Nissen, H. B., & Henriksen, B. (2006). Word class influence on word association test results. *International Journal of Applied Linguistics*, *16*(3), 389–408.
Pavlenko, A. (2009). Conceptual representation in the bilingual lexicon and second language vocabulary learning. In A. Pavlenko, (Ed.), *The bilingual mental lexicon: Interdisciplinary approaches* (pp. 125–160). Bristol: Multilingual Matters.
Paribakht, T. S., & Wesche, M. (1999). Reading and "incidental" L2 vocabulary acquisition. *Studies in Second Language Acquisition*, *21*, 195–224.
Princeton University. (2010). *About wordnet*. Retrieved June 2011, from http://wordnet.princeton.edu
Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 209–227). Amsterdam: John Benjamins.
Roininen, K., Arvola, A., & Lähteenmäki, L. (2006). Exploring consumers' perceptions of local food with two different qualitative techniques: Laddering and word association. *Food Quality and Preference*, *17*(1), 20–30.
Savický, P., & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, *9*, 215–231.
Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951.
Schulte im Walde, S., Melinger, A., Roth, M., & Weber, A. (2008). An empirical characterisation of response types in German association norms. *Research on Language & Computation*, *6*(2), 205–238.
Singleton, D. (1999). *Exploring the second language mental lexicon*. Cambridge: Cambridge University Press.
Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, *19*(5), 317–330.
Verhallen, M., & Schoonen, R. (1993). Lexical knowledge of monolingual and bilingual children. *Applied Linguistics*, *14*(4), 344–363.
West, M. (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longman.
Wettler, M., Rapp, R., & Sedlmeier, P. (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics*, *12*(2–3), 111–122.
Wolter, B. (2001). Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model. *Studies in Second Language Acquisition*, *23*, 41–69.
Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? *Second Language Research*, *23*(2), 123–153.
Zareva, A. (2011). Effects of lexical class and word frequency on the L1 and L2 English-based lexical connections. *The Journal of Language Teaching and Learning*, *1*(2), 1–17.
Zareva, A., & Wolter, B. (2012). The 'promise' of three methods of word association analysis to L2 lexical research. *Second Language Research*, *28*(1), 41–67.

7 If *Olive Oil* Is Made of Olives, then What's *Baby Oil* Made of?

The Shifting Semantics of Noun+Noun Sequences in American English

Jesse Egbert and Mark Davies

Introduction

Noun+noun constructions (NNs)—also known as pre-modifying nouns, nominal pre-modifiers, noun-noun sequences, and noun+noun compounds—are a topic of particular interest in the study of English for a number of reasons. NNs occur when a head noun is pre-modified by one or more nouns (e.g. *corpus linguistics*, *research design*, *book chapter*). However, the nature of the semantic relationship between a pre-modifying noun and a head noun is highly variable (Biber et al., 1999). Even for a given noun, there can be a wide range of meanings. Consider, for example the variety of semantic relationships that are possible between the noun *oil* and a pre-modifying noun:

Semantic relationship	*Examples*
oil is made from _____	olive oil, vegetable oil, coconut oil
oil is used for _____	baby oil, motor oil, cooking oil
oil is extracted from _____	shale oil, coal oil, tar sands oil
oil found at the location of _____	gulf oil, sea oil, ocean oil
oil belongs to _____	state oil, government oil

Each of the preceding examples come from the list of the 200 most frequent N + *oil* constructions in the Corpus of Contemporary American English (COCA). This demonstrates the tremendous amount of variation in the semantics of NNs. To further complicate matters, all of this variation exists without any grammatical cues to indicate the semantic relationship between nouns in NN pairs.

This intriguing phenomenon has been almost entirely ignored in the literature. Very little empirical research has been focussed on the semantics of NNs and how they have evolved over time. The objective of this study is to take a first step toward filling that gap by triangulating *use*-based corpus data and *user*-based classification data to investigate the

semantic relationships that are possible in NNs and how those categories are changing over time.

In the next section, we discuss research into diachronic changes in the use of NNs. After that we review previous literature on the semantics of NN sequences and their classification. We then introduce two approaches to linguistic data: use-based and user-based, and their relevance to the present study. Finally, we introduce the objectives and research questions for this study.

Diachronic Change in NNs

The use of NNs in English is increasing at an accelerated pace (Biber & Gray, 2013). It has recently been shown that this pattern is part of a larger trend in written English toward the use of more compression in the noun phrase and less elaboration in clauses (Biber & Gray, 2016). While NNs appear to be on the rise in English generally, the rate of this increase varies quite dramatically across registers (Biber & Gray, 2011; Biber, Egbert, Gray, Oppliger, & Szmrecsanyi, 2016).

NNs can simply represent a genitive relationship (e.g. FBI's director = director of the FBI = FBI director). However, there are many other semantic relations that are possible between two nouns in a NN structure, as we observed in the *oil* examples in the previous section. One of the reasons for the rapid rise in the use of NNs is the fact that this structure is extremely productive, not only in terms of the nouns that are used but also in the semantic relationships that are possible between the two nouns. Biber and Gray (2016) suggest that the list of possible semantic relationship between nouns in NNs is expanding over time, but an empirical investigation of these changes was beyond the scope of their study. While no research has looked at diachronic change in the semantics of NNs, the next section describes previous attempts at describing NN semantics in contemporary use.

Semantic Relationships in NNs

The multiplicity of possible semantic relationships between nouns in NNs is a phenomenon that has perplexed linguists for decades. In the earliest research on this topic, scholars working within a generative syntax framework were unable to agree on the best way to classify NNs based on their semantic properties (see Lees, 1970; Levi, 1974; Zimmer, 1971). Much of this research relied on paraphrasing NNs into clauses or prepositional phrases in order to determine their meaning (see also Lauer, 1995; Nakov & Hearst, 2006). Based in part on this lack of agreement, one scholar concluded that it is not possible to classify NNs into discrete semantic categories (Downing, 1977).

One line of more recent research has focussed on NNs as a third genitive variant (in addition to *of* and *'s* genitives) (see Rosenbach, 2006,

2007; Szmrecsanyi et al., 2016). However, it is also clear that many NNs cannot be paraphrased using an *'s* or *of* genitive. Moreover, many of the NNs that can be paraphrased using one of those two variants do not meet the traditional criteria for genitives (see Biber et al., 2016).

The Longman Grammar of Spoken and Written English is one of the only reference grammars of English to describe patterns of use for NNs (Biber et al., 1999). The authors of this book include a list of 12 semantic categories for NNs (three of which are subdivided into two types). However, they also acknowledge that the list they provide is by no means exhaustive or all-inclusive (Biber et al., 1999, p. 591). While this list is useful, it was developed on the basis of the authors' intuitions after surveying corpus data rather than on an empirical quantitative analysis.

More recently, Girju, Moldovan, Tatu, and Antohe (2005) developed a list of 35 semantic classification categories for NNs in a top-down fashion, based on their own experience and intuitions. It appears that the authors of this study developed this list independently of the list produced in Biber et al. (1999), as the lists look quite different and the Girju paper does not make any reference to Biber et al. (1999). Assuming that this is the case, there is a surprising amount of overlap between the two lists. For example, both lists contain categories for *location*, *purpose*, *source*, *time/temporal*, *agent/subjective*, *partitive/part-whole*, and *hypernymy/identity*. The longer list produced by Girju et al. is largely a result of an attempt to classify NNs at a more fine-grained level. For example, the authors of that study distinguish between *cause* (malaria mosquito = mosquito that causes malaria) and *make/produce* (shoe factory = factory that makes shoes).

The list developed in Girju et al. was used by coders in their study to classify a set of NNs into categories based on the semantic relationship between the two nouns. The primary coders were two Ph.D. students who were assigned to select one of the 35 categories for a set of NNs sampled from newspapers. The ultimate goal of Girju et al.'s study was to train computational algorithms to automatically classify NNs. As a result, the methods used by the authors, as well as the results they report, are limited in their usefulness for descriptive linguistic purposes. However, to our knowledge, this is the only previous study that has taken an empirical approach to the semantic relationships between nouns in NNs. The moderate agreement results they achieved are promising in that they suggest that (a) NNs can be classified based on semantics and (b) this can be done with at least some level of reliability.

The methods used in the Girju et al. study leave a number of important questions unanswered, such as: What is the best method for establishing a list of semantic categories for NNs? What is the best method for collecting a sample of NNs that represents the full range of semantic categories? Who should classify the NNs? How should this classification be performed? While some of these questions may have been beyond the scope

of Girju et al.'s study, they are central to the present study. Additionally, our study also attempts to learn about i) an empirically supported list of semantic categories for NNs; ii) the psycholinguistic reality of the semantic relations between nouns in NNs; and iii) diachronic change in the use of NN semantic categories. The present study is the first to attempt to answer those questions.

Use-based vs. User-based Approaches to Linguistics

In order to achieve the objectives of our study, it is necessary to analyze NNs using data from two different sources: use-based and user-based. In this section, we describe these two types of data and how they can be triangulated in linguistic research.

Use-based data includes any language actually produced by language users. Corpus data is a prime example of use-based data. However, corpora are not the only type of use-based data. Use-based data includes elicited language, sociolinguistic interviews, test responses, and any other type of recorded language data. The analysis of use-based data can provide insights into a wide variety of phenomena about how a language is actually used by its speakers, such as frequency of use, language choices, linguistic possibilities and probabilities, diachronic change, and sociolinguistic variation. This study will rely on use-based data to (i) generate a list of high-frequency NNs and (ii) measure the frequency of those NNs over time.

Unlike use-based data, which is produced by language users, user-based data is data about language that comes from language users. Examples of user-based data include reader/listener perceptions, perceptions of dialect features and accentedness, grammaticality judgments, text classification, and data from psycholinguistic experiments (e.g. sentence completion tasks, lexical decision tasks, eye-tracking, etc.). Technically speaking, linguistic researchers are often users of the language they are analyzing. However, these experts are excluded from our definition for user-based data, which includes only language-related data from non-expert language users. One source of non-expert user-based language data that is becoming increasingly common in linguistic research is crowdsourcing. Crowdsourced data is collected from a large number of people, usually via an internet-based crowdsourcing service (e.g. Mechanical Turk). Within linguistics, crowdsourcing has been used to collect data on reader perceptions (Egbert, 2014; Egbert, 2016), register classification (Egbert, Biber, & Davies, 2015; Biber & Egbert, 2016; Asheghi, Sharoff, & Markert, 2016), word sense disambiguation (Rumshisky, 2011; Jurgens, 2013), among others.

Traditional methods in historical linguistics have relied heavily on researcher intuition, judgment, identification, and classification. This approach has always raised questions about the reliability of data coding. However, until recently it was possible for a single researcher to code

all of the data for a study due to the relatively small data sets and corpora that have been used (typically just 1–2 million words for the most widely-used historical corpora of English). This is no longer possible in many cases. The 'lone researcher' approach is entirely impractical with the much larger historical corpora that are becoming widely used.

In this study, we propose a new approach that relies on user-based data from multiple coders. This makes it possible to (i) develop a list of semantic categories for NNs and a NN classification instrument; and to (ii) classify NNs into semantic categories. We believe this approach addresses the issue of reliability because we are able to calculate inter-coder reliability and agreement. Additionally, this approach is scalable to very large data sets, especially with the use of crowdsourcing technology.

Study Aims and Outline

In this study, we attempt to answer the following three research questions about the semantics of NNs:

1. What semantic relationships are possible between nouns in NNs?
2. Can non-expert language users reliably classify NNs into semantic categories?
3. How do the semantic categories for NNs develop historically?

This will be accomplished through triangulation of the previously listed use-based and user-based methods. In the next section, we describe these methods in detail. The Results and Discussion section contains the full results of our study and our discussion of them, organized according to the three research questions. We conclude this chapter with a summary of our findings and some reflective comments on the use of triangulation in this study.

Methods

This section contains a detailed description of the methods used in this study. We begin by describing the design of the corpus and the search queries we used. We then explain the methods used to establish a list of semantic relationship categories for NNs, develop an instrument for classifying NNs into these categories, and use that instrument to classify 1,535 NNs. Finally, we describe the methods we used to analyze this data in order to answer our three research questions.

Corpus

The use-based aspects of this study relied on the Corpus of Historical American English (COHA) (Davies, 2010). COHA contains just over

Table 7.1 Composition of COHA, by Register and Decade

Decade	*Fiction*	*Popular Magazines*	*Newspapers*	*Non-fiction Books*	*Total*
1810s	641,164	88,316	0	451,542	1,181,022
1820s	3,751,204	1,714,789	0	1,461,012	6,927,005
1830s	7,590,350	3,145,575	0	3,038,062	13,773,987
1840s	8,850,886	3,554,534	0	3,641,434	16,046,854
1850s	9,094,346	4,220,558	0	3,178,922	16,493,826
1860s	9,450,562	4,437,941	262,198	2,974,401	17,125,102
1870s	10,291,968	4,452,192	1,030,560	2,835,440	18,610,160
1880s	11,215,065	4,481,568	1,355,456	3,820,766	20,872,855
1890s	11,212,219	4,679,486	1,383,948	3,907,730	21,183,383
1900s	12,029,439	5,062,650	1,433,576	4,015,567	22,541,232
1910s	11,935,701	5,694,710	1,489,942	3,534,899	22,655,252
1920s	12,539,681	5,841,678	3,552,699	3,698,353	25,632,411
1930s	11,876,996	5,910,095	3,545,527	3,080,629	24,413,247
1940s	11,946,743	5,644,216	3,497,509	3,056,010	24,144,478
1950s	11,986,437	5,796,823	3,522,545	3,092,375	24,398,180
1960s	11,578,880	5,803,276	3,404,244	3,141,582	23,927,982
1970s	11,626,911	5,755,537	3,383,924	3,002,933	23,769,305
1980s	12,152,603	5,804,320	4,113,254	3,108,775	25,178,952
1990s	13,272,162	7,440,305	4,060,570	3,104,303	27,877,340
2000s	14,590,078	7,678,830	4,088,704	3,121,839	29,479,451
TOTAL	207,633,395	97,207,399	40,124,656	61,266,574	406,232,024

400 million words of published writing across four registers of American English between 1810 and 2009 (see Table 7.1). Roughly half of the words in COHA come from fiction (prose, poetry, and drama). The other half is composed of popular magazines (24%), non-fiction books, balanced across the U.S. Library of Congress classification system (15%), and newspapers (10%). A more complete description of COHA, along with a complete description of all 115,000 texts can be found at http://corpus.byu.edu/coha/.

Corpus Analysis

The first step in our study was to extract the most frequent NNs from COHA. COHA was tagged using the CLAWS part-of-speech tagger. Using these tags, we performed a database query that identified all occurrences of two adjacent nouns. In order to ensure that we represented NNs from across the time periods in COHA, we divided the corpus into six time periods (1810–1840; 1850–1880; 1890–1920; 1930–1950; 1960–1980; and 1990–2000) and required that at least the 400 most frequent NNs from each of the six time periods were included in our data set. We limited our analysis to the 1,535 most frequent NNs. Before

establishing our final list of NNs, we manually eliminated non-NNs that were in the list as a result of tagging errors. Frequency data (per million words) for each decade was recorded for each of these 1,535 NNS. These normed frequency counts were used for the analysis of diachronic change (Research Question 3).

Developing a User-based Instrument for NN Classification

The next two steps of our method, developing a list of semantic categories for NNs and an instrument for NN classification, are described together in this section. This is because these two steps were developed simultaneously in a bottom-up fashion through a series of pilot studies.

We used the list of semantic categories for NNs in the *Longman Grammar of Spoken and Written English* (LGSWE) as a starting point for our research (Biber et al., 1999, pp. 590–591). The LGSWE categories are:

1. Composition
2. Time
3. Location
4. Partitive
5. Specialization
6. Institution
7. Identity
8. Source
9. Purpose
10. Content
11. Objective
12. Subjective

Pilot Study 1

The first pilot study was performed by the two authors ($N = 2$). We began by randomly sampling 100 NNs from our list of 1,535 NNs. We then attempted to independently classify each NN into one of the 12 LGSWE categories. A comparison of the results revealed extremely low inter-rater agreement. After discussing the reasons for the disagreements we realized that we did not even agree on the distinctions between the semantic categories.

Based on this discussion, we made modifications to the list of categories. We also learned from this pilot study that selecting from a long list of semantic categories is a difficult task. This led us to develop a classification instrument in which coders select from a list of sentences, where each sentence presents the NN in question in the form of an explanation of the relationship between the two nouns (e.g. afternoon tea: 'tea is found or takes place at the time of afternoon'). Based on our experience

with the first pilot study, we believed this approach would be superior because it would make the task faster and more natural by eliminating the requirement for coders to memorize and interpret the definitions of the semantic categories.

Pilot Study 2

For the second pilot study we used the instrument described earlier. The coding was performed by university students (N = 42) enrolled in classes taught by the two authors. Each coder classified the same set of 30 NNs, which were randomly selected from the original list of 1,535 NNs. This was done using an online survey tool that presented each of the 30 NNs followed by 12 sentences with the instruction to select the sentence that best represented the meaning of the NN. The results of this study revealed that while perfect agreement among the 42 participants was rare, most of the NNs were coded into one semantic category by a large majority of the coders. This led us to the conclusion that more than two raters were needed to get reliable results for this task.

Based on our results, we made small modifications to the wording of some of the sentences. We also noticed that three of the more general semantic categories were being overused by some of the participants. This led us to modify the instrument so that it began with the question, "Do any of the following describe the meaning of _____?", followed by nine options. Then, after a section break, the survey presented the question, "IF NOT, do any of the following describe the meaning of _____?". This was done with the hope that coders would take the opportunity to select a more specific category, if it was the best choice, rather than repeatedly selecting more general categories that might also make sense within the rephrased sentence. For example, the categories of purpose, topic, and process were overused in the early pilot studies. The modification described here seemed to motivate participants to select the most appropriate option. The list of semantic categories used in this study, along with the rephrased sentences used in the classification instrument, is displayed in Table 7.2.

Pilot Study 3

In the final pilot study, we used the modified version of the survey used in Pilot Study 2. We recruited coders (N = 59) through Mechanical Turk. Together, these workers coded 150 NNs, randomly sampled from the original list of 1,535 NNs, and each NN was coded by eight independent workers. The results of this pilot study were encouraging, showing that most of the NNs were coded into a single semantic category by a majority of the eight coders. Based on these results, we decided that the list of semantic categories and the instrument were ready to be used on a large

Do any of the following describe the meaning of health care?

- ○ care is made from health, (example: *glass window*)
- ○ care is found or takes place at the time of health, (example: *Christmas party*)
- ○ care is found or takes place at the location of health, (example: *corner cupboard*)
- ○ care is one of the parts that make up a(n) health, (example: cat *legs*)
- ○ care is a person, health is what he/she specializes in. (example: *finance director*)
- ○ care is an institution, health is the type of institution, (example: *insurance company*)
- ○ care is owned by health, (example: pirate ship)
- ○ A(n) health care is a(n) health and it is also a(n) care, (example: *exam paper*)
- ○ health is the source of care, (example: *plant residue*)

If NOT, do any of the following describe the meaning of health care?

- ○ health is the topic of care, (example: *algebra textbook*)
- ○ care is a process related to health, (example: *eye movement*)
- ○ health is the purpose or use for care, (example: *pencil case*)

Figure 7.1 An Example of the Final Classification Instrument for NN Sequences

scale for our final analysis. A screenshot of our final instrument can be seen in Figure 7.1.

Classifying NN Sequences

After making several rounds of revision to the list of semantic categories and to the classification instrument, based on three pilot studies, we were prepared to collect data on a larger scale. We recruited a large number of coders (N = 255) through Mechanical Turk. These workers coded the full set of 1,535 NNs described in the Methods section. As with Pilot Study 3, each NN was coded by eight independent workers.

Data Analysis

Agreement and Classification

Agreement was measured using Fleiss' kappa, a statistic for measuring interrater agreement among two or more raters on categorical data. Like

Cohen's kappa, Fleiss's kappa accounts for chance agreement among raters, making it more robust than simple percent agreement. However, unlike Cohen's kappa, Fleiss' kappa does not require coders to be the same for each item, making it ideal for the design of our study. Fleiss' kappa calculations were performed in R, using the 'kappam.fleiss' function in the *irr* package (Gamer, Lemon, Fellows, & Singh, 2012).

After measuring inter-rater agreement, we set out to classify as many NNs as possible into a single semantic category. We began by calculating the number of coders that assigned each of the 1,535 NNs into each of the 13 categories (12 semantic categories plus an 'other' option). We determined that a NN sequence would be assigned to particular category if that category was selected by a plurality of the coders. In our study, we used the following definition for plurality. A particular NN sequence was classified as Category X by a plurality if:

a. It was classified as Category X by 5+ raters; or
b. It was classified as Category X by 4 raters, and no other category was selected by more than 2 coders; or
c. It was classified as Category X by 3 raters, and no other category was selected by more than 1 coder.

Quantitative Analysis

The subset of NNs that met the previously listed agreement criteria was included in this study. These NNs, along with normed rates of occurrence (per million words) for each of the six major COHA time periods, were stored in a spreadsheet. These data were used to compute frequency means for each of the 12 semantic categories in each time period. These means were used to measure diachronic change in the use of the 12 semantic categories. It was also used to perform a factorial ANOVA to measure the effect of time, semantic category, and the interaction between those two variables, on the frequency of use of the NNs in the data set. All statistical procedures were performed in R.

Results and Discussion

Semantic Relationships

Before analyzing the quantitative results of the study, we will first take a closer look at the semantic categories included in the final list. The complete list, along with the rephrased sentence used in the instrument and three examples for each, is displayed in Table 7.2.

These semantic relationships can be organized on a continuum that ranges from *more concrete* to *more abstract*. The categories of Composition, Partitive, and Location are quite concrete, whereas the categories of

Table 7.2 Semantic Categories of NNs, With Rephrased Sentences and Examples

Category	*Rephrased Sentence*	*Examples*
Composition	N2 is made from N1	*brass button* *grape juice* *paper towel*
Time	N2 is found or takes place at the time of N1	*Christmas gift* *autumn leaf* *summer air*
Location	N2 is found or takes place at the location of N1	*library door* *street light* *mountain stream*
Partitive	N2 is one of the parts that makes up a N1	*shirt collar* *chicken breast* *television screen*
Specialization	N2 is a person. N1 is what he/she specializes in	*college professor* *sales manager* *construction worker*
Institution	N2 is an institution. N1 is the type of institution	*police department* *oil industry* *law school*
Identity	A/an N1 N2 is a/an N1 and it is also a/an N2	*patron saint* *bow tie* *minority student*
Source	N1 is the source of the N2	*farm income* *man power* *drug problem*
Purpose	N1 is the purpose or use for N2	*assault weapon* *light bulb* *operating room*
Topic	N1 is the topic of the N2	*tax law* *world news* *science fiction*
Process	N2 is a process related to N1	*data analysis* *air conditioning* *population growth*
Ownership	N2 is owned by N1	*enemy plane* *family mansion* *merchant vessel*

Process, Purpose, and Topic are more abstract. The categories of Ownership and Source fall somewhere in the middle. Scholars in semantics have hypothesized that linguistic forms with concrete meanings tend to develop before forms with more abstract meanings (see Traugott, 1989; Heine, Claudi, & Hünnemeyer, 1991). Based on this, we could hypothesize that concrete NNs were adopted into English first, and more abstract NNs were adopted later.

The final list of semantic categories used in this study is by no means exhaustive. As we will see in the next section, there were many NNs that

coders could not agree on. One reason for these disagreements may be that the actual semantic category for the NN was not presented as an option in the instrument, and coders disagreed on what the next best option was. However, based on the results of our pilot studies and data collection, we believe that this list includes many of the important meaning relationships that can exist between two nouns in a NN. We hope that this list serves as a useful starting place for future research.

Classification Agreement

The overall Fleiss's kappa was .34 for the full set of 1,535 NNs with eight raters and 12 categories (12 semantic categories plus an 'other' option). This can be interpreted as an indication of "fair agreement" according to Landis and Koch (1977).

As discussed in in the Methods section, NNs were assigned to a category if they met our criteria for classification by a plurality of coders. Overall, 974 NNs met these criteria, allowing us to classify approximately 64% of the NNs. Table 7.3 shows the number of NNs that were classified into a semantic category at each level.

The data in Table 7.3 can be analyzed in many different ways. We could focus on the fact that coders were not able to agree on a single semantic category for 36% of the NNs. We were actually quite encouraged by these results despite being unable to achieve agreement on all of the NNs in our data set. A large and varied group of untrained coders managed to classify nearly two thirds of the NNs in our data set, without any situational or linguistic context to aid them. We should also keep in mind that many of the NNs in our data set were most frequent in earlier time periods (e.g. *chain stitch*, *salt pork*, *boon companion*, *tenement house*), increasing the likelihood that they would be unfamiliar to contemporary coders. We do, however, believe that the cases where agreement was not achieved raise important questions that must be addressed in future research. These questions include: What additional semantic categories for NNs exist? Are there NNs that represent hybrids (i.e. they can fit within multiple semantic categories)? Would modifications to the

Table 7.3 Classification Agreement Results

Raters	*NNs*	*Cumulative NNs*
8	133	133
7	171	304
6	174	478
5	238	716
4	230	946
3	28	974

coding instrument improve inter-rater agreement? Would some amount of coder training help improve inter-rater agreement? While the answers to these questions are beyond the scope of the current study, we believe they are important questions for future research that will add to our understanding of the semantics of NNs.

The remaining results in this study are based on the reduced data set of 974 NNs that were classified into a single semantic category. We computed Fleiss' kappa for this data set overall, as well as for each of the 13 semantic categories. The overall Fleiss' kappa was .47, revealing a marked, if unsurprising, improvement over the kappa results for the full data set.

For the reduced data set, we also computed Fleiss' kappa for each of the 13 categories separately. These results are given in Table 7.4. There is a wide range of variation in the inter-rater agreement across these categories, including three categories that could be interpreted as 'substantial agreement' (.61–.80) and three that would be interpreted as 'slight agreement' (.01–.20) (Landis & Koch, 1977).

After reviewing the agreement results, we were curious to know whether coders are more likely to agree on the semantic category of NNs that are more frequent in recent decades than in the distant past. As discussed earlier, in order to learn about diachronic change in the use of NNs, we included NNs that were much more frequent in historical time periods than in contemporary use. However, we believe this may have presented challenges to coders who are unfamiliar with the language used in those earlier time periods. While a comprehensive analysis of this relationship is beyond the scope of this study, we ran a series of simple correlations between the frequency of the NNs across semantic categories and the Fleiss' kappa agreement across those categories. This same correlation was

Table 7.4 Fleiss' Kappa Results Across the 13 Semantic Categories for the Reduced Set of 974 NNs

Category	*Fleiss' kappa*	*Interpretation*
specialization	0.731	Substantial
composition	0.727	
time	0.653	
location	0.569	Moderate
institution	0.542	
purpose	0.385	Fair
process	0.369	
partitive	0.262	
source	0.259	
ownership	0.231	
identity	0.206	Slight
other	0.197	
topic	0.162	

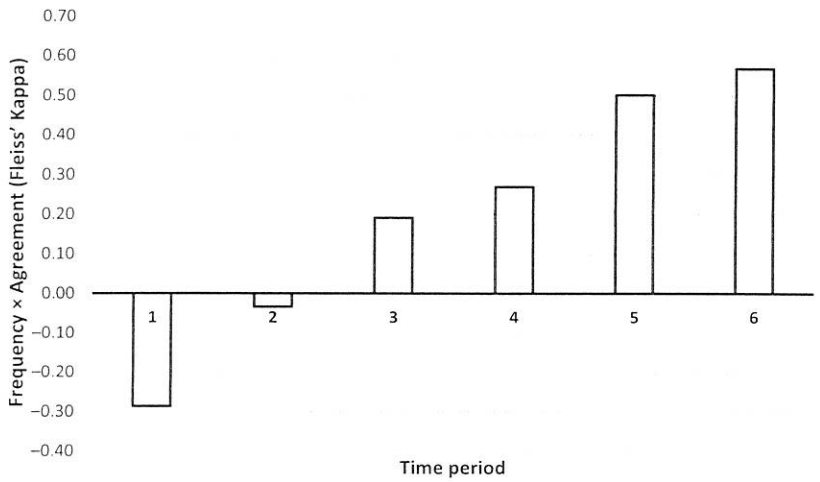

Figure 7.2 NN Frequency by Agreement Correlations Across Time Periods

run for each of the six time periods included in this study. The results are displayed in bar plot form in Figure 7.2. The six time periods are laid out on the *x*-axis and the correlations are plotted on the *y*-axis. It can be seen from this plot that there is a strong relationship between inter-rater agreement among the coders and the frequency of the NNs across time periods. Whereas there is a moderate negative correlation between these two variables in the earliest time period (1810–1840), there is a strong positive correlation in the two most recent time periods (1960–1980; 1990–2000). From these results we can draw two conclusions: (i) there is a relationship between NN frequency and inter-rater reliability; and (ii) this relationship is much more pronounced for recent data sets than for older ones.

Diachronic Change

To answer the third research question, we investigated diachronic changes in the use of NNs across the 12 semantic categories. Our results confirmed findings from previous research by showing that NNs are increasing in frequency over time. We found that this pattern holds true across all of the semantic categories (see Table 7.5).

However, our analysis of the data revealed that this general pattern only tells part of the story. There is a large amount of variability across semantic categories in their initial frequency in the earliest time period of the corpus, as well as in their frequencies in the most recent time period. Moreover, the rate of increase over time varies widely across semantic

Table 7.5 Normed Rates of Occurrence (per Million Words) of NN Semantic Categories Across Six Time Periods

Category	*1810–1849*	*1850–1889*	*1890–1929*	*1930–1959*	*1960–1989*	*1990–2009*
Institution	45.06	45.18	110.01	225.24	251.22	278.27
Location	83.81	101.43	152.57	196.94	188.77	207.03
Specialization	19.25	39.06	100.47	158.68	172.52	268.88
Purpose	23.60	20.74	51.00	157.75	214.94	256.35
Composition	85.52	98.13	95.78	115.87	103.39	127.06
Process	10.07	11.60	28.43	79.94	125.64	204.90
Source	33.98	29.02	35.86	56.66	45.12	66.04
Time	36.94	46.71	40.45	49.34	37.51	45.77
Topic	16.56	19.52	26.68	38.75	51.28	94.81
Identity	14.26	11.48	22.21	50.63	48.69	56.53
Partitive	3.53	3.12	7.42	20.21	34.67	70.27
Ownership	4.69	4.20	7.27	11.89	9.96	12.98
Other	16.29	12.62	15.11	26.07	28.22	35.19

category. In the next section we describe the results of a factorial ANOVA aimed at accounting for the effect of time and semantic category, as well as a possible interaction between them.

Statistical Results

We performed a 6 × 13 factorial ANOVA to determine the statistical effects of the variables of time and semantic category on variation in the frequency of NNs in the data set. The results of this analysis revealed a significant interaction effect between the variables of time and semantic category, $F(60, 5766) = 4.18$, $p < .0001$, $R^2 = .03$. This shows that diachronic change in the use of a particular NN is likely to depend on the semantic relationship between the two nouns. NN frequency was also predicted by semantic category, $F(12, 5766) = 3.50$, $p < .0001$, $R^2 = .006$, and time, $F(6, 5766) = 119.67$, $p < .0001$, $R^2 = .08$.

However, in the presence of an interaction, it is appropriate to investigate simple effects (i.e. differences between the levels of variable 1 within each level of variable 2, separately) rather than the overall main effects produced by the factorial ANOVA. Thus, our next step was to describe the diachronic trends in the use of NNs across each of the 13 semantic categories. Although each of the categories follows a slightly different trend, we managed to identify three major patterns of diachronic change:

1. Frequent → frequent
2. Infrequent → infrequent
3. Infrequent → frequent

In the next three sections we present descriptions of these patterns, including quantitative and qualitative findings from our data.

Pattern 1: Frequent → Frequent

The first pattern we discovered in our data set was that two semantic categories, location and composition, were more frequent than the other categories in the earliest time period (Figure 7.3). Whereas most of the categories occurred less than 50 times per million words (pmw), these two categories occurred about 85 times per million words. These categories share the characteristic of being highly concrete. Composition NNs (e.g. *stone wall*, *orange juice*, *gold watch*, *wood door*) include tangible objects, where N2 is a concrete head noun and N1 is another concrete noun that specifies what N2 is made from. In location NNs (e.g. *kitchen table*, *heart disease*, *forest fire*, *mountain resort*) N1 describes the place where N2 exists or takes place. Of all the semantic categories in our framework, these two categories are the most concrete. This suggests that the NN grammatical structure may have begun as a means of describing concrete objects and processes and was later adopted for the description of more abstract concepts and processes. This supports the previously mentioned hypothesis that semantic change often shifts from concrete meanings to abstract meanings (see Traugott, 1989; Heine et al., 1991).

The categories of location and composition have both increased in use over time. Based on the data in COHA, the rate of increase for the category of composition has been relatively modest over time, with a net increase of just over 40 occurrences pmw from time period 1 to time period 6. In contrast, the location category has increased in use much

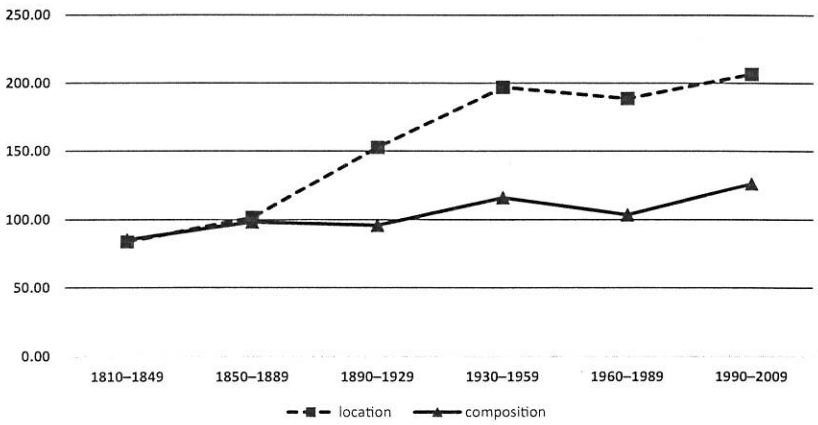

Figure 7.3 Diachronic Change in the Semantic Categories Within Pattern 1—Frequent → Frequent

more rapidly, with a net increase of more than 120 occurrences pmw. Although these two categories are not the most frequently used semantic categories for NNs, they are both more frequent than more than half of the semantic categories. It is interesting to note that location and composition NNs, combined, made up more than 40% of all NNs in our data set during the period of 1810–1849. This proportion shifted quite dramatically to a mere 19% in the most recent time period, showing the relatively rapid diachronic spread of NN constructions into other semantic domains.

Pattern 2: Infrequent → Infrequent

The second pattern comprises semantic categories that have occurred in every time period, but which have remained consistently infrequent relative to the other semantic domains (Figure 7.4). This list includes the following semantic categories: time (e.g. *summer day*), identity (e.g. *student teacher*), partitive (e.g. *family member*), topic (e.g. *algebra text*), source (e.g. *government policy*), ownership (e.g. *police car*), and other.

The semantic categories in Pattern 2 are less frequent than the categories in the other two patterns. All of these categories are increasing, just at different rates. Some of these categories, such as source, ownership, identity and time, are only about 2–3 times more frequent in the most recent time period when compared with the earliest time period. Identity NNs are four times more frequent, and topic NNs are six times more frequent. The semantic category in this pattern that is increasing most rapidly is the partitive category, which is nearly 20 times more frequent

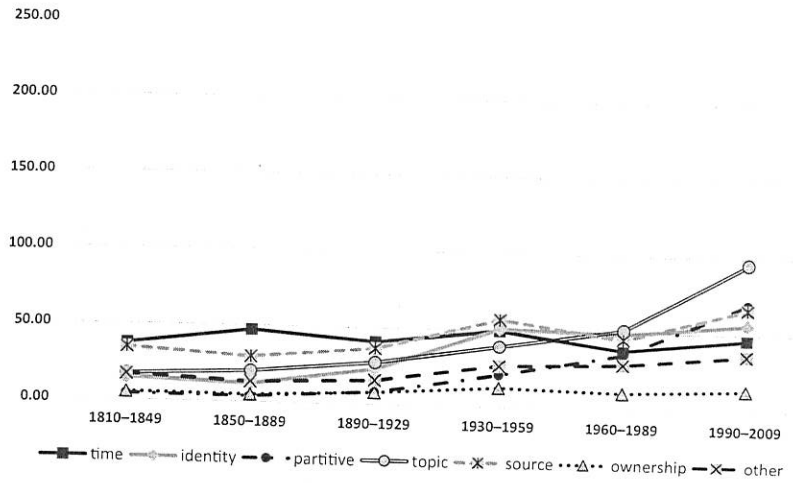

Figure 7.4 Diachronic Change in the Semantic Categories Within Pattern 1—Infrequent → Infrequent

in time period 6 when compared with time period 1. The semantic categories within this pattern represent the broad range of semantic relationships that can occur in NN sequences even if they are not nearly as frequent in contemporary American English writing as the categories in the other two time periods.

Pattern 3: Infrequent → Frequent

The semantic categories within Pattern 3 are particularly interesting since these NNs have undergone the most rapid changes in frequency during the past 200 years (Figure 7.5). The four categories in this pattern include: institution (e.g. *insurance company*, *stock market*), specialization (e.g. *police officer*, *government official*), purpose (e.g. *credit card*, *golf course*), and process (e.g. *tax cut*, *birth control*). On average, these categories are 13 times more frequent in period 6 than in period 1. Moreover, these categories comprise the four most frequent meaning relationship categories for NNs, occurring more than 250 times pmw, on average.

A closer look at the particular NN sequences within these four categories suggests that the rapid increase in the use of these categories reflects societal changes in the United States. The most salient of these changes seem to be related to specialization—of knowledge, labor, industry, and day-to-day processes. This shift in the semantics of NNs seems to correspond to the development of increasing specialization in scientific disciplines, government, commerce, job descriptions, and technology, among many others. Obviously, there is much more to the story since this shift

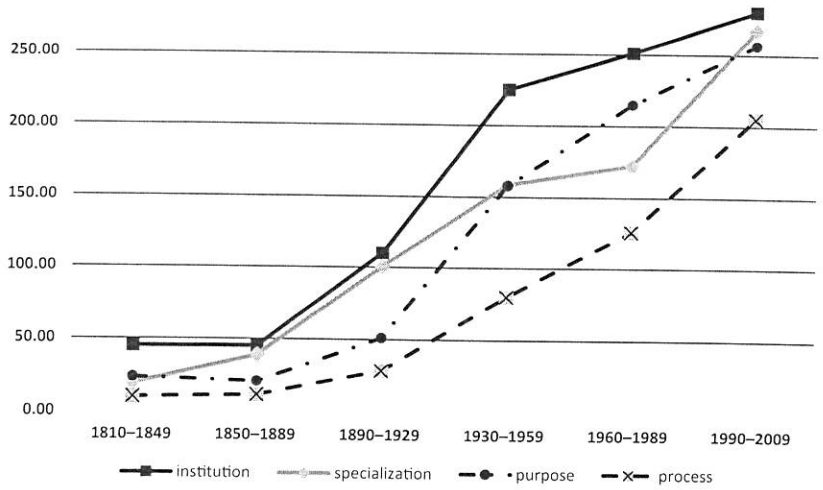

Figure 7.5 Diachronic Change in the Semantic Categories Within Pattern 1—Infrequent → Frequent

toward specialization could have been expressed in other ways in American English. The other piece to the puzzle seems to be a shift toward increased economy in the language that manifests itself in the use of compressed phrases rather than elaborated clauses, especially in writing (see, e.g. Biber & Gray, 2016).

Conclusion

Summary of Findings

Our end goal in this study was the description of diachronic change in the use of NNs across categories representing distinct semantic relationships between N_1 and N_2 (RQ 3). Before it was possible to answer that question, however, it was necessary to establish a list of the semantic relationships that are possible between two nouns in an NN (RQ 1) and determine whether non-experts could reliably classify NNs based on the meaning relationship between the two nouns (RQ 2).

To address RQ 1, we developed a list of 12 semantic categories for NNs. This list was initially based on Biber et al. (1999), but we made several revisions to it based on a series of pilot studies. During those same pilot studies, we refined an instrument that could be used by non-expert English speakers for classifying NNs into semantic categories. In answer to RQ 2—can non-expert language users reliably classify NNs into semantic categories?—we would answer 'yes'. Using this method, we managed to classify nearly two thirds of the NNs in our data set. However, we also found considerable variation in reliability across the semantic categories, suggesting that some of these categories are much better defined than others in the minds of language users. We also found an effect of time on inter-rater reliability. Raters were much more likely to agree on NNs that are more frequent in recent time periods than those that are more frequent in earlier time periods.

Our analysis of historical change in the use of NNs confirmed that the NN construction is increasing over time in written English. The results of this study have also shown that NNs are becoming increasingly productive, with the construction rapidly spreading across a wide range of semantic categories over the course of 200 years, a relatively short period of time in historical linguistics. We have also shown that different semantic categories have developed over time in very different ways. While all of the semantic categories are increasing over time, we showed a statistical interaction between time and semantic category. We then explored three underlying patterns of historical development. The first pattern includes two semantic categories that have been relatively frequent in all time periods. The second pattern includes seven categories that began with low frequencies and have experienced relatively small increases in frequency over time. The final category includes four categories that have

undergone rapid increases in frequency over time, beginning with relatively low frequencies and ending with the highest frequencies in the data set. We believe that one explanation for these changes is the hypothesis that semantic change moves from more concrete to more abstract meanings.

On Methodological Triangulation

The use-based corpus methods in this study were quite straightforward. The user-based methods, on the other hand, presented many challenges. Our experience was that non-expert raters perform best when the instrument relies on terminology and tasks they are already familiar with. In our case, that meant eliminating the names we had developed for the semantic categories and structuring the instrument in the form of a series of simple sentences. We had a similar experience in another project focussed on register classification in which coders performed better when register labels were replaced with familiar situational characteristics (e.g. spoken/written, interactive/non-interactive) (Egbert et al., 2015). We also found that it was not realistic to expect high inter-rater reliability between two raters in the task of classifying NNs. However, we were able to classify most of the NNs when we used eight raters and focussed on plurality agreement rather than inter-rater reliability. We had similar experiences in previous research that triangulated user-based and use-based corpus research methods (see Egbert et al., 2015; Egbert, 2014).

This study would not have been possible without the combination of use-based corpus data and the user-based semantic classification of human raters. High frequency NNs extracted from the corpus (use-based) were classified by raters (user-based) in a series of pilot studies that informed the development of a list of semantic categories and a NN classification instrument. This instrument was used by hundreds of raters (user-based) to classify 1,535 NNs extracted from the corpus (use-based) to represent patterns from six time periods. Finally, diachronic change in the frequencies of each semantic category of NNs (use-based) was measured for each of the categories that were established by raters (user-based). The iterative use-based and user-based stages of this research process combined to create a robust research methodology. This methodology provided data to address the question of historical change in English NNs across semantic categories.

The results of this study demonstrate the value of triangulating corpus linguistic methods with other methods in linguistics. Although it would have been possible for us as linguistic researchers to attempt to classify each of the NNs in our data set, we believe there were major advantages to using non-expert language users for the task. The most obvious advantage was that we were able to classify all of the NNs in our data set in a fraction of the time that it would have taken us. More importantly, we

learned a great deal in this study about the semantics of NNs that we could not have learned if we had attempted to classify them ourselves. We learned that:

1. users can think about the meaning relationship between two nouns in a NN by rephrasing it in the form of a sentence.
2. some semantic categories are much easier for users to agree on than others.
3. users are much more likely to agree on the semantic category of a NN if it is frequent in contemporary English.
4. users can usually identify the semantic relationship between two nouns in a NN without (i) any linguistic context or (ii) grammatical cues.

We believe that discoveries such as these are extremely valuable because they answer questions that are unanswerable using corpus data alone.

References

Asheghi, N. R., Sharoff, S., & Markert, K. (2016). Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, *50*(3), 603–641.

Biber, D., & Egbert, J. (2016). *Register variation online*. Cambridge: Cambridge University Press.

Biber, D., Egbert, J., Gray, B., Oppliger, R., & Szmrecsanyi, B. (2016). Variationist versus text-linguistic approaches to grammatical change in English: Nominal modifiers of head nouns. In M. Kyto & P. Paivi, (Eds.), *Handbook of English historical linguistics*. Cambridge: Cambridge University Press.

Biber, D., & Gray, B. (2011). The historical shift of scientific academic prose in English towards less explicit styles of expression: Writing without verbs. In V. Bathia, P. Sánchez, & P. Perez-Paredes (Eds.), *Researching specialized languages* (pp. 11–24). Amsterdam: John Benjamins.

Biber, D., & Gray, B. (2013). Being specific about historical change: The influence of sub-register. *Journal of English Linguistics*.

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge, Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Pearson.

Davies, M. (2010). *The Corpus of Historical American English (COHA): 400 million words, 1810–2009*. Available online at http://corpus.byu.edu/coha/.

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, *53*(4), 810–842.

Egbert, J. (2014). *Reader perceptions of linguistic variation in published academic writing* (Unpublished Ph.D. dissertation). Northern Arizona University, Arizona.

Egbert, J. (2016). Stylistic perception. In P. Baker & J. Egbert (Eds.), *Triangulating methodological approaches in corpus linguistic research*. New York, NY: Routledge.

Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, *66*(9), 1817–1831.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *Irr: Various coefficients of interrater reliability and agreement. R package version 0.84*. Retrieved from https://CRAN.R-project.org/package=irr

Girju, R., Moldovan, D., Tatu, M., & Antohe, D. (2005). On the semantics of noun compounds. *Computer speech & language*, *19*(4), 479–496.

Heine, B., Claudi, U., & Hünnemeyer, F. (1991). *Grammaticalization*. Chicago and London: The University of Chicago Press.

Jurgens, D. (2013, June). Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 556–562).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Lauer, M. (1995). *Designing statistical language learners: Experiments on noun compounds* (Unpublished Ph.D. Thesis). Macquarie University, Australia.

Lees, R. (1970). Problems in the grammatical analysis of English nominal compounds. In M. Bierwisch & K. E. Heidolph (Eds.), *Progress in linguistics* (pp. 174–186). The Hague: Mouton.

Levi, J. N. (1974). *On the alleged idiosyncrasy of non-predicate NPs*. Papers from the 10th regional meeting, Chicago, Chicago Linguistic Society (pp. 402–415).

Nakov, P., & Hearst, M. (2006). Using verbs to characterize noun-noun relations. In *International conference on artificial intelligence: Methodology, systems, and applications* (pp. 233–244). Berlin, Heidelberg: Springer.

Rosenbach, A. (2006). On the track of noun+noun constructions in Modern English. In C. Houswitschka, G. Knappe, & A. Müller (Eds.), *Anglistentag 2005 Bamberg: Proceedings of the conference of the German Association of University Teachers of English* (pp. 543–557). Trier: Wissenschaftlicher Verlag Trier.

Rosenbach, A. (2007). Emerging variation: Determiner genitives and noun modifiers in English. *English Language and Linguistics*, *11*(1), 143–189.

Rumshisky, A. (2011). Crowdsourcing word sense definition. In *Proceedings of the 5th linguistic annotation workshop* (pp. 74–81). Prague: Association for Computational Linguistics.

Szmrecsanyi, B., Biber, D., Egbert, J., & Franco, K. (2016). Towards more accountability: Modeling ternary genitive variation in late modern English. *Language Variation and Change*, *28*(1), 1–29.

Traugott, E. C. (1989). On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language*, *65*(1), 31–55.

Zimmer, K. E. (1971). Some general observations about nominal compounds. Working papers on language universals, *Stanford University*, *5*, 1–21.

8 Corpus Linguistics and Event-Related Potentials

Jennifer Hughes and Andrew Hardie

Introduction

Collocation can be defined as a "co-occurrence relation between two words" (McEnery & Hardie, 2012, p. 240), with collocation extraction being one of the key techniques used in corpus linguistics. Indeed, Gilquin and Gries (2009) find that 32% of a sample of 81 corpus research articles use collocation analysis. However, despite the prevalence of collocation analysis, it is not clear whether corpus-derived collocations are actually processed differently by the brain from non-collocational sequences, and therefore whether collocation can be seen as being a plausible psychological phenomenon.

There is growing recognition of the importance of combining corpus data with experimental work. For instance, Arppe, Gilquin, Glynn, Hilpert, and Zeschel (2010, p. 6) point out that "linguists have made relatively few efforts up until now to test the cognitive reality of corpora". Likewise, Gries (2014, p. 12) argues that "there will be, and should be, an increase of corpus-based studies that involve at least some validation against experimental data". Some psycholinguistic studies have attempted to ascertain the psychological validity of corpus-derived collocations using techniques such as eye-tracking and self-paced reading (e.g. Conklin & Schmitt, 2008; McDonald & Shillcock, 2003a, 2003b; Underwood et al., 2004; Huang, Wible, & Ko, 2012). The results of these studies reveal that sequences of words which form collocations are read more quickly and receive fewer eye fixations than sequences of words which do not form collocations. These studies therefore provide strong evidence in support of the validity of collocation as a psychological phenomenon. However, to ascertain whether or not there exists a *neural* correlate of corpus-derived collocations, it is necessarily to combine corpus data with neuroimaging (also known as brain imaging) techniques—and thus study processing activity directly rather than via proxy variables such as eye movements.

One such imaging technique is *electroencephalography* (henceforth EEG), where electrodes placed across the scalp detect some of the electrical