

Article

John Newman*

Revisiting N *waiting to happen*: word, construction, and corpus choices in a collocation analysis

<https://doi.org/10.1515/cllt-2024-0019>

Received February 22, 2024; accepted February 23, 2024; published online March 12, 2024

Abstract: In undertaking any collocation analysis, a researcher must make decisions concerning the properties of words, constructions, and corpora. Each of these crucial aspects of the analysis can be dealt with in alternative ways: words can be investigated as either lemmas or inflected forms; a construction can be characterized in alternative ways (reliance on semantics or syntax or some combination thereof, the span of the construction, etc.); the choice of corpus (or corpora) will be influenced by whether a researcher has an interest in different genres and varieties, whether the study is synchronic or diachronic, etc. I review various ways in which a researcher's decisions about words, constructions, and corpora are relevant to a corpus-based study of N *waiting to happen*, referencing throughout the collocation analysis of this construction by Stefanowitsch and Gries. The approach adopted here can be seen as supplementing Stefanowitsch and Gries' original collocation analysis. It illustrates how multifarious the results of a corpus-based study of constructions can be and serves as a reminder that no one corpus-based measure can possibly answer all the questions linguists might reasonably ask about a construction.

Keywords: collocation; corpus; inflection; lemma; genre

1 Introduction

The starting point for this chapter is Stefanowitsch and Gries' (2003) article, henceforth referred to as S&G (2003). As is well known, the authors' aim in that article was primarily to introduce *collocation analysis*, a corpus-based method of

***Corresponding author: John Newman**, School of Languages, Literatures, Cultures & Linguistics, Monash University, Menzies Building, 20 Chancellors Walk, 3800 Clayton, VIC, Australia, E-mail: john.newman@monash.edu

determining the degree of attraction or repulsion words (*collexemes*) have to a construction. As part of the explanation of the method, the authors offered a collostructional analysis of the noun (N) words that occur in an N *waiting to happen* construction of English. As the authors clearly state (S&G [2003: 219]), the primary aim of the analysis was to illustrate a method. In a sense, any relevant construction could have been used to illustrate the method. It would be unfair, then, to evaluate their analysis of the N *waiting to happen* construction as if it were being offered as a major linguistic study of the construction, which it clearly was not intended to be. Nevertheless, their analysis of the construction is of interest in its own right, not just as an illustration of a method, but as an original study of a construction that had hitherto not been the focus of attention in linguistics. The authors themselves point out what they consider to be interesting aspects of their results, remarking, for example, on the negative connotations attaching to most of the nouns in the construction.

My aim here is to revisit Stefanowitsch and Gries' analysis of the N *waiting to happen* construction, highlighting various ways in which the data might have been dealt with differently, leading to additional or somewhat different results. The discussion concerns three key aspects of their analysis: how the concept of "word" is to be understood in relation to the collexemes of the construction; how the construction should be characterized; and the choice of corpus underlying the study. Importantly, also, researchers can now avail themselves of a much larger corpus than was available in 2003 to investigate the construction and this will be helpful in confirming (or not) the findings from S&G (2003).

2 Collostructional analysis using COCA

One very obvious way to revisit the collostructional analysis of the N *waiting to happen* construction in S&G (2003) is to apply the method to the Corpus of Contemporary American English, abbreviated as COCA (Davies 2008–).¹ The version of COCA released in March 2020 contains more than one billion words of American English spanning the years 1990–2019. This version of COCA contains eight subcorpora: BLOG, WEB, TV-MOV (TV and movie subtitles), SPOK (spoken

¹ For an even larger corpus of English, one could make use of the Corpus of Global Web-Based English (GloWbE, Davies [2013]) containing 1.9 billion words of text from 20 different countries. For the present study I chose to focus on American English, and it was important for my purposes to use the largest (balanced) sample of American English texts available to me, which is COCA, containing more than 1 billion words. The American component of GloWbE consists of approximately 387 million words, about half of which are taken from blogs only, and as such represents a smaller and less balanced sample of American usage than COCA does.

subcorpus of transcripts of unscripted conversation from TV and radio programs), FIC (fiction), MAG (popular magazines), NEWS (newspaper text from ten newspapers in the USA), and ACAD (academic journals). COCA offers, therefore, the possibility of a much larger set of results concerning the construction compared with the British National Corpus (BNC) used in the S&G (2003) article. The BNC, it will be recalled, consists of 100 million words, containing just 35 instances of the construction. Applying the method to a larger data set, based on a corpus ten times the size of the BNC, provides, therefore, an opportunity to confirm or disconfirm trends noted in the BNC. In this section, I provide a collostructional analysis of the N *waiting to happen* construction, following closely the method employed by S&G (2003), but basing the analysis on COCA. I put aside for the time being questions about varieties, genres, and time periods reflected in the two corpora (see Section 5 for more on these matters).

As much as possible, the same procedures were followed that were used in S&G (2003). The nouns in N *waiting to happen* were identified manually from the context of instances of *waiting to happen* in the whole of COCA (an initial total of 735 hits).² Recall that the N in this construction is the noun that functions as a semantic head, modified by *waiting to happen*. Such heads could be part of the subject, e.g., *an accident is waiting to happen*, or not, e.g., *it's an accident waiting to happen*; they may be adjacent to *waiting* or not, e.g., *an accident just waiting to happen*; they may occur in the same sentence as *waiting to happen* or in an earlier sentence. A number of issues arise when trying to identify the head noun in the COCA data, issues which were not apparent when working with the BNC examples. In some cases, there is no convincing single noun that acts as the head of *waiting to happen*, as in (1a–b), which occur in the MOVIE subcorpus. In (1a), the topic of discourse is an imminent birth, but there is no obvious noun connected to *waiting to happen*. In (1b), it is difficult to say if *waiting to happen* refers to anything at all in the preceding discourse. The movie context of these scripts, or transcripts, presumably allows for a freer use of *waiting to happen*, with visual aspects of the scene and the general story line providing the necessary background information. Such cases were not included in the collostructional analysis.

² The initial total of hits is based on a count of the concordance lines retrieved through an online search of COCA with *waiting to happen* as the search term. There was a discrepancy between the frequency of hits as reported in the web interface (737 in January 2023) and the number of concordance lines actually returned (735). Repeating the online search in December 2023 revealed a similar discrepancy: a total frequency shown as 738 in the online interface but the same 735 actual concordance lines returned as before.

- (1) a. Breathe, breathe, breath, breathe. – No, John. John, I need to poop. No, no, it's the baby, okay? Don't poop. It's not poop, it's the baby. Oh, God, I need the drugs. I need drugs. I need something. – I can't do this. It hurts so much. – **Waiting to happen**. I need an epidural. Hey, wait. Can you get me an epidural? (COCA, MOVIE, 2015)³
- b. Listen, I grew up around these rednecks, and I think you ought to know who you're fucking with. And in case you got any interest, Gator eats breakfast nearly every day at Lions café just like me. Now, he don't look like much or nothing, but you remember what I told you, he ain't true. Crazy, **waiting to happen**. Hey, where are you? (COCA, MOVIE, 2013)

An interrogative pronoun can function as the head of *waiting to happen*, as in (2a). Since there is no noun acting as the head, such cases were also excluded from further analysis. The personal pronoun was used in a couple of instances like a noun, as in (2b), where *me* refers to a projected persona of the speaker, so these instances of personal pronouns were included in the analysis. In cases of conjunction or disjunction, as in (3a–b), the *waiting to happen* phrase refers to each of the nouns in the noun phrase rather than just one single noun. The decision was made to include these (relatively few) cases of multiple nouns as heads by counting each individual head word. After this manual editing of the initial hits, there were 735 head nouns (coincidentally, the same number of concordance lines identified before the manual editing) and these were the nouns used in the subsequent analysis.⁴

- (2) a. I shiver and sweat at the thought of what is **waiting to happen** tomorrow and where. (COCA, NEWS, 2008)
- b. You know, like, “Okay, I know I can commit suicide,” “because there’s a new me **waiting to happen**.” (COCA, MOVIE, 2011)
- (3) a. This event looks like another riot and disaster **waiting to happen**. (COCA, BLOG, 2012)
- b. her cooking is like a heart attack/diabetes/health problems just **waiting to happen** (COCA, WEB, 2012)

It is useful to remind ourselves of the results of the collostructional analysis in S&G (2003) and these are shown in Table 1. This table is essentially S&G's Table 5, p. 219,

³ For each example, I indicate the corpus (COCA), the subcorpus (BLOG, MOVIE, etc.), and the year of occurrence. Where appropriate, the head noun, or head noun phrase, associated with *waiting to happen* is underlined.

⁴ All the concordance lines, the noun phrases and nouns identified as being relevant to *waiting to happen*, the text type of each line, and the year of occurrence are available in the data package accompanying this article.

Table 1: Collostructional analysis for all lemmas in the N *waiting to happen* construction in S&G (2003), based on the BNC.

	Nouns	Freq	P Fisher Exact	Signif
1	accident	14	2.12e-34	*****
2	disaster	12	1.36e-33	*****
3	Welkom	1	4.46e-05	****
4	earthquake	1	0.00246	**
5	invasion	1	0.0071	**
6	recovery	1	0.0132	*
7	revolution	1	0.0168	*
8	crisis	1	0.0221	*
9	dream	1	0.0245	*
10	sex	1	0.0283	*
11	event	1	0.0692	ns

showing frequency and probability value according to the Fisher Exact test. I have added the statistical significance (“Signif”) indicated by asterisks thus: ***** = significant at $p < 0.00001$, **** = $p < 0.0001$, *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$. For the collostructional analysis based on COCA, the number of verb constructions required for the analysis was taken to be 100 million. Relying solely on the online interface to COCA, it was not possible to simply search for the number of verb forms in the whole of COCA. The figure of 100 million was a rough estimate and was arrived at by applying the same approximate proportion of (number of verb constructions)/(total number of words) as found in the BNC, namely 1/10, to COCA.^{5,6} Table 2 shows the results of the collostructional analysis based on COCA for all lemmas with frequency >4, amounting to a total of 457 (62 %) of the 735 inflected word forms occurring as collexemes.

As expected, the much higher frequencies found in COCA result in lower probability values. The main interest in such tables, however, lies in the ranking rather than the actual probability value, and it can be seen that the two words that rank highest in Table 1 (BNC), *accident* and *disaster*, also rank highest in Table 2 (COCA), with *accident* outranking *disaster* in both tables. In these ways, then, the analysis based on COCA provides additional support for the conclusions from S&G (2003).

5 The number of verb constructions need only be some realistic estimate for the purposes of collostructional analysis (cf. Gries [2015: 511]).

6 The collostructional analysis was carried out using the *collostructions* package v.0.2.0 (Flach [2021]) in R (R Core Team 2023). The R code used in the analysis is available in the data package accompanying this article.

Table 2: Collostructional analysis for the N *waiting to happen* construction, based on COCA (2020 version). Table includes all lemmas with frequency >4 in the construction.

	Nouns	Freq	P Fisher Exact	Signif
1	accident	159	0.00e+00	*****
2	disaster	129	4.75e-294	*****
3	wreck	22	3.17e-51	*****
4	tragedy	25	3.50e-45	*****
5	attack	28	1.77e-31	*****
6	lawsuit	16	5.50e-25	*****
7	catastrophe	10	2.93e-20	*****
8	nightmare	9	8.23e-14	*****
9	injury	8	2.55e-08	*****
10	crash	6	5.01e-08	*****
11	crime	6	8.10e-07	*****
12	scandal	5	1.10e-06	*****
13	incident	6	1.39e-06	*****
14	bomb	5	1.72e-05	****
15	crisis	5	0.000245	***
16	trouble	5	0.000561	***
17	problem	8	0.0351	*
18	case	5	0.101	ns

Furthermore, the negative connotations that the authors found in the BNC collexemes are also apparent in COCA. While Table 2 does not include all the collexemes, the words included in the table account for more than half of all the collexemes in that slot. And, arguably, every word in Table 2 has some degree of negative connotation. Note, though, that the negative connotation in either table is evident from inspecting the list of words, along with their frequencies. One doesn't have to depend on the statistical sophistication of their collostructional method to arrive at the conclusion that a negative semantic quality is present more often than not for either Tables 1 or 2. A list of all words with frequency above some cut-off value, accounting for the majority of instances (as in Table 2), is also sufficient in itself to arrive at some conclusion about negative connotations.⁷ In some cases, however, probability values may be a useful step in refining some finding about negative connotations (cf. the

7 The Fisher Exact measure of probability used in S&G (2003), and adopted here in the interest of following the same methodology, conflates (and masks) a number of different dimensions, including frequencies, as discussed in Gries (2019: 394–401). This means that ranking collostructional strength by the Fisher Exact test (and some other statistical tests such as the log-likelihood test) largely coincides with ranking by raw frequencies. The (log) odds test, on the other hand, is independent of the raw frequencies.

discussion of negative prosody of the *experience* N construction in Newman [2011: 542–546]). In the case of the BNC, Stefanowitsch and Gries point out that the words *dream*, *sex*, and *event* lack a negative connotation and one could mention in this regard that these three words are, in fact, the three lowest ranked collexemes in Table 1 relying on the Fisher Exact test rather than simply frequencies. So, the ranking of the words in Table 1 (specifically) by probability values turns out to have some correlation with the presence or absence of negative connotation in this case (although the authors themselves did not pursue this line of inquiry).

Although it is the collexemes in COCA with higher frequencies (>4) and greater statistical significance that we are most concerned with here, the lower frequency collexemes can also be brought into the discussion. It still appears to be the case that the lower-frequency head nouns, considered out of context, are mostly negative in their connotations (e.g., *trainwreck*, *war*, *riot*, *batshit*, *bottleneck*, *breakup*, *confusion*, *collision*, *cruelty*, *debt*, *depression*, *mess*), but nouns that would appear to be neutral or positive in connotation do occur. Neutral nouns include, arguably, *biscuit*, *calls*, *change*, *career*, *criterion*, and positive nouns might include *blessing* and *adventure*. The larger set of examples of the construction found in COCA enables a fuller appreciation of the minor pattern of positive connotations than is possible using the BNC. More discourse context is necessary to make more accurate judgments about the positivity or negativity of the uses of the head nouns (see the discussion of semantic prosody in Section 4), though pursuing this goal goes beyond the current study. The examples in (4) illustrate positive connotations associated with the use of the construction taken from COCA. One can see how a head noun, taken out of context, may seem quite neutral in connotation, whereas the larger context points to a positive connotation. Thus, *quilt* in (4b) is used as part of a metaphorical and uplifting way of talking about the variety of good experiences in life; *citizens* in (4f), as part of the phrase *good citizens*, points to a positive future.

- (4) a. A bare tabletop is nothing but a display opportunity waiting to happen. (COCA, MAG, 2010)
- b. In order to receive a response you can change your status in your blogger profile. LinkWithin Life is a Quilt Waiting to Happen # Montreal 2010 I See Quilts # Whether it be the birth of a new baby, the loss of a loved one, a spectacular sunset or the way the light hits the glass of a beautiful building ... life is a quilt waiting to happen. (COCA, BLOG, 2012)
- c. No question, Rebel Wilson is a big, funny personality, a TV star waiting to happen. (COCA, NEWS, 2013)
- d. Bateman uses the Force to convince yet another network executive that he's a star waiting to happen. (COCA, NEWS, 2003)

- e. And – and she really is. She’s like, you know, comedy **waiting to happen**. (COCA, SPOK, 1998)
- f. ... flirted disastrously with the idea that terrorist groups like Hizballah and Hamas are simply good citizens **waiting to happen**. (COCA, MAG, 2007)
- g. Packed like a bright graveyard with slabs of marble and men on horseback, Washington is a ceremony **waiting to happen**. (COCA, MAG, 1997)
- h. With “Teen Spirit,” Nirvana – minor celebrities in the Northwest underground – became superstar totems of a rock & roll America **waiting to happen**: Alternative Nation. (COCA, MAG, 2000)

3 The collexemes

The concept of “word” is open to different interpretations in so far as it applies to identifying the collexemes in a collostructional analysis. It should be pointed out that the ambiguity that attaches to the concept “word” is an issue of some importance within the whole field of corpus linguistics affecting any results pertaining to frequency lists, collocates, key words, etc. (cf. the discussion of inflected forms, lemmas, and counting “words” in Newman [2010]) and is not just an issue that arises when carrying out collostructional analysis.

In S&G (2003), the concordance lines included the example of *a business disaster waiting to happen* and the word *disaster* was taken to be the head noun of the *waiting to happen* construction. However, one might just as well have taken the compound noun *business disaster* as the head of the *waiting to happen* construction, rather than just the single orthographic word *disaster*, even if *disaster* functions as the head element of the noun compound. In the COCA data, one finds similar noun compounds consisting of two or more orthographic words such as *heart attack(s)* (24), *train wreck* (22), *antitrust case* (1), *murder plot* (1), *gender revolution* (1), *paper magazine fashion spread* (1), etc. Recognizing the whole noun compound does capture more of the noun semantics, while identifying just a single head noun of the compound leads to stronger generalizations involving fewer nouns. Either way seems reasonable with little effect on results in most cases. Note, though, the high frequencies of the compounds *heart attack(s)* and *train wreck* in the COCA data, which suggests that recognizing noun compounds as collexemes will impact the relative ordering of the collexemes by probability value. Identifying noun compounds that appear as distinct orthographic words may involve considerable additional effort by the researcher

unless the corpus is syntactically tagged for such compounds. Variation in orthographic conventions, such as *lawsuit*, *law-suit*, and *law suit*, only add to the problem.

A second point of interest about the collexemes is whether the word should be analyzed as an inflected word form or as a lemma. S&G (2003: 215) make clear that they are exploring the noun collexemes in the N *waiting to happen* construction at the lemma level and propose the lemma as the default level of investigation: “[...] collostructional analysis collapses these [= counts of singular and plural nouns] into one figure for each corresponding lemma unless there is reason to believe that the construction is associated with only one particular word form.” Thus, the word written as *accident* in Table 1 represents the lemma standing for both the inflected word forms *accident* and *accidents*. The same procedure was followed in reporting the findings of the analysis in Table 2. Including all the inflected word forms in the analysis, as Stefanowitsch and Gries did, has the advantage that one can further investigate the relative frequencies of the various inflected forms, though the authors chose not to do this.⁸ In the case of the N *waiting to happen* construction, this additional step in the analysis reveals large differences in the frequencies of the singular and plural forms: in the BNC data, 31 out of 35 of the N collexemes are singular (89 %); in the COCA data, 672 out of 735 collexemes are singular (91 %).

Clearly, there are more singular nouns than plural nouns functioning as collexemes in both analyses. The question arises, however, whether the singular nouns are significantly more frequent in this construction than in other constructions, given that singular nouns are more common overall in English (Biber et al. [1999: 291–292]). Singular nouns function as the unmarked form, being used, for example, as a noun modifier in compounds such as *accident investigation(s)*, *accident prone*, etc. In such cases, the number difference is neutralized and the plural form is unacceptable (**accidents investigation*, **accidents prone*). To test whether the singular/plural ratio is indeed higher in the N *waiting to happen* construction than elsewhere, a comparison was made with the singular/plural ratio in noun + verb sequences. It was important to ensure as far as is realistically possible that the base ratio being used for comparison excluded the many cases where the noun in question occurs as the modifier of a head noun to the right, since in these compounds the modifier only occurs in the singular form. There are many such compounds in the whole of COCA, e.g., *disaster relief* (1,020), *disaster recovery* (343), *disaster area* (309), *disaster [medical response]* (2), *accident victims* (145), *accident scene* (143), *accident investigation* (142), *lawsuit abuse* (48), *lawsuit settlement* (29), and many more. In addition, there are many compound adjectives with the target noun as the left modifier of a head: *accident prone* (55), *accident*

⁸ Although Stefanowitsch and Gries did not investigate inflectional differences in the head noun of the N *waiting to happen* construction, they did choose progressive, imperative, and past tense (inflected) categories of the English verb as case studies to further illustrate collostructional analysis elsewhere in the same article.

free (9), *disaster prone* (11), etc. Since our interest is in the choice of singular versus plural, it is desirable to exclude these cases where no such choice is available. Hence, the decision to limit the search for singular versus plural nouns to the context of a following verb form, e.g., *accident was/is/should/will/happen/occur* etc. Some of these sequences involve a head noun of the subject phrase (*the accident was horrible*) while others do not (*the cause of the accident is unknown*). The key requirements are that the search terms include many cases of the target noun (in the hundreds or thousands) in contexts where the number difference is not neutralized to the singular and a target noun followed by a verb as search term will accomplish this. I, therefore, compared the ratio of singular/plural in the N *waiting to happen* construction in COCA for the six most frequent lemmas with the ratio of singular/plural in N V sequences in COCA (excluding the N *waiting to happen* construction).⁹ Results are shown in Table 3. The difference in ratios (expressed as the relative proportion of the singular/plural ratios in the two constructions) together with the Fisher Exact probabilities show that the singular form is significantly more frequent than the plural form in the N *waiting to happen* construction, compared with other construction in which a noun is followed by a verb. This finding lends statistical support to the conclusion that the N *waiting to happen* construction concerns the expectation of a *single* event rather than multiple events.¹⁰

Table 3: SG/PL ratios in the N *waiting to happen* construction (N*waiting* Cx) and in the N V construction (N*verb* Cx, excluding N *waiting to happen*) in COCA for the six most frequent lemmas. Zero frequencies of occurrence of the plural (*wrecks* and *lawsuits*) were assigned the value of 1 as the denominator in the calculation of ratios.

SG/PL forms	N <i>waiting</i> Cx	N <i>verb</i> Cx	N <i>waiting</i> Cx/N <i>verb</i> Cx	P Fisher Exact	Signif.
accident/accidents	143/16 (8.9)	3,636/2,168 (1.7)	5.2	1.81e-14	*****
disaster/disasters	119/10 (11.9)	3,091/1,150 (2.7)	4.4	7.66e-08	*****
tragedy/tragedies	21/4 (5.3)	3,035/2,924 (1)	5.3	0.000947	***
attack/attacks	25/3 (8.3)	8,972/7,018 (1.3)	6.4	0.000210	***
wreck/wrecks	22/1 (22)	454/137 (3.3)	6.7	0.006853	**
lawsuit/lawsuits	16/1 (16)	5,407/2,370 (2.3)	7	0.004759	**

⁹ In COCA, the search terms are “accident VERB+,” “accidents VERB+,” etc.

¹⁰ A reviewer has pointed out that one could carry out a collostructional analysis based on inflected word forms rather than lemmas in order to establish the preference for singular forms among the collexemes and their statistical significance. Estimating the frequencies of relevant singular word forms in the whole corpus (limiting the scope of singular forms to sites of potential singular ~ plural alternation) remains a challenge, but even a collostructional analysis based on automatically assigned POS-tags for singular and plural nouns will be instructive.

4 The construction

The tendency for nouns with negative connotations to occur as the collexeme in the N *waiting to happen* construction is an intriguing result of S&G's analysis. The study of positive or negative nuances, or connotations, accompanying a word or phrase is commonly referred to now as *semantic prosody* (Begagić [2018]; Stewart [2010]). One may ask, then, how well a study of the specific construction N *waiting to happen* captures the relevant facts about the semantic prosody accompanying *waiting to happen*.

While the decision to focus attention on a single noun slot in a construction may be procedurally convenient, allowing the researcher to collect and evaluate many examples of the same part of speech functioning as the head of a construction, there is a risk that in doing so the researcher could overlook the presence of (positive or negative) connotations adhering to words other than the head noun. A semantic nuance may be expressed, after all, through a variety of words in context, a good illustration of this being Sinclair's analysis of the semantic prosody of the word *regime* (Sinclair [2003: 17–21]). On the basis of inspecting concordance lines with *regime* as the key word, Sinclair argued that this word attracted a negative prosody, suggesting unpopular governments or dictators, the violent use of power, etc. The specific words that formed the basis of this finding about *regime* occurred in a variety of parts of speech and syntactic roles: attributive adjectives (*appalling regime*), predicative adjectives (*the regime, which was not discriminatory*), verbs with *regime* as subject (*the regime has angered companies*), verbs with *regime* as object (*to blast the regime publicly*), head noun in a prepositional phrase (*condemnation of the military regime*), etc. The same kinds of considerations apply to the occurrences of *waiting to happen* in COCA. Some head nouns of *waiting to happen* seem neutral in connotation when examined alone, but accompanying adjectives may reveal negative connotations as in the examples in (5a) or positive connotations as in (5b).

- (5) a. the negative effects, a bad marriage, a horrible signing, a bad biscuit, a blood sugar spike, a failed policy, a red card
 b. a positive domino effect, a perfect miniseries, good citizens (as in (4f) above)

The presence of adverbials modifying the attributive adjectives can also create a stronger connotation, e.g., *a very serious injury*, *a preposterously expensive accident*. A similar accentuation of the negativity is achieved through a modifying phrase after the head noun in *a Hindenburg disaster times a hundred*. The use of *just* as an adverbial modifier in *just waiting to happen* (39 instances in COCA, two in BNC) could also be seen as accentuating the inevitability or imminence of the negatively viewed

event referred to by the head noun. A semantic prosody may only be evident from the larger, nonspecific context, as in (6), where the noun *Greece* is being viewed negatively in (6a) and where *storybook* is viewed as something positive in (6b).

- (6) a. you all will be living as the earthquake victims in Haiti and the people of india who live in slums and you will only have yourself to thank # Goodforall # Unfortunately for people who think like we do, the sheep can't see the forest for the trees – we are now a Greece waiting to happen. # SATCitizen # So let the hand outs begin. (COCA, BLOG, 2012)
- b. This horse is so cool. The horse has one eye. And it's very rare, extremely rare that you get a thoroughbred running in these kind of races. SHEINELLE-JONES# With one eye. MIKE-TIRICO# But he's still able to do it and he actually has a chance to win this race because he's bred from horses that have run well at a long distance. SHEINELLE-JONES# It's like a children's storybook waiting to happen. DYLAN-DREYER# Yeah. MIKE-TIRICO# And kids around the country fell in love with the story of Patch sending notes to the barn and all that stuff. (COCA, SPOKEN, 2017)

Given that semantic prosody associated with *waiting to happen* can manifest itself through such a variety of words belonging to different POS classes, functioning in a variety of syntactic roles and grammatical relations within and across sentences, it seems that any attempt to restrict attention to just one part of speech in one slot, like the head noun of *waiting to happen* is likely to be inadequate in some ways. On the other hand, one must recognize the challenges of investigating semantic prosody, especially the need to inspect some indeterminate number of words before and after *waiting to happen*, within the sentence and beyond. These challenges are daunting when undertaken on a large scale, and some simplification of methodology is surely justified. Focusing on the head noun in the N *waiting to happen* allows the researcher to capture the most pertinent data relevant to semantic prosody in a manner that is efficient and easily manageable (even when working with a one billion-word corpus). On balance, then, the N *waiting to happen* construction would seem to be an astute choice of construction, offering a revealing and at the same time a very practical choice of constructional elements to investigate semantic prosody.

It is worth noting that *happen* has its own prosody. An early reference to a negative semantic prosody associated with *happen* can be found in Sinclair (1991: 112), where the author observes: “Many uses of words and phrases show a tendency to occur in a certain semantic environment. For example, *happen* is associated with unpleasant things – accidents and the like.” Further references to the semantic prosody of *happen* are made by Partington (1998: 67), Bublitz (1996: 17), Stewart (2010: 9), and Begagić (2018: 70–71). Extended discussion can be found in Sinclair (2003: 117–125) and Partington (2004: 136–141). Sinclair (2003: 124) concludes that “the main

orientation of *happen* is the prospection of an unfortunate event happening; this often goes with expressions of doubt or vagueness. Occasionally, the word presages the opposite – a desirable event – and in such cases there are often expressions of certainty along with it.” It is obvious that the semantic prosody associated with *happen* overlaps with that of *waiting to happen*. One key point of difference, however, is that *happen* can occur with a “happen-by-chance” meaning that is absent from the *waiting to happen* construction. This meaning is found with *happen to* + infinitive verb, as in *I happen to have read that book*.¹¹

5 The corpus

It can be instructive to examine the distribution of a construction in the various text types that may be included in large, general corpora like the BNC and COCA.¹² To simply illustrate a method, as was done in S&G (2003), all that is necessary is *some* corpus, but if the goal of the researcher is to investigate the use of a construction, then the use of the construction in different varieties or different text types within a variety is directly relevant. Figure 1 shows the frequency per million words of *N waiting to happen* across the text types, based on the 735 instances of the construction identified earlier.¹³ It can be seen that TV-MOV, where entertainment value of content is high, shows the highest relative frequency, though confidence intervals show that there is no significant difference with a number of other categories. ACAD, where content is generally more sober and restrained, shows the lowest relative frequency. FIC and WEB occupy a middle ground, showing lower frequency of usage than TV-MOV (and mostly nonoverlapping confidence intervals), but higher frequency than ACAD (again, with non-overlapping confidence intervals).¹⁴ Some

11 To the extent that there is overlapping semantic prosody between *happen* and *waiting to happen*, one can view the prosody of the latter as being *motivated* by the former alongside the default *inheritance* of a bias toward negative prosody through the overriding abstract schema associated with the *waiting to happen* construction (cf. Booij’s discussion of motivation and inheritance in *Constructional Morphology* in Booij [2017]).

12 As mentioned in Footnote 1, GloWbE includes texts from 20 different varieties of English, enabling comparisons of the *N waiting to happen* construction across those varieties. The construction occurs in all varieties represented in the corpus, but a detailed study of its use in all these varieties is beyond the scope of the present study.

13 The 95 % confidence intervals included in the analysis were obtained by applying a proportion test.

14 As suggested by a reviewer, it may be that the *N waiting to happen* construction is regarded as too cliché for more frequent use in imaginative fiction, even if the head noun can be imaginatively chosen, e.g., *just another evil ex waiting to happen*, *a bad biscuit waiting to happen*, etc.

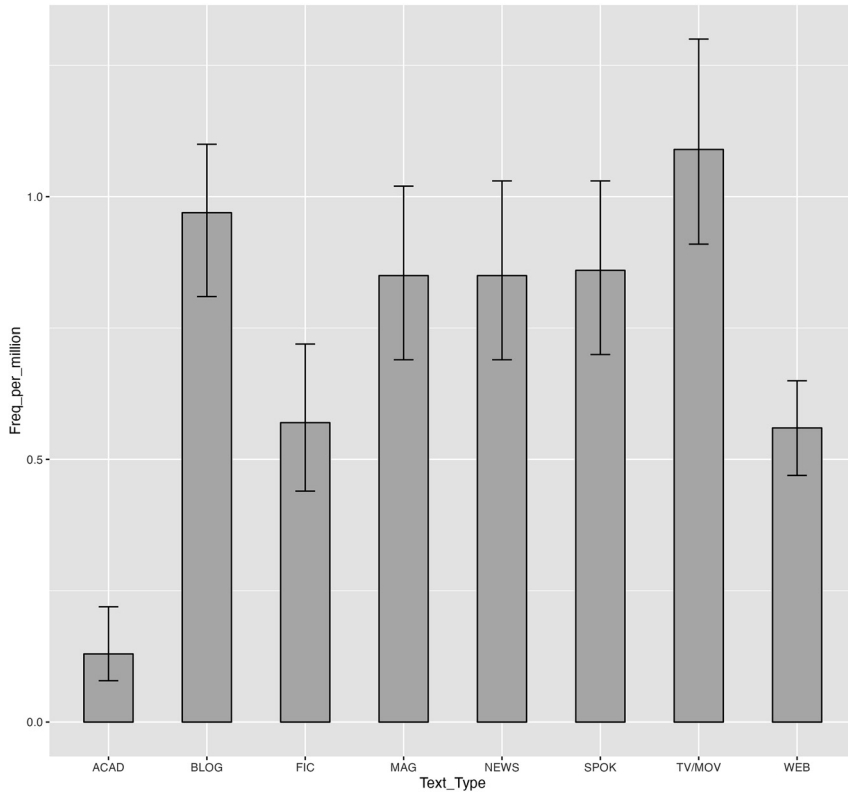


Figure 1: Relative frequencies of *N waiting to happen* in eight text types of COCA.

examples of the construction from TV and MOVIE dialogues, shown in (7), indicate a kind of playfulness with the construction. In all three examples, a speaker uses the construction to highlight some character attribute of the addressee. In (7a), the head noun phrase *another evil ex* clearly points to a negative appraisal of the addressee, as one might expect from the discussion of semantic prosody above. (7b) and (7c), on the other hand, contain seemingly innocent head noun phrases, *a gym teacher* and *a grad student*. Because of the overwhelmingly negative semantic prosody associated with the construction, however, the use of *waiting to happen* here has an overlay of a negative connotation imposed on it, resulting in a humorous, lightly teasing effect. In (7b), the implication is that a gym teacher is not a physically attractive person, the context being a beauty competition as an episode in a comedy show. In (7c), the instructor confirms that the student is likely to become a grad student, which the student clearly sees as something undesirable, saying “I hope not.”

- (7) a. I don't enjoy all this, Scott. In fact, I'm sick of it. I thought you might be more understanding. I just ... You're just another evil ex **waiting to happen**. (COCA, MOVIE, 2010)
- b. This competition's gonna be a piece of cake, Rose. Just look around. Ugly. No charisma. Now there's a gym teacher **waiting to happen**. Blanche, how can you say that? All these girls are adorable. (COCA, TV, 1991)
- c. Do you need an extension?
No, actually I wanted to say thanks. This has been a really cool class. I can't wait to get to more complicated stuff next term.
Really? You think this stuff is cool?
Yeah, but I'm kind of a dork.
You're like a grad student **waiting to happen**.
I hope not, but thanks again. (COCA, MOVIE, 2011)

The COCA data also enables us to compare the usage of *waiting to happen* across the time span of 1990–2019 in 5-year intervals, shown here as Figure 2. It should be noted that the BLOG and WEB categories are excluded from the data for calculating the

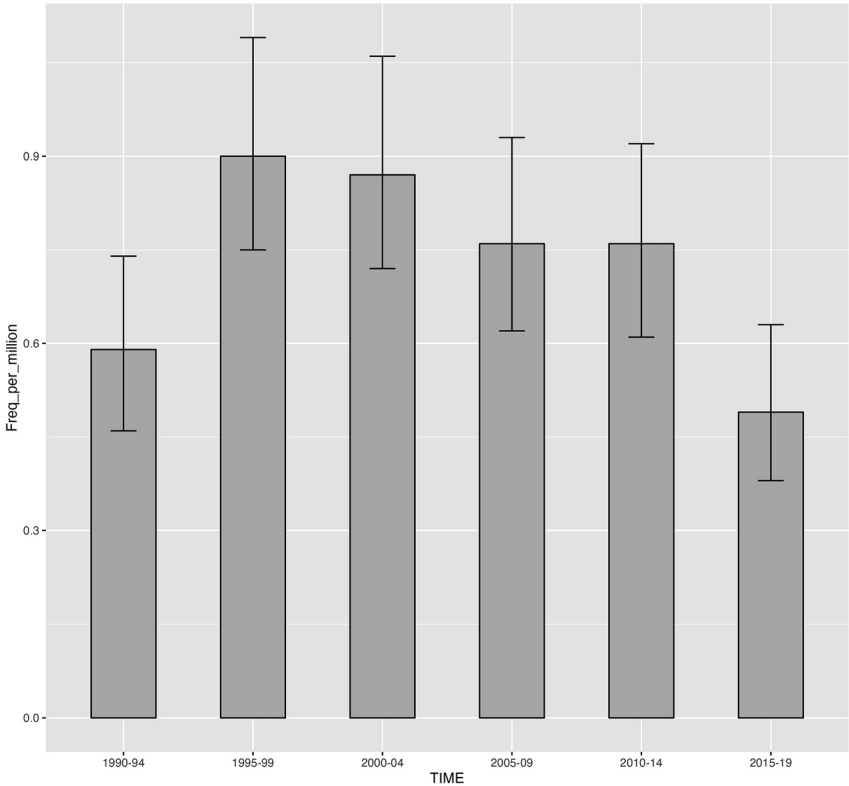


Figure 2: Relative frequencies of N *waiting to happen* in the period 1990–2019 in COCA.

values in Figure 2, resulting in a total of 540 instances for analysis rather than 735. As explained by the creator of the corpus, Mark Davies, in the online information pages of the corpus, the texts from these two categories were all added in October 2012 and they are excluded here to enable the data to be more comparable across all time periods. Figure 2 shows a rise in usage from 0.6 words per million in 1990–4 to 9.0 in 1995–99, with nonoverlapping confidence intervals associated with these two time periods. Usage remains above 0.6 up to and including the 2010–14 period, after which usage drops back to a similar rate as in the 1990–94 period.

One can also consider the productivity of the construction over the time span 1990–2019. One measure we may use involves type/token ratios, with greater productivity reflected through a higher ratio of the head noun types to tokens. Relevant information on type versus token frequencies of the head nouns in *N waiting to happen* in COCA is summarized in Table 4. Singular and plural forms are treated as distinct tokens in the counts in this table. There is little fluctuation in the 2000–2014 period, but the most recent five-year period of 2015–19 has a significantly higher type/token ratio (0.68, 95 % confidence interval is 0.55–0.79) than is found for the 1995–99 period (0.42, 95 % confidence interval is 0.33–0.52), pointing to an increased productivity of the *N waiting to happen* construction even while the most recent token frequency of the construction is less compared to the 1995–99 period. Another measure of productivity is the proportion of single-occurrence hapaxes relative to tokens, also shown in Table 4. The most recent period 2015–2019 shows a hapax/token ratio of 0.6 (95 % confidence interval is 0.47–0.72), which is higher than in the previous time periods and significantly higher than in the 1995–99 period (0.30, 95 % confidence interval is 0.22–0.40). Note also the most highly ranked (ranks 1–3) head nouns in each of the five-year periods, as shown in Table 5. It can be seen that singular *accident* and *disaster* consistently occupy first and second rank positions, respectively. In the most productive time period of 2015–19, these two head tokens accounted for just 25 % (15/60) of the head nouns, whereas in the period of lowest productivity 1995–99, they represent 44 % (50/113). In the 2015–19 period, then, the use of the *N waiting to happen* construction has become less frequent overall in COCA, while the set of nouns functioning as the head of the construction has become more diverse.

Table 4: Counts of head nouns in *N waiting to happen* in COCA 1990–2019.

	1990–94	1995–99	2000–04	2005–09	2010–14	2015–19
tokens	71	113	109	94	93	60
types	37	48	56	54	50	41
hapaxes	31	34	46	44	41	36
type/token	0.52	0.42	0.51	0.57	0.54	0.68
hapax/token	0.44	0.30	0.42	0.47	0.44	0.6

Table 5: Highest frequency tokens of head nouns in *N waiting to happen* in COCA 1990–2019. Frequencies are indicated in parentheses.

Freq. rank of N	1990–94	1995–99	2000–04	2005–09	2010–14	2015–19
1	accident (15)	accident (27)	accident (30)	accident (29)	accident (16)	accident (8)
2	disaster (14)	disaster (23)	disaster (12)	disaster (7)	disaster (11)	disaster (7)
3	disasters, accidents, tragedy (4)	attack (4)	wreck (5)	attack (6)	wreck (6)	attack (4)

One can turn to other corpora to explore the usage of *waiting to happen* over a longer time-frame than is possible with COCA. I turned to Google Books and the Google Ngram Viewer, also referred to as Google Books Ngram Viewer.¹⁵ I was unable to find an exact figure for the current size of the English language part of the Google Books corpus used by Ngram Viewer from the website, but at the time of its initial release in 2010 (before the addition of the 2011–2019 texts), the corpus contained 361 billion words of English (Michel et al. [2011: 176]). The information pages of Google Books describe the corpus as based on over 10 million books, but this includes languages other than English.¹⁶ The Ngram Viewer allows a user to inspect the use of words and phrases over spans of years, in American English or British English, and English fiction. There can occasionally be problems with the metadata associated with Google Books which a researcher needs to bear in mind (see Younes and Reips [2019: 2]). So, for example, a wildly wrong date of publication can sometimes be associated with an occurrence of a phrase (a 2011 publication containing the phrase *waiting to happen* was erroneously assigned a date of 1901, for example). While claims relating to any particular occurrence of *waiting to happen* in Google Books thus need to be double-checked for accuracy, the Ngram Viewer does provide useful snapshots of overall trends. Figure 3 shows the relative frequencies of selected expressions in the American English (2019) and the British English (2019) subcorpora of Google Books, accessed in November 2023, using the Ngram Viewer. The relative frequencies are computed by dividing the number of instances of the n-gram in a given year by the total number of words in the corpus in that year, expressed as a percentage (Michel et al. [2011: 176]). In the graph of the American corpus, Figure 3a, it

15 Google Books Ngram Viewer is accessible at <https://books.google.com/ngrams>.

16 The reference to more than 10 million books can be found at https://www.google.com/intl/en/googlebooks/about/free_books.html.

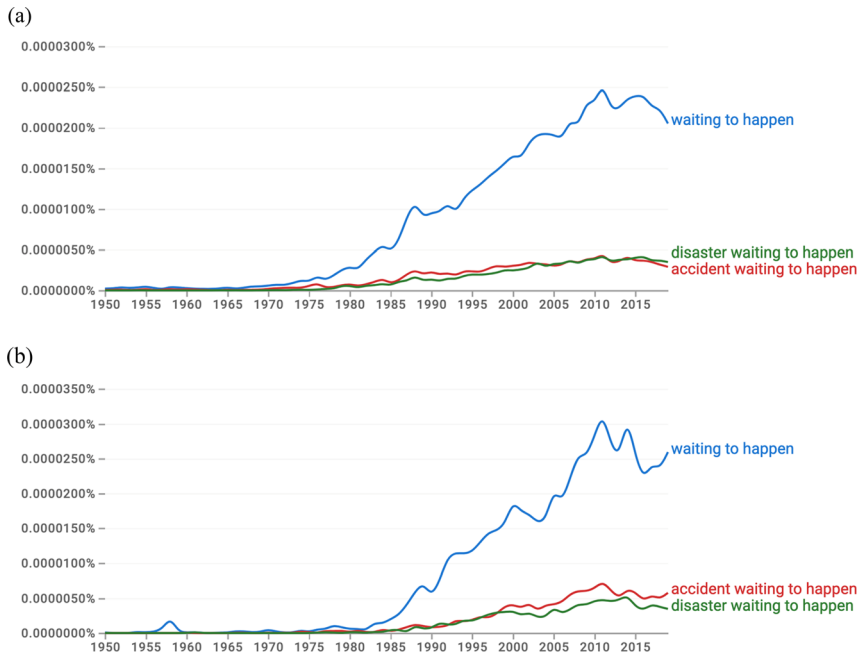


Figure 3: Screenshot from Google Books Ngram Viewer showing plots of relative frequencies of *waiting to happen* and its use with *accident* and *disaster* 1950–2019 in (a) American English and (b) British English.

can be seen that *waiting to happen* rose steadily in frequency of usage, beginning in the 1970s, peaking in 2010–11 (the precise year is obtained through interaction with the graph online), and dropping off a little since 2013. The graph of British English in Figure 3b shows a similar overall trend but indicates a later beginning to the increase, in the 1980s rather than the 1970s. The graphs also show the presence of the forms *disaster* and *accident* as relevant head nouns throughout this period, though there is increasing use of the construction overall. The Ngram Viewer results and the COCA results both point to a rise in the usage in the late 20th century/early 21st century, with a slight fall since the peak. The steady rise in American usage based on Ngram Viewer (and Google Books) is not evident in the time chart based on COCA, Figure 2 above, a reflection of the quite different kinds of corpora underlying the two graphs.

Using a web interface to Google Books to explore the earliest occurrences of *waiting to happen*, and double-checking dates of publication in light of the remarks about potential dating problems mentioned above, I was able to find the two

American examples in (8), dated 1929 and 1930, respectively. These were the earliest examples I was able to find in Google Books.¹⁷ While *accidents* occurs as the head noun in the 1929 example, supporting the idea that *accident(s)* plays a key role in the early usage of *waiting to happen*, it should be noted that even in this early period other nouns are possible, such as the compound *first degree murder* in (8b).

- (8) a. There is not much that we can do now about any of the accidents referred to in the foregoing. They have already happened. But there are a great many more accidents **waiting to happen** in the months to come...
(Pacific Telephone Magazine, 1929, Vol. 23, p. 15)
- b. Defendant was not entitled to a new trial because of Commonwealth's summation remarks that "every armed robbery is a first degree murder **waiting to happen**..." (Purdon's Pennsylvania Statutes, Annotated, 1930, p. 174)

6 Conclusion

The main findings from the discussion above are summarized in (i)–(iv):

- (i) Relying on the current one billion-word corpus of COCA, it is possible to arrive at more robust results from a collocational analysis of N *waiting to happen* than was the case in S&G (2003). The results from COCA confirm the negative connotations of the head nouns that had been remarked upon in S&G (2003), with *accident* and *disaster* being the collexemes most attracted to *waiting to happen* in both corpora.
- (ii) It is possible to gain more insight into the collexemes by exploring the relative frequencies of singular versus plural forms. Singular nouns are significantly more frequent than plural nouns as collexemes of *waiting to happen*.
- (iii) The combination of the specific inflected forms *waiting* and *happen* in the construction N *waiting to happen* means that the negative connotation associated with the head noun is particularly strong. Without the *waiting* element, *happen* allows for a "happen by chance" meaning, which does not have the preference for the negative connotation of *waiting to happen*. The construction N *waiting to happen* was, therefore, an astute choice of construction to study in the sense that it brought to light the negative connotations of the construction in a very direct and elegant manner.

¹⁷ The earliest example of *waiting to happen* in The Corpus of Historical American English (COHA, Davies [2010]), covering the period 1820s–2010s, is dated 1977, with *accident* as the head noun. The earliest example from the TIME Magazine Corpus (Davies [2007]), spanning the period 1923–2006, is dated 1976, also with *accident* as the head noun.

- (iv) Relying on COCA, it can be seen that the N *waiting to happen* construction is most frequent in the popular culture of TV, movies, and blogs and least frequent in academic journals. The construction rose in popularity in the late 20th century and early 21st century, with signs that its popularity has now passed its peak, while at the same time the construction has become more productive.

Given that the analysis of the N *waiting to happen* construction in S&G (2003) was never intended to be a full linguistic study of the construction, it is worth stressing how effective Stefanowitsch and Gries' analysis was in demonstrating the attraction of nouns with negative connotation to a construction that had not previously been subjected to such scrutiny. It was precisely because I found their analysis of this particular construction so compelling when I first read S&G (2003) that I have chosen to take a closer look at their analysis twenty years later.

As effective as their analysis of the construction was, it was necessarily influenced and constrained by particular decisions made in the course of their analysis, e.g., whether to focus attention on inflectional word forms or lemmas, how to characterize exactly the construction of interest, whether to take a synchronic or diachronic perspective, and whether to consider variation between text types. These are choices that have a bearing on any corpus-based analysis of constructions in English. My purpose here has been to show that there is much more to be learned about the N *waiting to happen* construction than was presented in S&G (2003), reflecting different choices that can be made by the researcher at various points in the procedure followed. This should not be surprising or shocking but is nevertheless worth pointing out. It echoes the point made by Gries (2019) who argues for a multifaceted approach to understanding the attraction of words to a construction, extending beyond reliance on any one statistical measure such as probability values in a collocation analysis. Gries was concerned specifically with quantificational measures such as frequency, effect size, association measures, etc. and the desirability of recognizing that all of these are potentially relevant in the study of words in constructions. I have tried to show the value of having a similar openness in how a corpus linguist might approach a study of words in constructions. In either case, it is unrealistic to expect that a full understanding of the role of a word in a construction will be achieved by one single method.

Acknowledgments: I am very grateful to two anonymous reviewers for their extremely helpful feedback on an earlier draft of this article.

Data: The original concordance lines taken from the BNC and COCA, text type and year of publication, the full lists of head noun tokens and lemmas, and the R code used to carry out the statistical analyses are all available at https://osf.io/vcs3q/?view_only=a90c456694cf4c849cee30ce4182a253.

References

- Begagić, Mirna. 2018. Semantic preference and semantic prosody: A theoretical overview. *Journal of Education and Humanities* 1(2). 65–88.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow, Essex: Pearson Education Limited.
- Booij, Geert. 2017. Inheritance and motivation in Construction Morphology. In Nikolas Gisborne & Andrew Hippisley (eds.), *Defaults in morphological theory*, 18–39. Oxford: Oxford University Press.
- Bublitz, Wolfram. 1996. Semantic prosody and cohesive company: Somewhat predictable. *Leuvense Bijdragen: Tijdschrift voor Germaanse Filologie* 85(1–2). 1–32.
- Davies, Mark. 2007. TIME Magazine Corpus. Available at: <https://www.english-corpora.org/time/>.
- Davies, Mark. 2008–. The Corpus of Contemporary American English (COCA): 1 billion words, 1990–present. Available at: <https://www.english-corpora.org/coca/>.
- Davies, Mark. 2010. The Corpus of Historical American English (COHA). Available at: <https://www.english-corpora.org/coha/>.
- Davies, Mark. 2013. Corpus of Global Web-Based English. Available at: <https://www.english-corpora.org/glowbe/>.
- Flach, Susanne. 2021. *Collostructions: An R implementation for the family of collostructional methods*. Package version v.0.2.0. Available at: <https://sfla.ch/collostructions/>.
- Gries, Stefan Th. 2015. More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3). 505–536.
- Gries, Stefan Th. 2019. 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics* 24(3). 385–412.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Peter Norvig, Jon Orwant, Steven Pinker & Martin A. Nowak. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014). 176–182.
- Newman, John. 2010. Balancing acts: Empirical pursuits in cognitive linguistics. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics*, 79–100. Berlin and New York: Mouton de Gruyter.
- Newman, John. 2011. Corpora and cognitive linguistics. *Brazilian Journal of Applied Linguistics* 11(2). 521–559.
- Partington, Alan. 1998. *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam & Philadelphia: John Benjamins.
- Partington, Alan. 2004. Utterly content in each other's company: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1). 131–156.
- R Core Team. 2023. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.
- Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, John. 2003. *Reading concordances*. Harlow: Pearson Education.
- Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.
- Stewart, Dominic. 2010. *Semantic prosody: A critical evaluation*. New York & London: Routledge.
- Younes, Nadja & Ulf-Dietrich Reips. 2019. Guideline for improving the reliability of Google Ngram studies: Evidence from religious terms. *PLoS One* 14(3). e0213554.