generate exemplars and follow different trajectories that may be attributed to grammaticalization and lexicalization over time.

2 Background

2.1 Classification of composite predicates

Although there is little consensus in previous literature on ways to classify CPs, there is general agreement on some basic assumptions.[1] Most studies converge on the principle that verbo-nominal phrases such as *make an assumption* are CPs that consist of multiple grammatical elements, with the verb's NP complement expressing the verbal

[1] See studies in Brinton & Akimoto (1999) or Claridge (2000) for an overview of the terminology associated with composite predicates in studies of earlier stages of English and the relationship between the verbo-nominal type discussed here and the broader category of multi-word verbal expressions that include phrasal verbs (e.g. *take off*), complex prepositions (e.g. *in relation to*), or verb–adjective combinations (e.g. *make sure*).

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

action of the phrase. Moreover, there is general agreement that verbs in CPs are in some way different from full verbs; they have been described as semantically deficient versions of full verbs (Bowern 2008: 163) whose 'contribution to the meaning of the predication is relatively small in comparison with that of their complements' (Huddleston & Pullum *et al.* 2002: 290). The most extensively discussed CP of this type in English involves verbs usually referred to as light verbs (e.g. *do*, *give*, *have*, *make* and *take*).[2] While many studies of CPs focus on only one or more of these five frequent verbs, like Brinton & Akimoto (1999) or Elenbaas (2013), others like Allerton (2003) and Ronan (2014) follow Quirk *et al.* (1985) and assume that lightness is a matter of degree. Allerton (2003: 187–91) includes verbs of medium token frequency (e.g. *feel*, *find*, *grant*) and low frequency (e.g. *add*, *lodge*, *launch*), alongside the high-frequency light verbs. Whether or not both light and heavier verbs should be categorized together, all of these verbs share characteristics of being relatively polysemous and occurring with a variety of NPs that carry the essential meaning of an action or event. In terms of the NP complement, many CPs include deverbal nouns (e.g. *take a walk*) or nouns that are derivationally related to a verb (e.g. *make a decision*) or other nouns that, as Quirk *et al.* (1985: 751) point out, share the characteristic that the NP complement contributes to the essential meaning of the CP (e.g. *make an effort*).[3]

As discussed in Brinton & Traugott (2005: 130) and Brinton (2008: 44), CPs generally fall into two broad categories. On the one hand, *take-a-look* CPs are often pairings of a light verb and an eventive noun. NP complements of this type are diverse, including some that are zero derivations (*take a walk*), bare NPs (*take care*), plurals (*make choices*), some that are preceded by indefinite articles (*make a wish*) or optionally by definite articles (*make the/a choice*), and some that display the possibility of adjectival modifiers (*take a (quick) look*). Moreover, the parts of this type of CP are, to some extent, interchangeable, on account of the general semantics of the verb, as in *take* or *have a drink* or *make* or *do a study* (Brinton 2008: 34), although it is not possible to replace light verbs with roughly synonymous heavier verbs (e.g. *give a kick* vs **bestow a kick*, **grant a kick*) (Brinton 2008: 46). Moreoever, as will be discussed below, light verbs occur with a greater variety of NP complements. On the other hand, *lose-sight* CPs consist of pairings with a greater variety of verbs but less diversity of grammatical options: the type of NP complement is most often a bare NP (*lose sight*), usually without the possibility of adjectival modification (**lose good sight*). It is more often the case that the verb in this type of CP cannot be easily substituted with another (e.g. *lose sight* but **find sight*) (Brinton 2008: 45) and occur with a narrow range of NP

[2] There are many different definitions and terms used to describe semantically lighter verbs in composite predicates. Verbs of this type have been called *thin verbs* (Allerton 2003) or *support verbs* (Ronan 2014). Jespersen (1942) first used the term *light verb* in his analysis of English. For simplicity, when referring to *do*, *give*, *have*, *make* and *take*, I use the term *light verb* and refer to pairings of a light verb + NP. complement as *light verb constructions* throughout this study. In the case of verbs other than these five in composite predicates, I refer to them simply as more lexically specific verbs.

[3] Section 3 provides further discussion on methods used to identify CPs, including the approach used here and in Bonial (2014).

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

compelements. Unlike the *take-a-look* type of CP, these are non-compositional and generally have a more idiomatic usage. As discussed in Sundquist (2020: 355) and Claridge (2000: 73), the differences between the two types of CPs are often unclear, with exceptions to rules based on various grammatical characteristics. However, general distinctions are easier to capture: the first type exhibits greater flexibility of grammatical options within the NP, while the second is more idiomatic and exhibits less variability within the NP. The characteristic that allows for the most clearcut distinction between the two is the verb: *take-a-look* CPs contain one of the five most frequent light verbs and *lose-sight* CPs contain more lexically specific transitive verbs of lower frequency.

2.2 *An exemplar-based Construction Grammar model*

Both types of CPs lend themselves well to an analysis within a Construction Grammar theoretical framework (Goldberg 2003, 2006; Croft 2001; Croft & Cruse 2004). Any approach within this framework is useful for analyzing verb–noun pairings in CPs, since it captures their idiosyncrasies in an attempt to explain why some pairings are possible while others are not.[4] The basic notion of a construction used in the study at hand follows from Goldberg (2003) and the idea that constructions are 'stored pairings of form and function, including morphemes, words, idioms, partially lexically filled and general linguistic patterns' (2003: 219). As Bybee & Eddington (2006: 327) note, the core concept of storage in this definition is important: grammar and the lexicon are closely connected when multiple grammatical elements occur frequently together, and any usage-based Construction Grammar approach assumes that these co-occurrences are stored in human memory through repetition. Associations that are most frequently reinforced become the building blocks for retention and subsequent use in new contexts. As patterns emerge through speaker experience, the constructions become entrenched, with greater linguistic strength of association and are more readily retrieved as a single unit (Clausner & Croft 1997: 252). In the case of verb–noun pairings in CPs, a light verb like *make* or *take* occurs with a variety of moveable parts (Bybee & Eddington 2006), as in *make* + NP, whose open slot may be filled with any number of NPs in the complement position.

The importance of frequency in the process of categorization in exemplar models has been discussed widely in Bybee & Eddington (2006) and Bybee (2006, 2010, 2013).[5] Following Pierrehumbert (2001), Bybee (2013: 53) defines exemplars as categories formed from tokens of experience that are judged to be the same, are structured by similarity and frequency, and exhibit prototype effects. Simply put, speakers map incoming tokens, or instances of actual usage, onto similar existing representations, or

[4] See also Wierzbicka (1982) for further discussion of the idiosyncrasy of certain light verb + NP complement combinations.

[5] See Divjak (2019) for a useful overview of exemplar-based approaches in her discussion of type and token frequency and memory.

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

exemplars, and if these representations are already present and stored, they are strengthened. Thus, token frequency is the defining characteristic in the formation of exemplars, since it is through repetition of linguistic experiences that prototypical characteristics become stored with the greatest strength of association (Bybee 2006: 714). The mapping process involves evaluation of the degree of similarity between the existing representation and probes of new linguistic experience. In this way, exemplar clusters of relatively similar linguistic data are formed, with some members that are more central to the overall category (Bybee 2010: 18).

As an example, building on this notion of categorization in psychological literature (Nosofsky 1988), Bybee & Eddington (2006) apply an exemplar-based model to the idiosyncratic patterns of combination of four different Spanish verbs of *becoming* with certain adjectives. The combinations appear idiosyncratic despite some commonalities across verbs and adjectives. For instance, *ponerse nervioso* is 'to get nervous', *ponerse furioso* is 'to get angry' and *ponerse pesado* is 'to become annoying'; all convey similar negative emotions (e.g. nervousness, anger or annoyance), but a similar adjective *loco* 'crazy' combines with a different verb altogther, *volverse loco* ('to go/become crazy') while ?*ponerse loco* is only marginal (Bybee & Eddington 2006: 330–1). Even though it might not be possible to capture all the uses of a particular verb of becoming and the characteristics of all adjectives that go with each of the four verbs, as Bybee & Eddington (2006: 324) note, it is possible to identify high-frequency tokens of use as more central than others, especially when taking frequency into consideration along with varying degrees of semantic similarity. Subjects in the experimental portion of their study tended to accept low-frequency pairings that were semantically similar to high-frequency exemplars; in comparison, subjects were significantly less accepting of low-frequency items if they are not semantically similar to high-frequency pairings.

2.3 *Exemplars and verb–noun pairings in composite predicates*

There are a number of similarities between the Spanish *becoming* construction and English verb–noun pairings in CPs with respect to the possibility that exemplars shape patterns of use. As Bonial (2014: 125) points out, in both cases, there are two separate grammatical elements that form a single, meaningful unit, and the ways in which these elements combine with each other is at times more idiosyncratic than systematic. While it is possible to combine *make* with *recommendation* along with similar nouns like *suggestion*, the combination of *make* with another similar noun like *advice* is not well formed. On the other hand, there does appear to be a certain degree of regularity with respect to categorization of some groups of nouns that occur with the same light verb. According to North (2005: 2), for instance, most nouns of motion combine with *take* to form CPs like *take a stroll*, *take a run*, but nouns that convey non-verbal expressions do not (e.g. **take a groan*, **take a smile*). Moreover, the relative frequency of certain verbs + noun combinations is important. Much like the exemplars of Spanish verb + adjective pairings, the presence of both low-frequency pairings and a handful of

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

high-frequency verb–noun combinations allows for the evaluation of the role that frequency plays in the production of CPs.

Bonial (2014) demonstrates a way to apply the exemplar-based model of Bybee & Eddington (2006) to the English data on light verb constructions. She examines several families of nouns that share semantic properties and occur with the same light verb (e.g. *make a statement*, *announcement*, etc.). Using the lexical semantic database FrameNet (Fillmore *et al.* 2002) and the 1.7-billion-word Gigaword corpus, she identifies nouns in the corpus that occur with the same light verb and which are part of the same semantic frame – a description of a type of event, relation or entity, and the participants in it. Results of the study provide further support for the claim that high-frequency exemplars significantly affect the acceptability of low-frequency, semantically similar light verb–noun combinations (Bonial 2014: 154). There are several factors at play in this result, including the relative size of a noun family, the overall frequency band from which the majority of its tokens come, and the family's distributional profile (i.e. long tail vs split tail families).[6]

2.4 *An exemplar-based diachronic approach to CPs*

Constructional change has been the subject of much discussion in recent diachronic research (e.g. Hilpert 2013; Barðdal & Gildea 2015; Perek 2020). However, few studies specifically address the notion of exemplars from a diachronic perspective. As noted by Bybee (2010: 66–7), a usage-based approach that takes language change into account must consider the relative strength of exemplar representations measured by their token frequency. However, Pierrehumbert (2001) also points out the impact of temporal factors in the increased strength of exemplars. In her discussion of the perception of phonetic parameter values, she states that, 'Exemplars encoding frequent recent experiences have higher resting activation levels than exemplars encoding infrequent and temporally remote experiences' (2001: 141). Ettlinger (2007) applies this same principle to historical sound change, taking this a step further to analyze chain shifts and discussing the strength of exemplars as a function of how recently an exemplar is activated through repetition (2007: 686). Thus, successful exemplars exhibit a higher token frequency than other similar representations and stand out uniquely from the group, but more importantly in a diachronic analysis, this higher frequency is also related to time by means of the recency of repetition.

As discussed extensively in previous diachronic Construction Grammar literature, token frequency is essential to our understanding of the process of entrenchment.[7] Although the concept of entrenchment is wide-ranging, as Schmid (2017a) points out,

[6] A frequency band is a token frequency range with a specific threshold that takes into account the distribution of members' token frequencies within a group or family. Also, Bonial (2014) defines 'long tail' families as those with only one leading exemplar and 'split tail' families as those that have two competing alternatives whose frequency stands out from the rest. See section 3 for further description of these methods.

[7] See the contributions in Schmid (2017b) for an overview of entrenchment, including the chapters by Hilpert & Diessel (2017) on entrenchment in diachronic Construction Grammar and De Smet (2017) on language change.

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

there are several core aspects emphasized in cognitive approaches that can be traced back to Langacker's (1987) original use of the term: novel structures become progressively entrenched to the point that they may become a unit that is variably entrenched, depending on the frequency of subsequent occurrences (1987: 59). Israel (1996), in particular, describes this process in his diachronic exemplar-based analysis of the *way*-construction (e.g. *make/dig/claw your way*), in earlier periods of English, noting the importance of token frequency information as part of speakers' prior knowledge during the categorization and mapping process (1996: 227). In a similar way, as we see in Bonial (2014), token frequency of certain verb–noun pairings in CPs provides us with a clear way to identify exemplars, to evaluate the extent to which they stand alone compared to other semantically similar pairings, and to track how these higher-frequency combinations continue to separate themselves over time.

Equally important to token frequency is type frequency in diachronic analyses based on exemplars. In approaches to language change within a usage-based framework, the type frequency of a pattern determines its degree of productivity (Bybee & Thompson 1997: 384).[8] New forms are created based on analogy to previously experienced utterances: through the categorization of similar expressions, these new forms take on independent meanings and in new contexts (Bybee 2010: 57).[9] In terms of exemplars with verb–noun CPs, for instance, high-frequency verb–noun pairings in CPs may serve as the basis for novel utterances in which a semantically similar noun can be paired with the same verb. By means of analogical extension based on semantic properties of the noun in NP complements, the number of different nouns that combine with the same verb may increase over time. This process resembles what Goldberg (2019) recently calls *coverage:* 'A potential productive use of an existing construction (a COINAGE) is acceptable to the degree that the category which would be required to include the previously attested examples *and* the coinage is well attested within the hyper-dimensional conceptual space in which exemplars cluster' (2019: 62–3). Coverage in this sense relates to (i) a construction's type frequency, (ii) semantic and phonological variability and (iii) the similarity of a given coinage to previously attested types (Goldberg 2019: 63). Novel expressions may lose out in the long term due to stronger associations to more conventional formulations, though these properties of coverage allow for the creation and development of novel expressions in the short term. In the context of verbo-nominal CPs, coverage is a function of the number of unique nouns that are already paired with a particular verb (i.e. its type frequency), its semantic variability (i.e. the polysemy of the verb) and the closeness in meaning between the new and already-existing pairings.

[8] For a helpful overview of how type frequency relates to productvity and previous research within diachronic Construction Grammar, see Barðdal (2008).

[9] See Bybee (2010: 57–76) for in-depth discussion on the role of analogy that is based on exemplars within a usage-based, diachronic analysis. Also, Himmelmann (2004) provides insight on the role of host-class expansion in the discussion on grammaticalization and host-class reduction in lexicalization.

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

Thus, in the case of the present analysis of verbo-nominal CPs, there are three aspects of exemplars to consider when evaluating their development diachronically. First, comparatively higher token frequency indicates that some individual verb–noun pairings stand out from other similar ones. Secondly, this higher token frequency rises over time; although there may be some temporary ups and downs, the long-term trend is represented by an overall increase. Lastly, greater type frequency is an indication of higher productivity: exemplary verb–noun pairings serve as models in the formation of new pairings that are more closely bound by their semantic similarities.

3 Methodology

3.1 The corpus

This study examines verb–noun pairings in CPs in a subcorpus of the larger *Corpus of Historical American English* (COHA, Davies 2010). COHA contains approximately 475 million words from over 115,000 texts written between 1820 and 2009. These texts represent four different genres: magazines, newspapers, fiction and non-fiction. The corpus has a balanced design to ensure that all decades have a similar percentage of texts from each genre. A subcorpus of 114 million words was used in this study that was equally balanced across magazines, fiction and non-fiction for texts written between 1820 and 2009.[10] All words are tagged in COHA by context for part of speech and parsed to allow for consideration of grammatical features. For the purposes of this study, COHA is useful for tracking frequency trends among CPs. Using each verb as the starting point in the identification of CPs in the subcorpus, one is in a position to evaluate the type frequency of each verb, measured by the number of different nouns that occur in NP complements in CPs with that verb, and the token frequency, measured by the overall instances in which a verb is paired up with nouns in NP complements.

3.2 Identification of verb–noun pairings in CPs

Using the full text format available through COHA, I ran SQL search queries on the subcorpus to find possible candidates of verb–noun combinations. First, searching by lemma, I collected all examples of verb–noun combinations with *make* and *take* as representative of the *take-a-look* type of CP. Next, in order to compare these results with those from *lose-sight* CPs, I ran the same search query on a medium-frequency

[10] Texts for the subcorpus were selected from the larger COHA corpus according to a combinatorial optimization algorithm that takes into account each individual text's word count, genre and decade (Sundquist & Rothwell 2019). The algorithm's code is provided here: https://github.com/corpus-based-research-lab. Thus, unlike the complete COHA corpus, the subcorpus is carefully balanced across nineteen decades, and each decade has the same word count of 6 million, including 2 million words from each of the three genres (e.g. magazines, fiction and non-fiction). Newspaper sources were not included in this study, since the earliest decades in COHA lack texts from this genre. Equal sample sizes from each decade were necessary in order to avoid skewed results that arise from the use of uneven sample sizes in longitudinal analyses of type frequency. See Perek (2018) for a discussion on normalization and the effects of uneven sample size on frequency counts per decade in COHA.

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

verb (*lose*) and a lower-frequency verb (*bear*).[11] These latter two verbs are more lexically specific; yet they have been known to share properties with typical light verbs when paired with nouns in CPs (Claridge 2000: 122; Allerton 2003: 174–91). For each verb lemma, I conducted an initial search that yielded all lemmatized forms of nouns that occur within five slots after the verb in the subcorpus.[12] For a highly frequent verb like *make*, this yielded over 15,000 different lemmas of nouns per verb; in the case of the verb *bear*, there were as few as 4,898 different lemmas. This approach is in line with synchronic analyses of English like Quirk *et al.* (1985) and Bonial (2014), as well as diachronic studies like Claridge (2000) and Ronan (2014), in which a variety of verbs and nouns in CPs are identified by means of semantic and grammatical properties.

In order to get an accurate count of type and token frequencies, several steps were necessary to filter out a large number of verb–noun pairings that are not CPs. First, I used concreteness scores as listed in Brysbaert *et al.* (2014) to include only abstract nouns while temporarily excluding concrete ones. By crosschecking the nouns in the verb–noun combinations from COHA with concreteness scores and using a cutoff score of 3.5, I was able to shorten the list to temporarily include only nouns that were more abstract.[13] For instance, pairings like *take* + *house* (concreteness score = 5.0), *make* + *bread* (4.9) or *make* + *word* (3.6) were filtered out at this step. Secondly, I used WordNet's (Fellbaum 1998) lexical file information to check which nouns fit the categories of stative and eventive nouns, as suggested by Chen *et al.* (2015) and listed here in table 1.

A noun's WordNet status allowed for additional truncating of the original list by including only those nouns that shared the semantic properties represented by the lexical file types listed in table 1. For instance, the noun *offer* in the CP *make an offer* would be considered an eventive noun and categorized as such by virtue of it being a noun of relation (e.g. 'noun.relation' in table 1) in WordNet (Chen *et al.* 2015). Thus, all pairings of the lemmas of *make* and *offer* would be included. Additional examples are *make* + *appearance* (noun.action), *make* + *improvement* (noun.event), *take* + *ownership* (noun.possession) or *take* + *swing* (noun.action). Pairings that were omitted because they are not eventive or stative nouns included, for instance, *take* + *fact* or

[11] The verbs' rankings among all verbs in the subcorpus are as follows (based on frequency counts of lemmatized forms of all verbs excluding copulas): *make* (2), *take* (9), *lose* (53) and *bear* (120). *Lose* and *bear* were selected because they provide interesting comparative data with the *take-a-look* CPs, are of the *lose-sight* type, and they occur at both a medium and low frequency relative to the two highly frequent light verbs.

[12] Here is a sample query using the COHA web interface: TAKE_v + [nn*] within 5 slots; www.english-corpora.org/coha/?c=coha&q=99998064 This query retrieves all instances of the lemmas of verbal forms of *take* in which a lemma of any noun may occur within five slots after the verb. This is a collocational search that displays all instances of these combinations according to each decade between 1820 and 2009. Also, although it is possible to search slots filled by nouns *before* a verb in COHA, in passive constructions (e.g. *mistakes were made*) or some relative clauses (e.g. *the assumption that he made*), this was not done here due to the high number of false positives that it would create.

[13] 3.5 was chosen as a conservative cutoff score that allowed for a large number of slightly more concrete nouns for consideration while staying close to the middle ranking of 3 on the 5-point scale established by raters in Brysbaert *et al.* (2014). As discussed below, a manual check using the context of the verb–noun pairing in COHA was performed after filtering of the data, following guidelines in Bonial *et al.* (2015).

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

Table 1. *WordNet lexical file information types of interest for eventive and stative nouns (from Chen* et al. *2015)*

Name	Nouns denoting…
noun.act	acts or actions
noun.cognition	cognitive process
noun.communication	communicative process
noun.event	natural events
noun.feeling	feelings and emotions
noun.location	spatial position
noun.motive	goals
noun.phenomenon	natural phenomena
noun.possession	possession and transfer
noun.process	natural processes
noun.relation	relations between things
noun.state	stable states of affairs

make + *basis*, which do not fall into any of the WordNet categories listed in table 1. Thirdly, a low-frequency cutoff was used. Any noun collocates that occurred alongside a verb at a rate of .05 words per million or less (5 tokens or fewer in the entire COHA subcorpus) were filtered out. This cutoff allowed for the exclusion of those verb–noun pairings that occurred at such a low rate that they appeared only sporadically.[14] For instance, the pairing of *make* + *estimation* occured only 5 times and was not considered in the empirical analysis, and many of the low-frequency pairings for sounds (e.g. *peep*, *plop*, *plunk*) that occurred with *make* fewer than 5 times in total in the subcorpus for the period under investigation were excluded. Finally, a manual check with crosschecking of the context in COHA was used. This was especially important in the case of nouns that were not listed in Brysbaert *et al.* (2014) or in WordNet annotations. The guidelines set up in Bonial *et al.* (2015) for light verb constructions were used to check for false positives that may have been included mistakenly or for any false negatives with nouns that, for various reasons, may have been unwittingly initially filtered out.

3.3 *Identification of noun families*

The next step involved identifying families of semantically similar nouns that occur in CPs with the same verb. I used FrameNet (Fillmore *et al.* 2002) to identify families of

[14] It should be acknowledged that by using a minimum threshold of .05 words per million, we rule out the possibility of including very rare verb–noun pairings or even *hapax legomena*. These single, isolated occurrences are relevant in discussions on changes in productivity, as pointed out by Baayen (2009), Perek & Hilpert (2017). However, because of the Zipfian distribution of tokens (Zipf 1949), the number of single occurrences of some verb–noun pairings is extremely high (i.e. several thousand isolated instances). It is not feasible to include single occurrences at this step of statistical analysis, because of the high number of individual items that would have to be checked manually. As will be demonstrated in section 5, however, some extremely low-frequency verb–noun pairings are considered in *post hoc* analysis of some families.

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

noun collocates that occur with *make*, *take*, *bear* and *lose*. FrameNet allows for the classification of semantically similar nouns that are all part of the same semantic frame. It displays each frame along with a definition, the required and optional frame elements, and all lexical units that evoke this frame, with human annotations based on how words are used in actual texts. Although each frame evokes lexical units of varying parts of speech, nouns are the only lexical units of direct relevance to this study. In many cases, a family of nouns may consist of close synonyms in FrameNet, but in other instances, the words are related more generally by virtue of belonging to the same semantic frame (e.g. *safari*, *pilgrimage* and *odyssey* in the 'Travel' semantic frame). As Bonial (2014: 176) notes, while it is possible that other factors may play a role in the semantic similarity of nouns that occur with the same verb, FrameNet comes close to capturing the semantic similarity between lexical items and identifies their shared membership in a real-world context.

After identifying nouns with common membership in frames according to FrameNet, I selected a number of families for comparative analysis based on several variables. First, I considered family size, operationalizing size according to the overall number of nouns that are members of a frame and which occur with the same verb in the COHA subcorpus. A 'small' family was defined as including 3 to 9 nouns, while a 'large' family contains 10 or more.

Secondly, I examined a range of families according to the frequency band to which the largest percentage of each family's tokens belong. After determining the aggregate token count of all members in a family from the decades between 1820 and 2009, I identified the frequency bands according to the following scale: high frequency (≥1000 tokens), medium frequency (300–999 tokens) or low frequency (≤ 299 tokens). For instance, consider the nouns in the 'Similarity' frame that occur with *make* in table 2. This group is considered a high-frequency family, because the largest percentage of tokens comes from the high-frequency band (i.e. greater than 1,000 tokens). *Difference* occurs with *make* 1,836 times in the subcorpus and makes up the largest percentage of tokens in the family. For comparison, consider table 3 and the low-frequency family of nouns like those from the 'Reveal Secret' frame which occur with *make*. The tokens in this family include verb–noun pairings that occur in the low-frequency band. All tokens occur at a rate of less than 300 in this family.[15]

Thirdly, I examined families with different distributional patterns among the highest-frequency members. Bonial (2014) compares families with a 'long tail' distributional pattern in which a single noun stands out from the other family members. This pattern differs from another pattern that she refers to as a 'split tail', with two

[15] Frequency bands are a useful tool for examining the relative distribution of token frequency within a group (e.g. a family of similar nouns that occur in pairings with the same verb). They are used extensively in Bonial's (2014) study of light verb constructions as well as in Bermel & Knittl (2012) in their corpus-based study of Czech morphosyntactic variants and acceptability judgments. Unlike simple type frequency counts (e.g. family size), frequency bands address the challenge posed by Zipf's law (Zipf 1949), allowing for the comparison of families with only a few high-frequency members and many low-frequency members with smaller families whose members' frequencies are more evenly distributed (Bonial 2014: 158).

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

Table 2. *A high-frequency family of verb–noun pairings with nouns from the 'Similarity' frame and their token frequency and percentage of tokens in the family*

Verb–noun pairings	Token frequency	% tokens in family
make + difference	1,836	81.3
make + distinction	415	18.4
make + deviation	7	0.3
	2,258	100

Table 3. *A low-frequency family of verb–noun pairings with nouns from the 'Reveal Secret' frame and their token frequency and percentage of tokens in the family*

Verb–noun pairings	Token frequency	% tokens in family
make + concession	296	45.2
make + confession	199	30.4
make + admission	74	11.3
make + revelation	48	7.3
make + disclosure	38	5.8
	655	100

Table 4. *A split tail family of verb–noun pairings with nouns from the 'Statement' frame and their token frequency and percentage of tokens in the family*

Verb–noun pairings	Token frequency	% Tokens in family
make + statement	700	23.2
make + remark	672	22.2
make + claim	311	10.3
make + concession	296	9.8
make + mention	216	7.1
make + declaration	159	5.2
make + announcement	136	4.5
make + assertion	127	4.2
make + proposition	121	4.0
make + explanation	109	3.6
make + admission	74	2.4
make + proclamation	46	1.5
make + pronouncement	24	0.8
make + exclamation	19	0.6
make + contention	7	0.2
make + allegation	6	0.2
	3,023	100

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

members that have a proportionately higher token count than all other family members. Consider the difference between the family associated with *make* from the 'Statement' frame in table 4 and the family from the 'Similarity' frame in table 2. The 'Statement' family in table 4 exhibits a split tail, with similar token counts for both *statement* and *remark*. In comparison, the 'Similarity' frame nouns in table 2 exhibit a classic long-tailed distribution, with one dominant noun (e.g. *difference*) and other family members that exhibit a low token frequency.

4 Results

4.1 Type and token frequencies of CPs by verb

After identifying CPs with the methods of detection that are outlined above, I compiled the results of the frequency analysis for *make*, *take*, *lose* and *bear* (see figure 1). Both *make* and *take* exhibit high type frequencies that increase gradually between 1820 and 2009, while *lose* and *bear* exhibit comparatively low type frequency that rarely increases. A non-parametric test for significance (Mann–Whitney U Test) was conducted and revealed that there are significant differences in type frequency between *make* and *take* on the one hand and *bear* and *lose* on the other (Z-score is 7.495, p-value <.00001, significant at the level $p < .05$). Both *make* and *take* occur with a greater variety of nouns in NP complements in CPs, and this variety increases over the period under investigation. Unsurprisingly, *lose* and *bear* have a more narrow expressive range, with less variety of nouns in NP complements in CPs.

The pattern of token frequency of these four verbs is similar to that of type frequency. As indicated by figure 2, *make* and *take* are very different from *lose* and *bear*. As expected, the number of instances in which *make* and *take* occur in CPs is much higher than it is for the more lexically specific verbs. A Mann–Whitney U test indicates that the two pairs of verbs (*make* and *take* vs *bear* and *lose*) are significantly different from each other in terms of their token frequency (Z-score is 7.501, p-value <.00001, significant at the level $p < .05$). The frequency of *take* in light verb constructions undergoes some ups and downs but increases generally over time; *make* has a prominent and steady increase followed by a slight decrease. The token frequency of verbo-nominal CPs with *lose* remains relatively flat between 1820 and 2009, and in the case of *bear*, the number of tokens undergoes a consistent, gradual decline.

Another quantitative measurement that combines both type and token frequencies reveals noteworthy differences between these two groups of verbs over time. Following Sundquist's (2020) overview of measurements of lexical diversity, I use Margalef's Richness Index, a statistic that incorporates type and token frequencies for comparison on an equal scale over multiple time periods. Margalef's Richness Index allows for an analysis of how many unique verb–noun combinations occur over time and avoids issues of scaling which may arise when comparing high token frequency verbs like

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

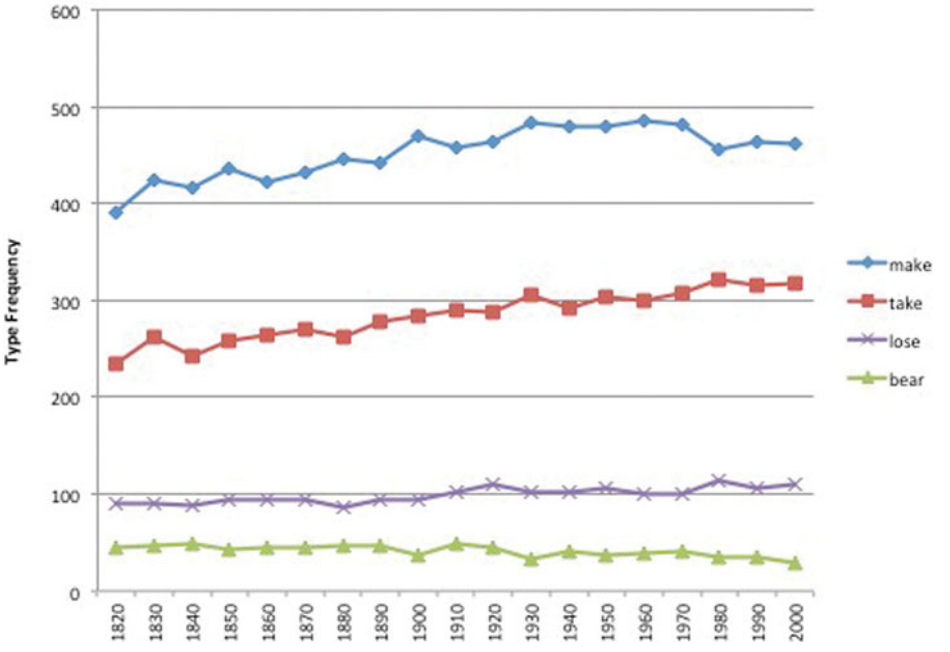

Figure 1. Type frequency of CPs with *make*, *take*, *lose* and *bear*

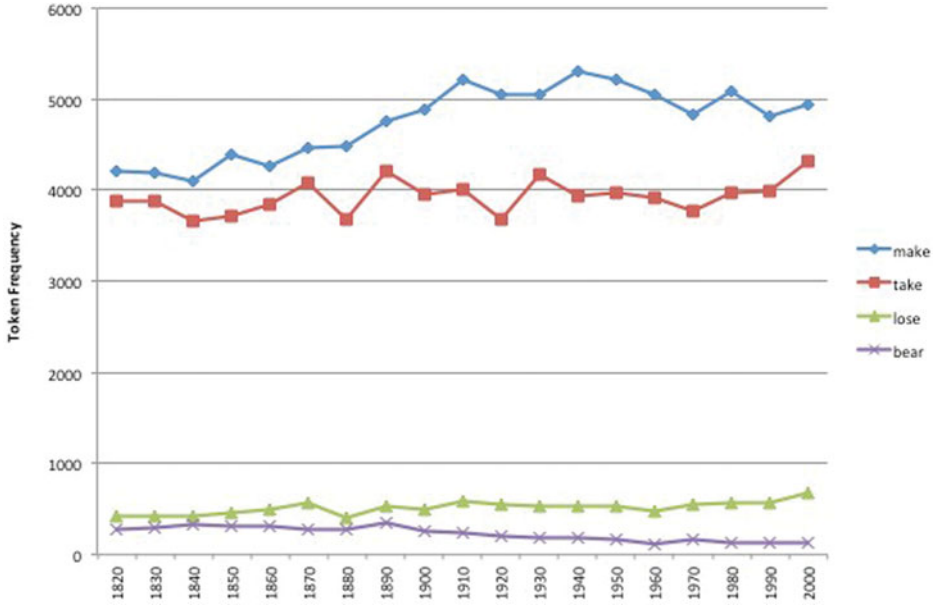

Figure 2. Token frequency of CPs with *make*, *take*, *lose* and *bear*

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

take or *make* with lower token frequency verbs like *bear* or *lose*.[16] Results of this test indicate that *make* and *take* increase in richness over the period between 1820 and 2009 (where *R* increases from 6.03 to 6.56 for *make* and 3.43 to 4.63 for *take*), while *lose* (4.42 to 4.21) and *bear* (2.66 to 2.61) decrease in terms of the unique combinations of NP complements with these verbs.

4.2 Family size

As pointed out in section 3.3, families of nouns vary in size, including small ones like the seven members of the 'Perception Active' frame in figure 3. The noun *look* occurs most frequently with *take*, as in the light verb construction *take a look*, while others like *glance*, *glimpse*, *observation*, etc. remain consistently and relatively infrequent. Some new nouns like *gander* appear with *take* for the first time later (e.g. in 1940) and occur only sporadically without any noticeable increase before the final decade.

An example of a large family is associated with the 'Statement' frame, as noted in table 4. Frequency data from this family are depicted in figure 4 and display changes in token frequency over time. There are 16 members of this family that occur with *make* in the

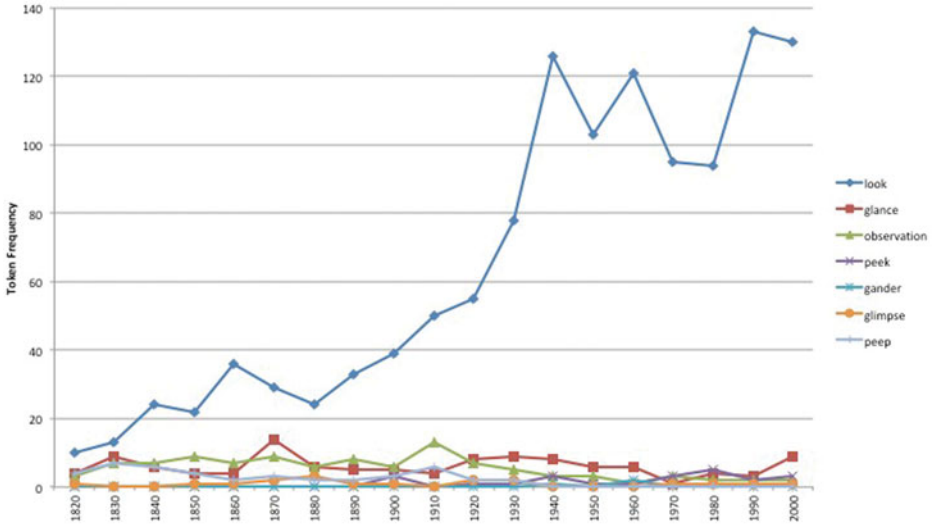

Figure 3. A small family of nouns from the 'Perception Active' frame that occur with *take*

[16] See Margurran (2013) for an overview of diversity measurements that are common in the field of ecology but can be applied to corpus linguistic data, including Margalef's Richness Index: $R = (S\text{-}1)/\ln N$, where S is the number of species (e.g. types) and N is the total number of organisms (e.g. tokens). Also, see Jarvis (2013) for an overview of similar empirical tests of lexical diversity that avoid sample-size issues that occur with simple Type–Token Ratio measurements.

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

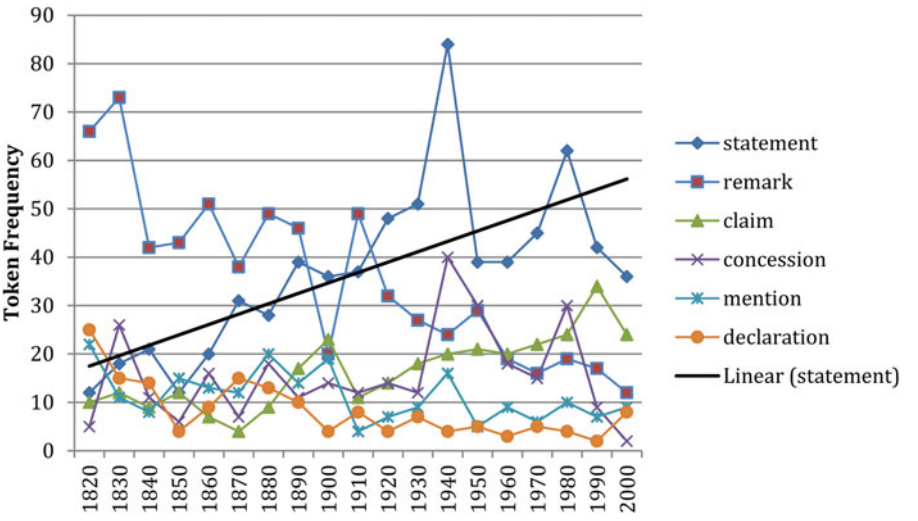

Figure 4. A large family of nouns from the 'Statement' frame that occur with *make*

COHA subcorpus, (e.g. *make a remark* or *make a claim*).[17] *Remark* and *statement* have the highest token counts, although they exhibit divergent trends, with *statement* leading the way. The frequency of the other nouns, despite some temporary increases (e.g. *claim* in the 1990s) and decreases (e.g. *declaration* in 1800s), remains relatively low.

4.3 Frequency bands

Following the method described in section 3.3, high-, medium- and low-frequency families were identified. An example of a high-frequency family was mentioned in figure 3, with CPs like *take a look* associated with the 'Perception Active' frame. Another example of a high-frequency family is given in figure 5 and includes nouns in the 'Similarity' frame that occur with *make*. The noun *difference* occurs in 81.3 percent (1836/2,258) of all the tokens in this family, putting the family in the high-frequency band. As figure 5 shows, *difference* continues to increase over time, while *distinction* and *deviation* remain at a low frequency.

A family from the medium-frequency band includes nouns from the 'Travel' frame that occur with *take*, including nouns like *trip*, *journey* and *excursion*. In figure 6, the most frequent and consistently increasing pairing (*take* + *trip*) continues to rise while others remain lower or decline.

Lastly, an example of a low-frequency family includes nouns that evoke the semantic frame 'Change position on a scale' in FrameNet, as seen in figure 7. These include, for example, the light verb construction *take a fall*. The largest portion of tokens in this

[17] Note that lowest-frequency pairings are excluded here in order to improve the readability of figure 4. Also, a trendline is provided for the leading pairing (*make* + *statement)*. Other nouns in the 'Statement' frame and their raw frequencies with *make* are listed in table 4.

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

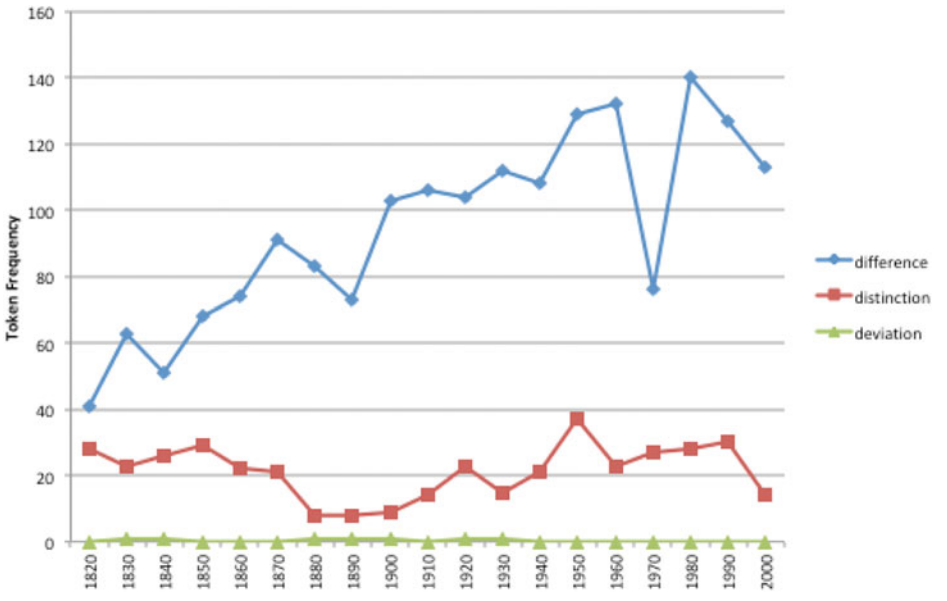

Figure 5. A high-frequency family of nouns from the 'Similarity' frame with *make*

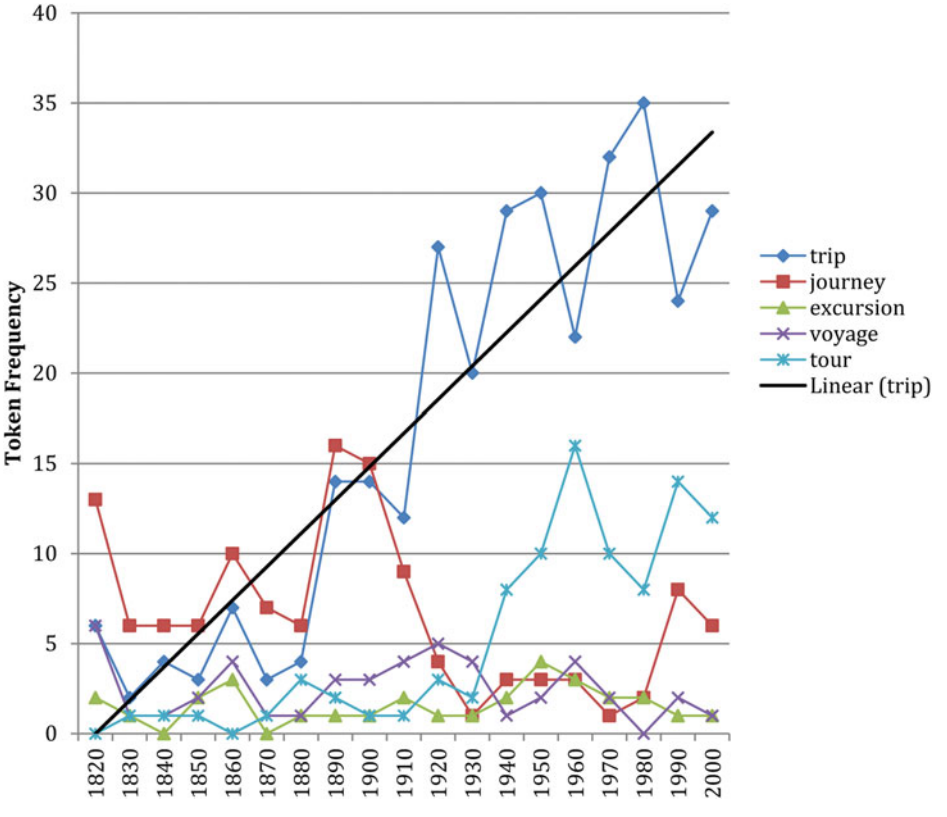

Figure 6. A medium-frequency family of nouns from the 'Travel' frame with *take*

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

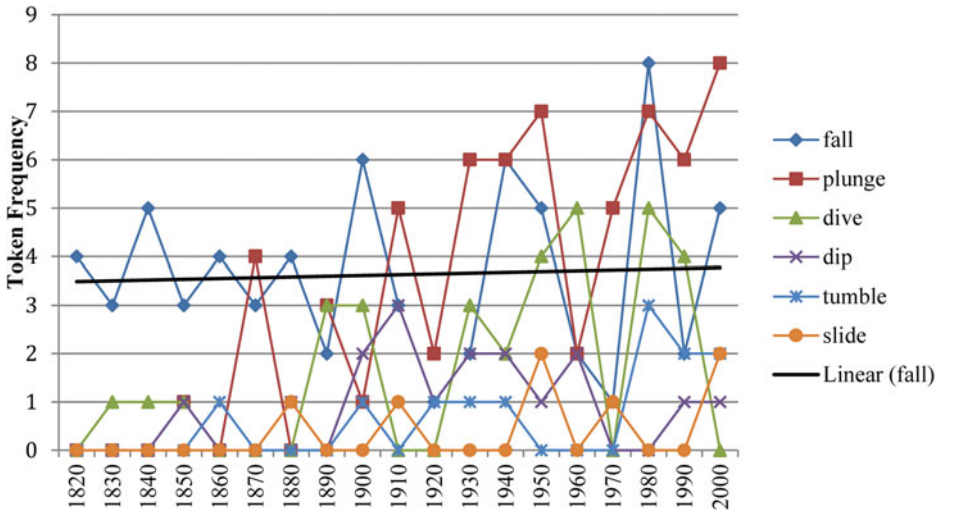

Figure 7. A low-frequency family of nouns from the 'Change of Position' frame with *take*

group involves verb–noun pairings that are in the low-frequency band. Although *plunge* exhibits an increase in the last several decades, none of the verb–noun pairings in this family show a noticeable trend toward becoming increasingly dominant, and all six members remain at a low token frequency between 1820 and 2009.

Thus, families whose majority of tokens occurs in medium- or high-frequency bands have at least one increasingly high-frequency pairing that stands out, while those families whose tokens mainly come from the low-frequency band do not. The diachronic data here also indicate that all low-frequency families in this study have fewer than ten family members. While there are both large and small families from medium- or high-frequency bands, there are no large families whose majority of tokens belongs to low-frequency bands.

4.4 Long-tailed and split-tailed families

Data on the distributional patterns of high- and low-frequency verb–noun pairings indicate that most families have a long-tailed distribution, like those presented in figures 3, 5 and 6. A single verb–noun pairing separates itself from the others and increases in frequency. Some families, however, exhibit a split-tailed distribution, like nouns in the 'Sounds' frame that occur with *make*, as depicted in figure 8. This is a medium-frequency, large family with 15 members in a split-tailed distributional pattern. Pairings with both *noise* and *sound* separate themselves over time and continue to increase in frequency, while the other members remain at a low frequency. If extremely low-frequency nouns in this family were included in a *post hoc* analysis (e.g. *blast*, *crunch*, *peep*, *plunk*, *plop*, etc.), the membership of the family would be even greater. There are 12 additional nouns in the COHA subcorpus that occur below

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

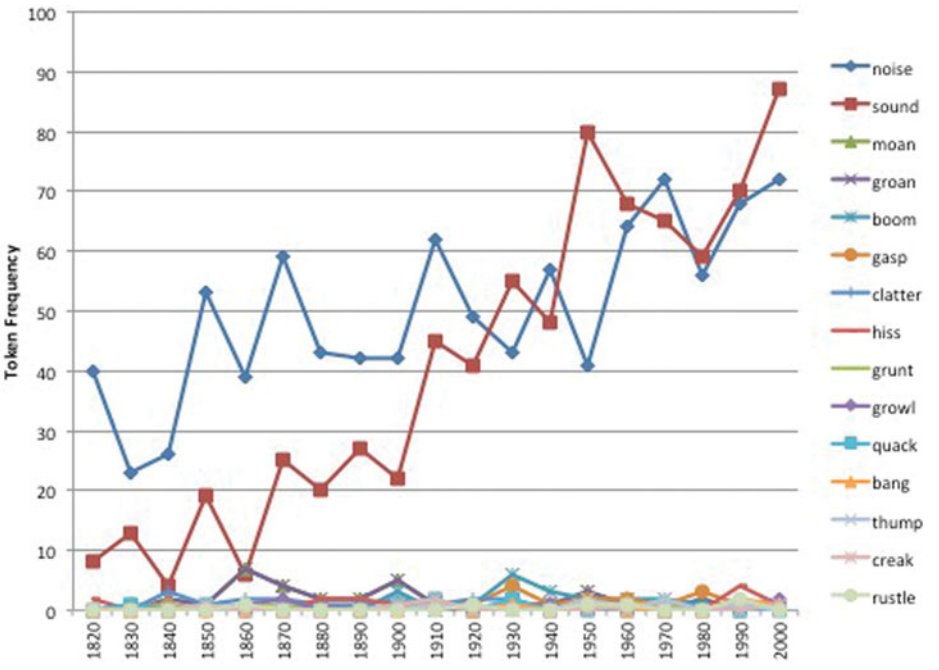

Figure 8. A split-tailed family of nouns from the 'Sounds' frame with *make*

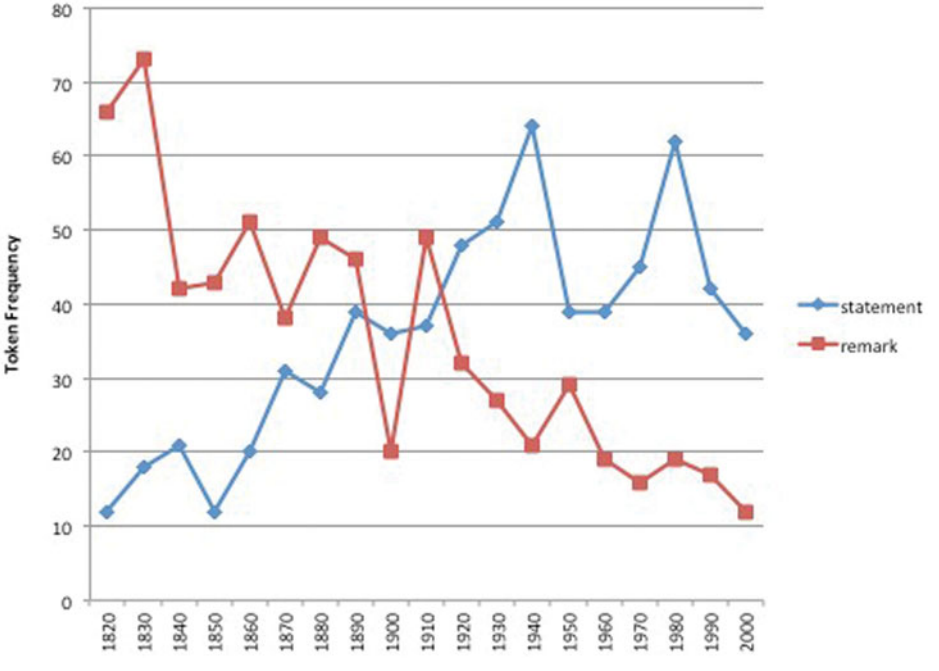

Figure 9. Pairings of *statement* and *remark* from the 'Statement' frame with *make*

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

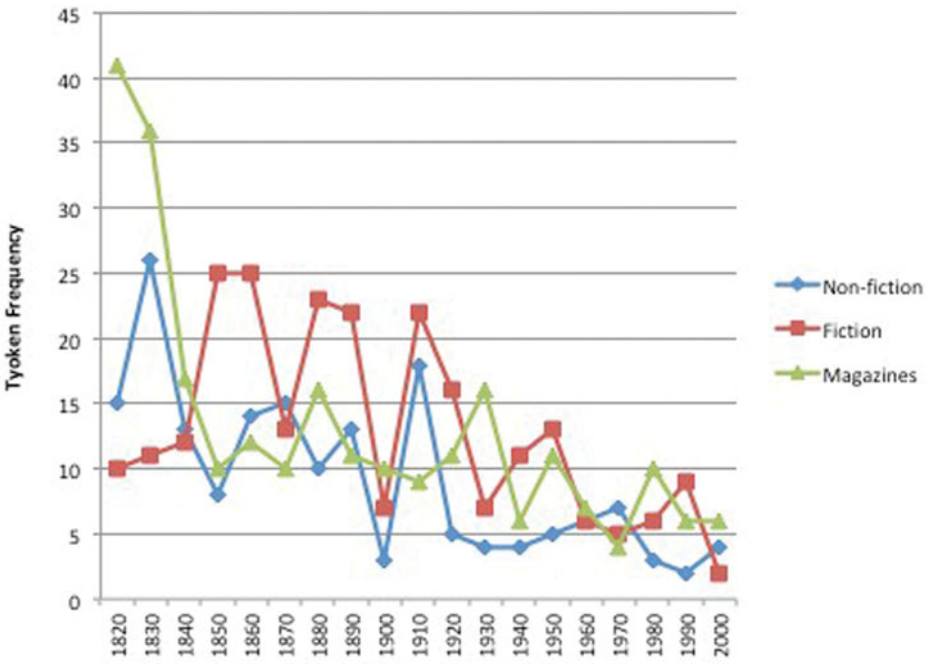

Figure 10. A family of nouns from the 'Statement' frame with *make* in the COHA subcorpus (1820–2009) divided by genre

the minimum threshold of .05 words per million. Even with these additional members, the distributional pattern of frequency within the family stays the same: *make + sound* and *make + noise* continue to separate themselves as high-frequency pairings.

A similar tendency is seen in another large family, namely, the 16 nouns in the 'Statement' frame that were described in figure 4. In order to highlight the pattern of development of the two most frequent members, the data in figure 3 are repeated in figure 9 without the other members of the family. Although the family exhibits a split-tailed distribution, the most frequent member of the family switches midway near the beginning of the twentieth century. *Remark* undergoes a gradual drop in frequency between 1820 and 2000, while the frequency of *statement* increases.

A closer analysis of *make + remark* indicates that the most noticeable drop in frequency occurs in texts from one genre, namely, magazines (see figure 10). Although the same declining trend is evident in texts from the non-fiction and fiction genres, the most noteworthy change comes from texts in this single genre, particularly in the decades from 1820 to 1850. Without these few early data points, the frequency of *make + remark* would not be much different than other pairings, and the family would not have a split tail. *Make + statement* would be the only high-frequency exemplar in the group, following a gradual increase in frequency that is typical of long-tailed distributional patterns.

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

4.5 Similarities and differences between CPs with different verbs

The data presented thus far are limited to examples of families of nouns that occur with *make* and *take*. Results from the analysis of these families indicate that there are no noticeable differences between these two verbs with respect to variables like family size, frequency bands and distributional patterns. Both *make* and *take* occur with a wide variety of families, including those that are both small and large whose majority of tokens occur in any of the three frequency bands. In addition, there are split-tailed and long-tailed families of nouns that occur with *make* and *take.*

In contrast, examples of families that occur in CPs with *lose* and *bear* are rare. In some cases, the semantic frame with which a particular collocate noun is associated does not include enough other nouns in the corpus to achieve family status, or in other cases, the noun does not belong to any frames that are annotated in FrameNet. For example, *bear* occurs with *risk*, a noun that belongs to the 'Daring' frame and occurs in a well-formed CP in the COHA subcorpus (e.g. *bear the risk*). Even though there are several other nouns in this semantic frame with *risk* (e.g. *chance*, *audacity*, etc.), none of them are CPs with *bear* or occur in the study's subcorpus. Other nouns that occur in pairings with *bear*, like *grudge*, do not belong to a frame in FrameNet at all. Moreover, in some cases there might be a noun in a CP that shares a frame with another noun, but there are no other family members attested in the subcorpus. For example, *control* belongs to the same frame as *command*, and both occur frequently enough with *lose* to be considered for further analysis. However, two nouns are not enough for 'family' status, according to the data collection procedures outlined above that specify a minimum of three members.

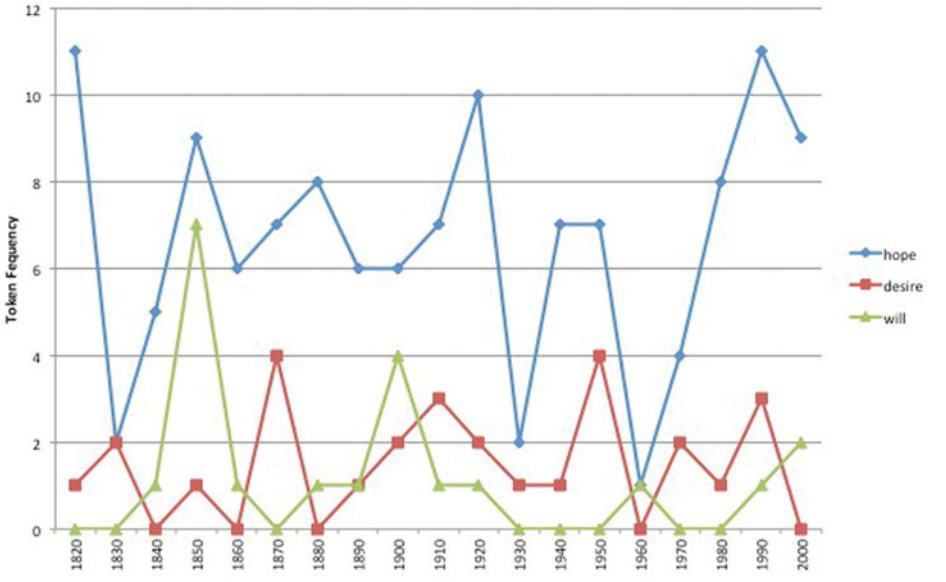

Figure 11. Nouns from the 'Desiring' frame with *lose*

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

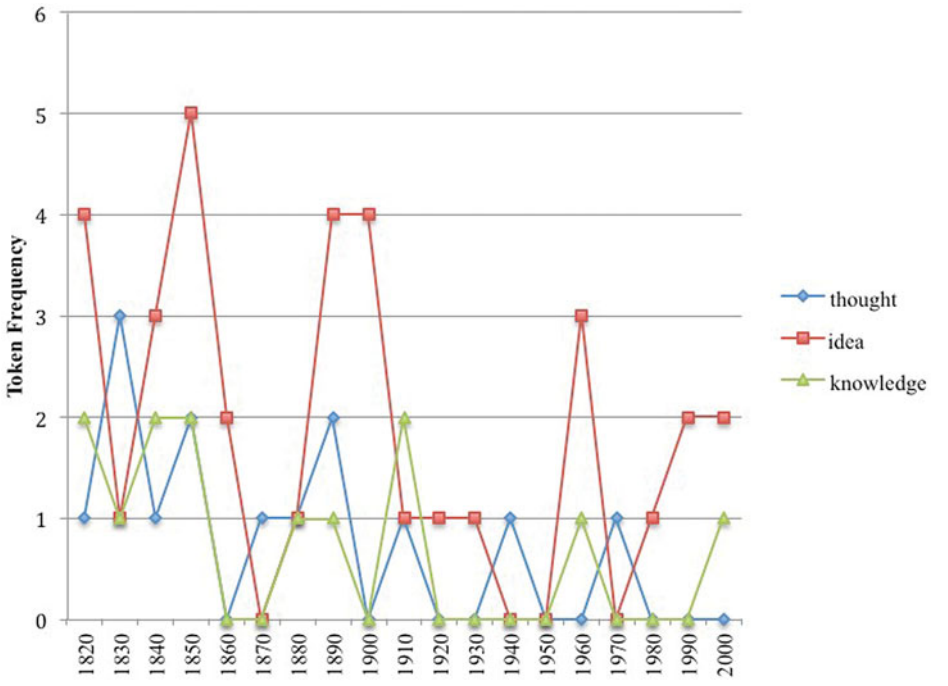

Figure 12. Nouns from the 'Awareness' frame with *bear*

There are some exceptions in the case of both *lose* and *bear*. For instance, there is one family of three nouns associated with the 'Desiring' frame in figure 11. However, this is a small family from the low-frequency band with only *hope*, *desire* and *will* with *lose*. While *lose* + *hope* is the most frequent member, its token frequency is sporadic throughout the period under investigation and remains low, with fewer than 10 instances per decade.

In a similar way, families with *bear* are also rare. One small family from the 'Awareness' frame includes three nouns that occur in CPs with *bear*, as seen in figure 12. This is another small family from the low-frequency band (i.e. the majority of nouns in this family when paired with *bear* have a token frequency ≤ 299). Although *idea* has the highest total token count, its frequency fluctuates at a low level and decreases. The frequency of the other members of the family also remains low and decreases.

5 Discussion

This study poses the question whether light verbs in CPs are more likely to generate what might be considered exemplars than more lexically specific verbs. It also considers which factors are relevant to the evolution of exemplars in a diachronic analysis. Recall that three characteristics of exemplars of verb-nominal CPs with semantically lighter verbs are considered in this study: how certain verb–noun pairings separate themselves by nature

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

of their high token frequency, how their token frequency consistently increases rather than decreases, and how many unique pairings with semantically similar nouns occur with the same verb. Moreover, the previous literature on multi-word verbal expressions notes general differences between types of CPs, differentiating those like *take a look* with a light verb and others like *lose sight* with a more lexically specific verb. Results indicate that there are several families with semantically similar nouns that combine with *take* and *make* to generate CP exemplars, while those that pair with *lose* and *bear* do not. CPs of the *take-a-look* type include exemplars with *make* (e.g. *make* + *sound* or *make* + *statement*) and *take* (e.g. *take* + *trip* or *take* + *look*), where a high-frequency pairing separates itself, increases in token frequency over several centuries and, as will be addressed below, exhibits an expanding range of semantically similar nouns that may be paired with the verb. Data reveal that *lose-sight* CPs follow a different pattern: other than a few isolated high-frequency combinations with *bear* or *lose* (e.g. *bear witness* or *lose sight*), there are no high-frequency pairings that consistently increase in token frequency and function as successful exemplars that could be the basis for analogy over a long period of time.

By focusing more specifically on the distribution of high token and type frequencies among these families for the two different CP types, I was able to determine which factors shape the generation of exemplars or lack thereof. Findings from the analysis of representative types of families indicate that family size and the distributional pattern of tokens within a family do not have a noticeable effect on the presence or success of exemplars. In the case of family size, for example, families with as few as three members or as many as 16 verb–noun pairings have exemplars that stand out and increase consistently in token frequency. The number of family members does not affect whether one pairing separates itself from others or not. Moreover, the distribution of tokens within a family does not play a noticeable role: families that exhibit either a long-tailed distribution or split-tailed distribution have pairings that have characteristics of exemplars. In the case of some split-tailed distributional patterns, one of two high-frequency pairings may replace another over time (e.g. 'Statement' nouns with *make*), or in other instances, two pairings rise in tandem with each other (e.g. 'Sounds' nouns with *make*). However, there is no difference between split-tailed families and long-tailed families with respect to the presence or absence of a leading exemplar. In sum: in the case of most CPs, neither the family size of nouns nor the distribution of tokens within the families has an effect the presence of certain exemplars.

The frequency band of a family's tokens, on the other hand, does have a noticeable effect on whether a family will include a certain pairing that separates itself from the others. Recall that families whose majority of tokens is in the high- and medium-frequency bands are more likely to have pairings whose token frequency consistently rises. High-frequency bands of tokens like those in the *make* + *difference* family or even medium-frequency bands (e.g. *take* + *trip*) yield examples of singular verb–noun pairings that set themselves apart – regardless of the size of the family. Moreover, these pairings continue to increase over time. Verb–noun pairings in

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

low-frequency bands (e.g. *take a fall*), even those CPs with light verbs, fail to separate themselves or show any increases at all.

This pattern is expected if we consider their exemplars to be cases of simple entrenchment. In terms of the high- and medium-frequency bands of tokens among families of nouns, the more frequent and readily accessible pairings become, the more likely it is that they will be used more frequently. In a sense, 'the rich get richer', and the most frequent pairings are the ones that will continue to increase over time.[18] In other words, once an exemplar begins to separate itself, it continues along this trajectory – with only temporary decreases. During this same time, low-frequency pairings in high-frequency bands occur but never increase consistently like the exemplar(s) in the family. Novel expressions lose out in the long term due to stronger associations to the more conventional formulations (Goldberg 2019: 61). Moreover, as we see in the case of the low-frequency bands of verb–noun pairings, there are no conventional formulations that provide the basis for strong associations for entrenchment in these cases. No pairings separate themselves from the other in terms of token frequency.

Specific data on frequency bands in *take-a-look* CPs vs *lose-sight* CPs indicate noteworthy differences in the trajectory of high-frequency pairings in these two types of CPs. Recall that both *take* and *make* occur with a wider variety of noun families, most of which have a high-frequency verb + complement combination or two. As indicated in section 4, *take a look* provides a clear example: *look* occurs frequently with *take* as the lone example in a family of many similar nouns; the token frequency of this single pairing continually rises while other nouns fail to increase in CPs with *take*. In the case of *lose* and *bear*, there are simply no examples in which a verb–noun pairing's token frequency both separates itself from others and increases over time. Instead, tokens of pairings in the few families with *lose* and *bear* occur primarily in the low-frequency bands, if at all. Any nouns in high-frequency pairings that occur with these verbs are isolated instances that show no long-term increases and do not belong to a semantically similar family of other nouns in CPs.

More fine-grained analysis of individual pairings with *lose-sight* CPs reveals the idiomatic, non-compositional nature of this type of CP. For instance, both *lose* + *sight* (910 total tokens) and another more fixed expression, *bear* + *witness* (379 tokens), exhibit high frequencies throughout the whole period. However, *bear* + *witness* and *lose* + *sight* undergo occasional decreases rather than any kind of long-term, prominent increases. In the case of *bear* + *witness*, mutual information (MI) scores in COHA indicate high collocational strength in each decade between 1820 and 2009 (mean MI score = 8.55, $SD = 0.41$ for co-occurrence within five slots), with a slight decrease in MI score from 9.09 in 1820 to 9.04 in 2009. For *lose* + *sight*, the MI score is also consistently high in each decade (mean MI score = 7.74, $SD = 0.36$ for co-occurrence within five slots), but with a slight decrease over time (7.96 in 1820 and 7.50 in 2009).

[18] See Sundquist (2020) for more discussion of the concept of 'the rich get richer' in light verb constructions and in patterns of linguistic variation and change in general.

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

In other words, the collocational strength mirrors the high but gradually decreasing token frequency of these fixed expressions in the eighteenth and nineteenth centuries. The idiomatic use of these pairings is reflected in their specific meanings: neither *witness* nor *sight* belongs to any kind of larger families of similar nouns that occur with the same verb in other CPs. In both cases, the verbo-nominal CP contains a bare NP and is attested in COHA with no other modifications within the NP (e.g. **bear a good witness*, **bear witnesses* vs *bear witness*). Moreover, the nouns in these CPs are irreplaceable (e.g. **lose vision*, **lose view* vs *lose sight*). Their token frequency remains relatively high, but they are clearly more formulaic expressions that are used only in specific contexts (Brinton 2008: 45). In the case of *bear* + *witness*, the CP is often used in legal or religious contexts. *Lose* + *sight* has developed a metaphorical meaning, as in 'not able to keep fresh in one's mind', alongside the more literal meaning in which one is 'no longer able to see' (Brinton 2008: 45). In other words, CPs of this type are more often restricted to specific idiomatic usages and lack interchangeability of parts. On the other hand, light verbs like *take* or *make*, while they do occur in some more idiomatic phrases (e.g. *make a killing*, *take the reins*), occur with a wide range of NP complements with more substitutability of component parts.

Differences between the two types of CPs underscore the importance of frequency bands in explaining the generation of exemplars. As Bybee (2010: 38–9) points out, the most frequent exemplars are the most accessible and promote faster recognition of and greater clustering with other similar forms; constructions with high type frequency will be more likely to be used simply because of the ever-increasing strength of association and the greater number of candidates on which to base analogy (2010: 95). In the case of frequency bands like those in *take-a-look* CPs, exemplars and other high-frequency similar pairings continue this trend by continually providing more evidence in favor of the stronger associations. Similar findings provided by Bonial (2014) in her analysis of perception of novel light verb constructions support these findings. Conversely, while *lose-sight* CPs do occur with high-frequency individual pairings (e.g. *lose* + *sight* or *bear* + *witness*), these pairings do not group together to strengthen association with other similar forms. Any group of similar pairings like these isolated, idiomatic pairings remains small and fails to grow over time to be a part of a larger and more robust pattern that is typical among high-frequency bands.

These differences between verbs shed light on issues of the relative productivity of the different types of CPs and the role of exemplars. Assuming type frequency to be a strong indicator of productivity as discussed in section 2.4, it is not surprising that *make* and *take* are associated with a greater number of families – and larger families. In the same way, the paucity of families with *bear* and *lose* is expected if we consider their low type frequencies and more limited expressive range. As Allerton (2003: 173) describes it, there is a niche that is created by the semantic preciseness of lexically specific verbs.

Differences in the relative productivity of each type of CP relates to the notion of coverage outlined in Goldberg (2019: 61). CPs with *make* and *take* exhibit wider coverage by virtue of their generalized meaning, their high type frequency (e.g. the number of unique co-occurring NP pairings), the great variability of semantically

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

related noun families that can be associated with them, and the similarity with which a new verb + noun coinage might resemble others that include nouns from previously established families. In the case of *take a look*, several similar pairings with *take* occur at a low frequency as early as 1820, including *glance*, *peak*, *glimpse* or *observation.* This family appears to expand to include novel pairings at a low frequency, such as *gander*, which first appears with *take* in the 1940s in the COHA subcorpus, after the steep rise of *take a look* in the late 1800s (figure 3). *Take a gander* fades out of use, along with other low-frequency CPs like *take a peep* by the 1930s. Even though these novel creations are short-lived in this subcorpus, they provide evidence that *take* + *look* is a productive pairing that functions as the basis for new pairings that share some overlap in meaning. In a similar way, *make* occurs with nouns similar to *difference* as early as 1820. A new pairing like *make a deviation* occurs for the first time shortly thereafter in 1830 before it continues at a low rate and ultimately disappears by 1930. Moreover, novel pairings similar to *make* + *sound* and *make* + *noise* occur throughout the eighteenth and nineteenth centuries, showing further temporary expansion: *make* + *boom* first occurs in 1880 and other newer combinations like *make* + *bang* or *make* + *thump* come on the scene in 1910, in the middle of the increase of these two exemplars in the late nineteenth and early twentieth centuries. There are many low-frequency pairings with *make* that occur between 1820 and 2009, including rare pairings with *blast*, *crunch*, *peep*, *plunk* or *plop* that provide further evidence that this is an expansive family of nouns which resembles the two exemplary pairings *make* + *sound/ noise.* While many of these coinages remain at a low frequency or die out, this more fruitful breeding ground for exemplars is made possible by the extensive coverage provided by CPs with generalizable, semantically lighter verbs.

On the other hand, CPs with *bear* and *lose* do not show this kind of coverage. One of the few examples of a family of similar nouns that occur with the same verb is nouns from the 'Awareness' frame with *bear*. The highest-frequency pairing in this group is *bear* + *thought*, and the family associated with this frame contains only three members; yet all three members exhibit declining frequency. Moreover, no novel pairings similar to *bear* + *thought* appear throughout the entire period of 200 years. High-frequency idiomatic pairings with lexically specific verbs (*bear* + *witness*, *lose* + *sight*) remain isolated instances with few similar expressions that share semantic properties. Unlike *take* or *make* which allow for a wider variety of eventive nouns, *bear* and *lose* generate no clear exemplars and do not follow this pattern of productivity over time.

6 Conclusions

Results indicate that an exemplar-based model may be useful in any study of differing types of CPs from a diachronic perspective. This approach sheds light on the ways in which changes to type and token frequencies play out over a longer period of time: one is better able to evaluate the dynamic relationship between high- and low-frequency CPs from different points in time rather than being limited to a snapshot of current usage. Quantitative analysis of families of semantically similar nouns reveals that it is common

for one or two verb–noun pairings with *take-a-look* CPs to separate themselves from others and continue to become more frequent while at the same time allowing for the introduction of novel or low-frequency pairings within the same family. On the other hand, lexically specific verbs in CPs like *lose sight* generate few exemplars that are part of larger families. Various explanations for the findings were offered, including discussion of entrenchment, coverage and productivity of these two types of CPs.

Analysis of several variables reveals that the frequency band from which the majority of tokens in a family occurs and the verb itself (i.e. *make/take* vs *bear/lose*) contribute to the success of certain exemplars. The size of the family does not affect the presence and proliferation of these high-frequency verb–noun pairings, although the distribution of tokens can have an effect on the success of some exemplars if the families have a large membership and split tail. Several different kinds of families were considered representative of these variables across verb–noun combinations that include each of these four verbs.

These findings have theoretical implications for our understanding of how an exemplar-based model of composite predicates might relate to grammaticalization and lexicalization.[19] In particular, the data here support the view in Brinton & Traugott (2005) and Brinton (2008) that the historical trajectories of CPs are not all alike, and that some are the product of wholly different diachronic processes. *Take-a-look* CPs exhibit characteristics typical of grammaticalization, since they exhibit host-class expansion and greater grammatical flexibility. Exemplary verb–noun pairings with light verbs continue on a consistent path of entrenchment with expanded coverage. *Lose-sight* CPs, on the other hand, exhibit typical characteristics of lexicalization: when such CPs are considered as a whole unit, they tend to become more idiomatic and non-compositional. The greater lexical specificity of the verbs leads to host-class reduction rather than expansion. The more idiomatic expressions in CPs with such verbs as *lose* or *bear* remain non-productive as they follow a path typical of lexicalization where high-frequency exemplars and ever-growing families of semantically similar pairings are lacking.

Future research of this topic could explore several new avenues of discovery related to CPs and exemplars from a diachronic perspective, addressing some of the limitations of the study. First of all, only four verbs were analyzed here. It would be interesting to include other light verbs, such as *give*, *have* or *do*, as well as other more lexically specific verbs. In this way, one would be able to find additional support for the view here that verbs in some CPs are more likely to generate exemplars than others. Moreover, we would also be able to examine more families of semantically similar nouns to analyze the variables of family size, token distribution and frequency bands, and the interaction of these variables. This study is limited to just a handful of families that exemplify some of the effects of these variables, and a greater number of families is necessary for more robust hypothesis testing. A wider selection of verbs would allow for more in-depth analysis of split- and long-tailed distributions within some families in order to gain understanding of what

[19] An in-depth discussion of lexicalization and grammaticalization is beyond the scope of the present study; however, see Brinton & Traugott (2005) for further discussion, and in the specific context of composite predicates, see Brinton (2008).

https://doi.org/10.1017/S1360674321000344 Published online by Cambridge University Press

factors are at work. Lastly, a qualitative analysis of certain verb–noun pairings would help explain the varying degrees of productivity and divergence of these two CP types over time. Cross-family comparisons might reveal similar characteristics among exemplars or uncover reasons for differing paths of diachronic development. Such follow-up studies may shed new light on the link between exemplars and lexicalization, grammaticalization and productivity of CPs in a variety of cross-linguistic analyses.

Author's address:

School of Languages and Cultures
Purdue University
640 Oval Dr.
West Lafayette, IN 47906
USA
jsundqui@purdue.edu

References

Allerton, David J. 2003. *Stretched verb constructions in English*. London: Taylor & Francis.
Baayen, Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, vol. 2, 899–919. Berlin: De Gruyter.
Barðdal, Jóhanna. 2008. *Productivity: Evidence from case and argument structure in Icelandic*. Amsterdam: John Benjamins.
Barðdal, Jóhanna & Spike Gildea. 2015. Diachronic construction grammar: Epistemological context, basic assumptions and historical implications. In Jóhanna Barðdal, Elena Smirnova, Lotte Sommerer & Spike Gildea (eds.), *Diachronic construction grammar*, 1–49. Amsterdam: John Benjamins.
Bermel, Neil & Luděk Knittl. 2012. Corpus frequency and acceptability judgments: A study of morphsyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory* 8(2), 241–75.
Bonial, Claire Nicole. 2014. Take a look at this! Form, function and productivity of English light verb constructions. PhD dissertation, University of Colorado Boulder.
Bonial, Claire, Julia Bonn, Kathryn Conger, Jena Hwang, Martha Palmer & Nicholas Reese. 2015. *PropBank annotation guidelines*. Center for Computational Language and Education Research, University of Colorado Boulder.
Bowern, Claire. 2008. The diachrony of complex predicates. *Diachronica* 25, 161–85.
Brinton, Laurel J. 2008. Where grammar and lexis meet: Composite predicates in English. In Elena Seoane, María José López-Couso & Teresa Fanego (eds.), *Theoretical and empirical issues in grammaticalization*, 33–53. Amsterdam: John Benjamins.
Brinton, Laurel J. & Minoji Akimoto (eds.). 1999. *Collocational and idiomatic aspects of composite predicates in the history of English*. Amsterdam: John Benjamins.
Brinton, Laurel J. & Elizabeth Closs Traugott. 2005. *Lexicalization and language change*. Cambridge: Cambridge University Press.
Brysbaert, Marc, Amy Beth Warriner & Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46(3), 904–11.
Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4), 711–33.
Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, Joan. 2013. Usage-based theory and exemplar representations of constructions. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford handbook of construction grammar*, 49–69. Oxford: Oxford University Press.
Bybee, Joan & David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82(2), 323–55.
Bybee, Joan & Sandra Thompson. 1997. Three frequency effects in syntax. *Annual Meeting of the Berkeley Linguistics Society* 23, 378–88.
Chen, Wei-Te, Claire Bonial & Martha Palmer. 2015. English light verb construction identification using lexical knowledge. *Proceedings from the Twenty-Ninth AAAI Conference on Artificial Intelligence*. www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewPaper/9925 (accessed 2 February 2020).
Claridge, Claudia. 2000. *Multi-word verbs in early Modern English: A corpus-based study.* Amsterdam: Rodopi.
Clausner, Timothy C. & William A. Croft. 1997. Productivity and schematicity in metaphors. *Cognitive Science* 21(3), 247–82.
Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.
Croft, William & D. Alan Cruse. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press.
Davies, Mark. 2010. *The Corpus of Historical American English*. www.english-corpora.org/coha/ (accessed 22 June 2019).
De Smet, Hendrik. 2017. Entrenchment effects on language change. In Schmid (ed.), 75–99.
Divjak, Dagmar. 2019. *Frequency in language: Memory, attention and learning*. Cambridge: Cambridge University Press.
Elenbaas, Marion. 2013. The synchronic and diachronic status of English light verbs. *Linguistic Variation* 13, 48–80.
Ettlinger, Marc. 2007. An exemplar-based model of chain shifts. In Jürgen Trouvain & William John Barry (eds.), *Proceedings of the 16th International Congress of Phonetic Sciences*, 685–8. Saarbrücken: Universität des Saarlandes.
Fellbaum, Christiane. 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press. wordnet.princeton.edu (accessed 22 June 2019).
Fillmore, Charles J., Christopher R. Johnson & Miriam R. L. Petruck. 2002. Background to FrameNet. *International Journal of Lexicography* 16(3), 235–50.
Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7, 219–24.
Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.
Goldberg, Adele E. 2019. *Explain me this*. Princeton, NJ: Princeton University Press.
Hilpert, Martin. 2013. *Constructional change in English: Developments in allomorphy, word formation, and syntax*. Cambridge: Cambridge University Press.
Hilpert, Martin & Holger Diessel. 2017. Entrenchment in Construction Grammar. In Schmid (ed.), 57–74.
Himmelmann, Nikolaus P. 2004. Lexicalisation and grammaticalization: Opposite or orthogonal? In Walter Bisang, Nikolaus P. Himmelmann & Björn Wierner (eds.), *What makes grammaticalization? A look from its fringes and its components*, 21–42. Berlin: Mouton de Gruyter.
Huddleston, Rodney & Geoffrey K. Pullman *et al*. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
Israel, Michael. 1996. The way constructions grow. In Adele Goldberg (ed.), *Conceptual structure, discourse, and language*, 217–30. Stanford, CA: CSLI.
Jarvis, Scott. 2013. Capturing the diversity in lexical diversity. *Language Learning* 63, 87–106.

Jespersen, Otto. 1942. *A modern English grammar on historical principles VI*. Copenhagen: Allen and Unwin.
Langacker, Ronald W. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*. Palo Alto, CA: Stanford University Press.
Margurran, Anne. 2013. *Measuring biological diversity*. Oxford: Blackwell.
North, Ryan. 2005. Computational measures of the acceptability of light verb constructions. MS thesis, University of Toronto.
Nosofsky, Robert M. 1988. Similarity, frequency, and category representation. *Journal of Experimental Psychology: Learning, Memory and Cognition* 14, 54–65.
Perek, Florent. 2018. Recent change in the productivity and schematicity of the *way*-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory* 14(1), 65–97.
Perek, Florent. 2020. Productivity and schematicity in constructional change. In Lotte Sommerer & Elena Smirnova (eds.), *Nodes and networks in diachronic construction grammar*, 141–66. Amsterdam: John Benjamins.
Perek, Florent & Martin Hilpert. 2017. A distributional semantic approach to the periodization of change in the productivity of constructions. *International Journal of Corpus Linguistics* 22, 490–520.
Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–57. Amsterdam: John Benjamins.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
Ronan, Patricia. 2014. Light verb constructions in the history of English. In Kristin Davidse, Caroline Gentens, Lobke Ghesquière & Lieven Vandelanotte (eds.), *Corpus interrogation and grammatical patterns*, 15–34. Amsterdam: John Benjamins.
Schmid, Hans-Jörg. 2017a. A framework for understanding linguistic entrenchment and its psychological foundations. In Schmid (ed.), 9–38.
Schmid, Hans-Jörg (ed.). 2017b. *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge*. Berlin: De Gruyter.
Sundquist, John D. 2020. Productivity, richness, and diversity of light verb constructions in the history of American English. *Journal of Historical Linguistics* 10(3), 349–88.
Sundquist, John D. & Andrew Rothwell. 2019. Combinatorial optimization and genetic algorithms in text selection for corpus-based historical linguistics research. Unpublished manuscript. www.johnsundquist.com/research (accessed 29 July 2021).
Taylor, John. 2012. *The mental corpus: How language is represented in the mind*. Oxford: Oxford University Press.
Wierzbicka, Anna. 1982. Why can you have a drink when you can't *have an eat? *Language* 58(4), 753–99.
Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Oxford: Addison-Wesley.