Chapter 18

Diachronic collostructional analysis meets the noun phrase

Studying *many a noun* in COHA

Martin Hilpert

1. Introduction

This chapter presents the approach of diachronic collostructional analysis (Hilpert 2006, 2008, forthcoming), a method for studying semantic and stylistic change in grammatical constructions. It is used here on the basis of the Corpus of Historical American English (COHA) in order to analyze recent change. Whereas much work in collostructional analysis has addressed verb-headed constructions, such as modal auxiliaries, causatives, the ditransitive construction, or the passive, relatively little attention has been paid to the nominal domain. The focus here is on the English *many a noun* construction, which is illustrated in (1), and which undergoes a recessive change over the past 200 years.

(1) Many a day will pass before this construction is properly understood.

This construction has been chosen because it deviates from more canonical noun phrase patterns and shows several restrictions: first, quantifiers do not typically precede determiners. Second, the construction is limited to *many*; semantically related elements such as *few*, *little*, *much*, or *lots* do not form analogous patterns. Third, as observed by Huddleston and Pullum (2002: 394), no elements may intervene between *many* and *a*.

Diachronic collostructional analysis, which has been adapted from Gries and Stefanowitsch (2004), uses temporally ordered corpus data to track historical shifts in collocational patterns. For instance, the English *be going to V* construction changes diachronically with respect to the items that typically fill the verb slot (Hilpert 2008). Shifts such as these indicate developments in constructional meaning—as the construction changes

semantically, it comes to be used with different collocates. Newly incoming collocates not only show that some change is underway; their lexical meanings further indicate how the construction changes semantically. Analyses of this kind can be used for exploratory studies, as well as for analyses that test existing hypotheses about semantic change, for instance in the area of grammaticalization studies.

The remainder of this chapter is structured as follows. Section 2 walks the reader through the steps of a diachronic collostructional analysis; section 3 offers an interpretation of the results that are obtained; and section 4 concludes with pointers toward theoretical issues and problems that have to be kept in mind.

2. Collostructional analysis: Methodology

The term "collostructional analysis" refers to a family of methods for the study of interrelations between grammatical constructions and their lexical collocates (Stefanowitsch and Gries 2003, 2005; Gries and Stefanowitsch 2004). The common goal of these methods is to identify lexical elements that are "typical" of a given grammatical construction. To illustrate, the *many a noun* construction can be used with almost any noun of the English language. While there is a large amount of variation in the noun slot, there are also robust tendencies of certain nouns appearing more often than others. In just about any large corpus of English, a raw concordance of the construction will reveal that nouns such as *time*, *day*, *man*, and *year* occur most frequently. Before it can be concluded that these are in fact the lexical elements that are most typical of the construction, their overall text frequency needs to be controlled for. The observed frequency of *many a time* thus has to be compared against the overall frequency of the noun *time* in the corpus that is used. If this control is implemented, it may emerge that the most frequent nouns are not necessarily the ones most typical for the construction. Table 1 shows the ten most frequent nouns of the *many a noun* construction in COHA, along with their overall text frequencies. The time expressions *time*, *day*, *year*, *night*,

and *hour* clearly dominate the picture in terms of raw frequency, but the element that occurs most often in the construction, given its corpus frequency, is the noun *mile*. The percentages in the third column of Table 1 show the differences between the nouns. Collostructional analysis identifies those lexical elements that occur with disproportionately high frequency and thus determines which ones are most typical.

[INSERT TABLE 18.1]

The main purpose of a collostructional analysis then is a semantic study of a construction via its most typical collocates. For the *many a noun* construction, the salience of time nouns suggests that the meaning of a recurrent or prolonged situation is a deeply entrenched semantic schema. This characterization however cannot be the whole story: frequent patterns with persons, as in *many a man*/*woman*/*heart* would have to be explained as alternative schemas. A synchronic collostructional analysis would aim to capture these schemas and to assess their relative importance and semantic interrelations.

Applied to diachrony, a collostructional analysis also investigates associations between lexical elements and grammatical constructions. However, as an additional layer of complexity, it considers diachronic shifts in these associations. Do some associations become weaker or stronger over time? Are the most strongly attracted lexical elements from an initial corpus period still typical of the construction at some later historical stage? The value of these questions lies in the view that shifting collocational patterns reflect semantic change. Looking at collocational change allows the analyst to study meaning change in real time.

Methodologically, a diachronic collocational analysis involves two steps. A first, computational step determines the most typical collocates for each corpus period. In a second step, these lists of lexical elements are interpreted semantically. The researcher has to select criteria that may be compared across the periods in order to reveal differences between them. The choice of these criteria is necessarily subjective and open-ended.

Diachronic collostructional analysis is an adaptation of distinctive collexeme analysis (Gries and Stefanowitsch 2004), which was originally designed to compare two or more constructions in their synchronic collocational behavior. Instead of making comparisons across two constructions, a diachronic collostructional analysis focuses on just one construction, comparing the frequencies of its lexical collocates across sequential corpus periods. If there are diachronic differences in the typical collocates of a construction, these can reveal how a construction changed semantically.

The following paragraphs detail the working steps of a diachronic collostructional analysis, using the example of the *many a noun* construction. After the first step of data collection, it is described how the data is divided into periods. A computational analysis of the partitioned data produces results for subsequent interpretation.

2.1 Data collection

A near-exhaustive concordance of the *many a noun* construction was performed by searching COHA for the form *many*, followed by the indefinite determiners *a* or *an*, up to two optional modifying elements, and a subsequent head noun. Spot checks show that this procedure generates a tolerably low number of false positives and double hits. The resulting database contains approximately 15,000 examples of the construction; the examples were produced between the 1810s and the 2000s. The text frequency of the construction steadily declines during that time.

[INSERT FIGURE 18.1]

The database used for all subsequent analyses contains the raw frequencies of all attested noun types within their respective decades of production. Overall, the construction occurs with 3,340 noun types, ranging from *aborigine* to *zloty*. Table 2 shows an excerpt of that data. While such a database already holds the information which types appear most frequently overall or in a given decade, the driving question of a diachronic collostructional

analysis is, of course, whether some noun types vary over time in their typicality of the construction. Put simply, did a phrase such as *many a time* have a heyday during which it was more in fashion than at other times? In order to find out, comparisons between different historical periods have to be made. COHA holds data from 20 decades; a diachronic comparison could thus simply contrast these. However, comparing 20 sets of collocates that exhibit a fair amount of semantic overlap would be an unwieldy exercise—a smaller number of periods is desirable from a practical perspective. The next section discusses how these periods are chosen.

[INSERT TABLE 18.2]

2.2 Data periodization

Gries and Hilpert (2008, this volume) argue that it is useful to divide one's data in a way that reflects the phenomenon that is being studied. In the case of *many a noun*, one possible way of dividing the data uses the frequency development shown in Figure 1, grouping together data points that show relatively similar frequencies, and creating period breaks where there are substantial changes. To this end, a hierarchical clustering algorithm (*Variability-based Neighbor Clustering*, VNC) merges neighboring data points, starting with the most similar. Further technical detail is offered in Gries and Hilpert (this volume, cf. also the supplementary web materials). When applied here, the VNC algorithm produces the periodization shown in Table 3.

[INSERT TABLE 18.3]

The periods in Table 3 reflect mutual similarities of frequency and are thus not equidistant, although comparable in length and size. An obvious exception is the first period, which holds only one decade. For that reason, and since collostructional analyses are relatively data-intensive procedures, the subsequent analysis only uses VNC periods two to five.

2.3 Data processing

The information on which a diachronic distinctive collexeme analysis is based is a table that holds the frequencies of all noun types for each of the four periods. Table 4 shows the 10 most frequent elements for each of the four periods.

[INSERT TABLE 18.4]

For each cell in the table, the analysis determines whether the observed frequencies differ from the expected frequencies. An exact binomial test determines whether the observed frequency is significantly higher (or lower) than expected. The exact p-value that is returned by the test is taken as a numerical measure of how distinctive a given lexical item is for the given period of the *many a noun* construction. This means that the p-values are not just used as categorical indicators of significance; nouns with more uneven distributions and lower p-values are judged to be relatively more distinctive of the respective periods. Sorting all nouns by their p-values yields a hierarchy of distinctive elements. The overall output of a diachronic distinctive collexeme analysis is thus a collection of lists that shows the significantly distinctive elements for each of the corpus periods. The computation is performed by a script for the software package R (Gries 2004, cf. the companion web page for this volume).

2.4 Results

Table 5 shows the 15 most distinctive elements for each of the four corpus periods. All nouns shown are significantly distinctive at $p < .01$ (Coll.Str > 2) or $p < .001$ (Coll.Str > 3).

[INSERT TABLE 18.5a AND 18.5b]

In each case, more distinctive elements could have been reported, but the elements that are shown give a good enough representation of the respective periods and of the changes that have taken place. The next section offers a qualitative interpretation of the numerical findings, going over each individual period before characterizing the semantic development as a whole.

3. Interpretation of the results

The distinctive collexemes of the first period (1820s–1860s) include the nouns *heart*, *tear*, *sigh*, and *pang*, which relate to the domain of human emotion. As the following COHA examples show, also *eye*, *spirit*, *bosom*, *thought*, and *prayer* are found in emotionally charged contexts.

(2) a. Still smiles the sun;—but many an eye shall weep

b. The war may have solved the problem for many a desperate spirit.

c. a sigh of relief went up from many a Republican bosom

d. Around it clings many a thought of desperate battle, of hope and fear

e. many and many a prayer of gratitude burst from Darina's lips

There is thus evidence that the *many a noun* construction used to be closely connected to the frame of human emotions. This idea is corroborated by examples with the most frequent distinctive collexeme of the first period—*hour*. While this element may appear to just instantiate the time noun pattern that still prevails today, examples with *hour* in the first period are typically modified by adjectives such as *weary*, *happy*, *pleasant*, and *lonely*, thus connecting the time period to an emotional experience.

The second period (1870s–1900s) marks the rise of time nouns as the predominant lexical class in the *many a noun* construction: *time*, *day*, and *year* are the three most distinctive collexemes, and also by far the most frequent items. Examples with these nouns show that usually a repeated or prolonged experience of a human being is at issue.

(3) a. I've thought that many a time myself.

b. Why, that girl's face will haunt me for many a day.

c. Jahez Gorham had been for many a year a successful manufacturer of jewelry.

The remaining items of that period do not form a single coherent class, but nouns such as *frolic*, *struggle*, *escapade*, and *reverie* still attest to the earlier tendency of the construction to occur with emotion-evoking nouns.

The third period (1910s–1940s) represents yet another developmental stage. Ten out of the 15 distinctive collexemes denote human beings. Within this broad category, the dimensions of politics, profession, and society seem particularly important: *citizen*, *bigwig*, and *Republican* are understood against the backdrop of US politics; *scientist*, *manufacturer*, and *educator* are vocational terms; and the nouns *reader* and *observer* are used to describe consumers of information as responsible citizens, as shown in the examples below. What these examples share with the ones presented above is the fact that they denote a human experience. The collocational shift to human beings is not only a semantic change but also reflects a change in style: The construction is no longer primarily used in literary texts, but it gains popularity in journalistic writing, especially magazines (cf. Biber and Gray, this volume).

(4) a. But many a newspaper reader was skeptical or confused.

b. the riddle continues to puzzle many a political observer.

The semantic focus on human experiencers in sociopolitical roles is kept up in the fourth period (1950s–2000s), during which the construction becomes ever less frequent. The distinctive collexemes *investor*, *businessman*, *politician*, *executive*, and *conservative* attest to this continuing trend, and to the continuing shift into journalistic writing. Beyond that, the nouns *moon* (as in *many a moon will pass*) and *weekend* illustrate the persisting schema of time nouns; the remaining distinctive collexemes show little semantic unity and low raw frequencies. It thus appears that the construction does not spawn any new semantic offshoots during this time; it merely retains old schemas while the overall pattern gradually becomes less frequent.

Overall, the progression of the four periods shows that a change in collocational interdependencies has taken place and that this change involves the semantic clusters of emotionally charged nouns, time nouns, and nouns denoting human experiencers. To the present-day speaker of American English, the phrases *many a time* or *many a day* will most readily come to mind as examples of the *many a noun* construction because they represent the schema with the highest token frequencies. However, the other two schemas remain productive and continue to carry semantic import. A phrase such as *many a dog-owner* thus has a slightly different meaning than the alternative phrase *many dog-owners*: due to the human experiencer schema, a phrase such as *many a dog-owner* is relatively more likely to continue with a statement about typical experiences, preferences, or emotional responses. The unmarked alternative is more likely to be followed by a matter-of-fact statement. While the difference is subtle, writers appear to exploit it and thus occasionally use the construction as a marker of style.

From a variationist perspective, the availability of two similar structures (*many a dog-owner*, *many dog-owners*) for the expression of similar meanings raises questions regarding the demise of the *many a noun* construction. Did the canonical pattern subsume functions that were previously conveyed by the non-canonical construction? Was there a frequency trade-off between the two patterns? Data from the COHA suggests that such effects, if existing at all, are minor (cf. supplementary web materials). The *many a noun* construction was, from its beginning, a stylistically marked, peripheral grammatical device. Hence, its diachronic demise does not cause substantial ripple effects.

4. The bigger picture

The preceding sections have presented a case study of a diachronic collostructional analysis. To conclude this chapter, it will be useful to consider a few general questions about the methodology, especially in the context of recent change.

A first concern about the collostructional approach that could be raised would be its reliance on subjective assessments. The semantic categorization of the distinctive collexemes and the extent to which full examples are taken into consideration is clearly a matter of qualitative analysis. The shorter and more recent the time span that is investigated, the more problems may arise in this regard. In the case of the *many a noun* construction, the four different corpus periods arguably show discernible differences in their respective distinctive collexemes, but not everyone looking at the results would arrive at the same conclusions. The purpose of the collostructional approach therefore cannot be to obliterate this kind of work. Its value lies in the fact that it uses quantitative data to make qualitative phenomena available for inspection that would otherwise remain inaccessible.

Second, it has to be kept in mind that distinctive collexeme analysis works on the basis of raw frequencies rather than normalized frequencies. Hence, if certain noun types are unevenly distributed across the COHA periods, this will affect the results. The question that arises is whether for instance *eye* and *heart* are distinctive for the first period because the *many a noun* construction changed, or just because these nouns are generally more frequent in that period. As well-balanced as COHA is, there will always be artifacts of sampling. It is, for instance, no wonder that the noun *war* occurs as a distinctive collexeme in the period between the 1910s and 1940s.

A related, third issue concerns the variability that is inherent in corpora that comprise different genres. Collocations such as *many a heart*, which are typical for early uses of the construction, differ from collocations such as *many a businessman* not only semantically but, of course, also with regard to the registers in which they are usually found. A frequency analysis of the retrieved examples indicates that the *many a noun* construction has over time become substantially more frequent in magazines and nonfiction texts (cf. supplementary

web materials). The differences between the present-day collocates and the historical collocates should thus be interpreted in terms of concurrent stylistic and semantic change.

Fourth, it is worth considering the application of collostructional analyses to matters of linguistic theory. The present chapter has limited itself to an exploratory study of semantic and stylistic change. A more difficult, but also more rewarding, application of collostructional methods is to bring them to bear on theoretical hypotheses. For example, Hilpert (2008: 183–86) investigated whether shifting collocational preferences of Germanic future constructions developed in accordance with preexisting claims about grammaticalization paths of these constructions, thus corroborating some earlier accounts while falsifying others. With regard to the topic of recent change, one particular strength of the collostructional approach is its potential to uncover processes of incipient grammaticalization: the method tracks the influx of new lexical types, and when newly attracted types violate earlier selection restrictions of the construction, this is evidence for a trajectory toward more abstract, grammatical meaning.

As a fifth and final point, another asset of the approach presented here is that it brings into focus the important role of the lexicon in the domain of recent change. Whereas diachronic corpora typically show the results of morphological and syntactic change only with a certain delay, due to the normative effect of writing, collocational change frequently proceeds under the radar of prescriptive influence, and is thus recorded immediately. Summing up, this chapter has hopefully shown that diachronic collostructional analysis can be made useful for the study of recent change in a variety of ways.

References

Gries, Stefan Th. 2004. Coll.analysis 3. A program for R for Windows. Last modified 25 December 2010. http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/.

Gries, Stefan Th., and Martin Hilpert. 2008. 'The Identification of Stages in Diachronic Data: Variability-based Neighbor Clustering'. *Corpora* 3: 59–81.

Gries, Stefan Th., and Anatol Stefanowitsch. 2004. 'Extending Collostructional Analysis: A Corpus-Based Perspective on "Alternations"'. *International Journal of Corpus Linguistics* 9: 97–129.

Hilpert, Martin. 2006. 'Distinctive Collexeme Analysis and Diachrony'. *Corpus Linguistics and Linguistic Theory* 2: 243–57.

———. 2008. *Germanic Future Constructions: A Usage-based Approach to Language Change*. Amsterdam: Benjamins.

———. 2012. 'Diachronic Collostructional Analysis: How to Use It and How to Deal with Confounding Factors'. In *Current Methods in Historical Semantics*, ed. Kathryn Allan and Justyna Robynson,133-60. Berlin: De Gruyter Mouton.

Huddleston, Rodney D., and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Stefanowitsch, Anatol, and Stefan Th. Gries. 2003. 'Collostructions: Investigating the Interaction of Words and Constructions'. *International Journal of Corpus Linguistics* 8: 209–43.

———. 2005. 'Covarying Collexemes'. *Corpus Linguistics and Linguistic Theory* 1: 1–46.

Figure 1.

Frequency development of the *many a noun* construction

Table 1.

Noun frequencies in COHA and in the *many a noun* construction

Noun	Corpus frequency (COHA)	Construction frequency (*many a noun*)	Construction frequency/corpus frequency
time	640,580	1,114	0.17%
day	336,077	707	0.21%
man	607,742	703	0.12%

year	199,094	585	0.29%
hour	87,190	319	0.37%
night	220,685	242	0.11%
heart	157,527	166	0.11%
woman	178,428	161	0.09%
mile	18,422	155	0.84%
eye	72,262	91	0.13%

Table 2.

Noun frequencies in the *many a noun* construction by COHA-decades

Noun	1810s	1820s	1830s	1840s	1850s	...
time	2	43	50	78	74	...
day	1	9	52	58	66	...
man	5	14	39	38	61	...
year	3	23	30	54	57	...
hour	6	13	32	34	38	...
night	2	9	13	15	19	...
...	...	...	...	...	...	...

Table 3.

VNC periods of the *many a noun* construction

Periods	COHA size (in million words)	Tokens of *many a noun*
1810s	1.05	286
1820s–1860s	67.50	5,699
1870s–1900s	78.08	4,651
1910s–1940s	97.95	3,583
1950s–2000s	155.44	1,633

Table 4.

Partial input for a diachronic distinctive collexeme analysis of *many a noun*

Noun	1820s–1860s	1870s–1900s	1910s–1940s	1950s–2000s
time	354	430	242	86
day	247	281	144	34
man	190	237	192	79
year	208	218	100	56
hour	153	100	43	17
night	79	63	67	31
heart	106	42	12	6
woman	49	60	38	13

mile	68	60	22	4
eye	68	12	6	4
...	...	...	...	...

Table 5a.

Distinctive collexemes of *many a noun*, 1820s–1860s and 1870s–1900s

1820s–1860s					1870s–1900s			
Noun	Observed	Expected	Coll.Str		Noun	Observed	Expected	Coll.Str
eye	68	33	13.22		Time	430	256	9.62
heart	106	61	11.99		Day	281	163	7.91
tear	46	25	6.99		Year	218	134	4.27
league	35	18	6.46		Recitation	8	2	4.20
sigh	25	11	6.17		Door	9	3	3.25
spirit	20	9	5.44		Turn	14	5	2.97
hour	153	115	5.22		Frolic	5	1	2.62
scene	36	20	5.15		Trout	5	1	2.62
pang	33	18	4.91		Bit	10	3	2.45
form	19	9	4.15		Poet	9	3	2.41
tale	44	27	4.07		Struggle	13	5	2.16
bosom	11	4	3.90		Escapade	4	1	2.10
thought	27	15	3.87		Reverie	4	1	2.10
prayer	23	12	3.64		Scratch	4	1	2.10
gem	15	7	3.26		Wave	4	1	2.10

Table 5b.

Distinctive collexemes of *many a noun*, 1910s–1940s and 1950s–2000s

1910s–1940s					1950s–2000s			
Noun	Observed	Expected	Coll.Str		Noun	Observed	Expected	Coll.Str
citizen	29	9	10.34		Moon	9	1	5.73
reader	19	7	5.56		Investor	5	1	4.90
war	10	3	4.23		Businessman	7	1	4.16
Jew	6	1	3.83		Marriage	7	1	3.87
bigwig	7	2	3.66		weekend	3	0	2.94
scientist	8	2	3.65		corporation	4	1	2.82
state	11	4	3.23		politician	7	2	2.81
manufacturer	5	1	3.19		newspaper	6	1	2.72
Republican	8	3	3.18		deal	4	1	2.49
dollar	9	3	2.94		executive	4	1	2.49
city	17	9	2.75		world	4	1	2.49

game	9	3	2.64		dinner	5	1	2.47
observer	9	3	2.64		conservative	3	0	2.37
cowboy	4	1	2.55		river	4	1	2.22
educator	4	1	2.55		company	5	1	2.10