International Journal of Multidisciplinary Research and Growth Evaluation.

On fill in and fill out: A corpora-based analysis

Namkil Kang
Far East University, South Korea

* Corresponding Author: **Namkil Kang**

Article Info

ISSN (online): 2582-7138
Volume: 03
Issue: 03
May-June 2022
Received: 05-04-2022;
Accepted: 20-04-2022
Page No: 108-112

Abstract
The main purpose of this paper is to demonstrate that *fill in* and *fill out* are low similarity synonyms. With respect to the Movie Corpus, it is worth noting that *fill in* was always preferable to *fill out* in the movies of six countries from the 1930s to the 1970s, whereas *fill out* was preferable to *fill in* from the 1980s to the 2010s. With respect to the COCA, it is significant to note that there is a high degree of similarity between *fill in* and *fill out* in the blog genre, whereas there is no similarity between them in the other genres (7 genres). Simply put, *fill in* is 12.5% the same as *fill out*. A further point to note is that *fill in* is the nearest to *fill out* in the blog genre. Quite interestingly, the COCA clearly shows that *fill in gaps* and *fill out forms* are the most preferable ones (61 tokens vs. 96 tokens) for Americans. When it comes to the collocations of *fill in* and *fill out*, 1.81% of fifty five nouns are the collocation of both *fill in* and *fill out*. This in turn suggests that *fill in* and *fill out* are low similarity synonyms.

Keywords: MC, COCA, token, fill in, fill out

1. Introduction

As Murphy (2016, 2019) points out, *fill in* and *fill out* are used interchangeably. The main goal of this paper is to compare *fill in* with *fill out* in the Movie Corpus and the Corpus of Contemporary American English. First, we consider the diachronic aspects of *fill in* and *fill out* in the Movie Corpus. More specifically, we compare *fill in* and *fill out* from the 1930s to the 2010s. Second, we consider the genre frequency of *fill in* and *fill out* in the COCA. To be more specific, by examining eight genres, we observe the similarity between *fill in* and *fill out*. In addition, we measure the distance between *fill in* and *fill out* in the eight genres of the COCA. Finally, we observe the collocations of *fill in* and *fill out* in the COCA. By comparing the collocations of *fill in* and *fill out*, we observe how much alike they are. The organization of this paper is as follows. In section 2, we argue that the film writers of six countries preferred using *fill out* rather than using *fill in*. We further argue that *fill in* was always preferable to *fill out* in the movies of six countries from the 1930s to the 1970s, whereas *fill out* was preferable to *fill in* from the 1980s to the 2010s. In section 3, we show that there is a high degree of similarity between *fill in* and *fill out* in the blog genre, whereas there is no similarity between them in the other genres (7 genres). This in turn implies that *fill in* is 12.5% the same as *fill out*. We also show that *fill in* is the nearest to *fill out* in the blog genre. In section 4, we contend that *fill in gaps* and *fill out forms* are the most preferable ones (61 tokens vs. 96 tokens) among Americans. We have also maintain that 1.81% of fifty five nouns are the collocation of both *fill in* and *fill out*. This in turn indicates that *fill in* and *fill out* are low similarity synonyms.

2. Fill in and Fill out in the Movie Corpus

In the following, we aim to consider the diachronic aspects of *fill in* and *fill out* from the 1930s to the 2010s. Table 1 shows the use of *fill in* and *fill out* from the 1930s to the 2010s:

An important question is “Which type was the preferable one among the movie writers of six countries?” Table 1 clearly shows that *fill out* (783 tokens) was slightly preferable to *fill in* (782 tokens) in the movies of six countries. More specifically, the overall frequency of *fill in* is 782 tokens, whereas that of *fill out* is 783 tokens.

This in turn suggests that the movie writers of six countries preferred using *fill out* rather than using *fill in*. The following graph shows the diachronic use of *fill in* and *fill out* from the 1930s to the 2010s:

Table 1: Frequency of fill in and fill out in the Movie Corpus

Period	Fill in	Fill out
1930s	12	7
1940s	41	14
1950s	55	29
1960s	50	26
1970s	49	46
1980s	55	56
1990s	101	118
2000s	194	211
2010s	225	276
All	782	783
US/CA	581	703
UK/IE	156	37
AU/NZ	16	10
Misc	29	33

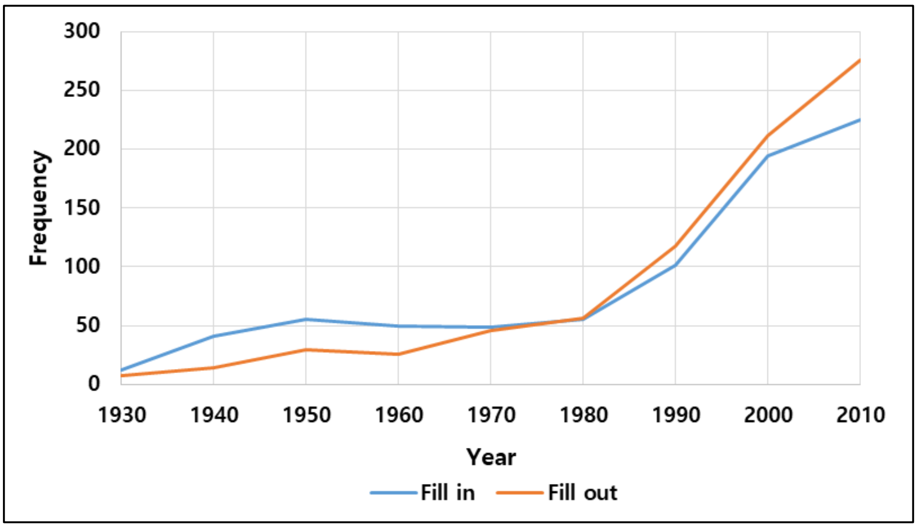

Fig 1: Frequency of fill in and fill out from the 1930s to the 2010s

It is worth noting that there was a steady increase (a rise of 43 tokens) in the frequency of *fill in* from the 1930s to the 1950s. More interestingly, there was a sudden decline (a fall of 5 tokens) in the frequency of *fill* in in 1960s. Similarly, the frequency of *fill* in continued to decrease to 1 token in the 1970s. It is interesting to point out that there was a gradual rise (an increase of 276 tokens) in the frequency of *fill in* from the 1970s to the 2010s. It is important to note that *fill in* had the lowest frequency (12 tokens) in the 1930s, whereas it had the highest frequency (225 tokens) in the 2010s. It is worth pointing out that *fill in* was the most preferred by American and Canadian movie writers (581 tokens), followed by British and Irish ones (156 tokens), and Australian and New Zealand ones (16 tokens).

It is probably worthwhile pointing out that there was a steady increase (a rise of 22 tokens) in the frequency of *fill out* from the 1930s to the 1950s. More interestingly, there was a sudden fall (a decline of 3 tokens) in the frequency of *fill out* in the 1960s. It is interesting to note that there was a gradual rise (an increase of 30 tokens) in the frequency of *fill out* from the 1960s to the 1980s. Quite interestingly, there was a steady increase (a rise of 158 tokens) in the frequency of *fill out* from the 1990s to the 2010s. Most importantly, *fill out* had the highest frequency (276 tokens) in the 2010s, whereas it had the lowest frequency (7 tokens) in the 1930s. Most interestingly, *fill out* was the most preferred (703 tokens) by American and Canadian movie writers, followed by British and Irish ones (37 tokens), and Australian and New Zealand ones (10 tokens). It is noteworthy that *fill in* was always preferred over *fill out* by the movie writers of six countries from the 1930s to the 1970s, whereas *fill out* was preferred over *fill in* by those of six countries from the 1980s to the 2010s.

3. Fill in and fill out in the COCA

In what follows, we compare *fill in* with *fill out* by examining eight genres in the COCA. Table 2 shows the genre frequency of *fill in* and *fill out* in the COCA:

Table 2: Genre frequency of fill in and fill out in the COCA

Genre	All	Blog	Web	TV/M	Spok	Fic	Mag	News	Acad
Fill in	5,063	887	798	574	505	451	862	500	486
Fill out	4,599	771	764	708	496	347	583	640	290

An important question is “Which type is the preferable one for Americans?” Table 2 clearly shows that *fill in* is preferable to *fill out* in America. To be more specific, the overall frequency of *fill in* is 5,064 tokens, whereas that of *fill out* is 4,599 tokens. This in turn suggests that Americans prefer using *fill in* rather than using *fill out*.

It is worth observing that *fill in* and *fill out* rank first (887 tokens vs. 771 tokens) in the blog genre. Quite interestingly, *fill in* shows the same property as *fill out* in rank-one, hence revealing a high similarity in the blog genre. It should be noted, however, that *fill in* is favored over *fill out* in the blog genre. The frequency of *fill in* (887 tokens) is much higher than that of *fill out* (771 tokens). It can thus be inferred that American bloggers prefer using *fill in* rather than using *fill out*.

It is worthwhile mentioning that *fill in* ranks second (862 tokens) in the magazine genre, whereas *fill out* ranks second (764 tokens) in the web genre. More importantly, *fill in* does not show the same pattern as *fill out* in rank-two, thus revealing a low similarity. With respect to the magazine genre, it should be pointed out that the frequency of *fill in* (862 tokens) is much higher than that of *fill out* (583 tokens). This in turn suggests that American journalists prefer using *fill in* to using *fill out*. It must be noted, on the other hand, that the frequency of *fill in* (798 tokens) is higher than that of *fill out* (764 tokens) in the web genre. From this, it is evident that *fill in* is preferable to *fill out* in the web genre.

It would be worthwhile mentioning that *fill in* ranks third (798 tokens) in the web genre, whereas *fill out* ranks third (708 tokens) in the TV/movie genre. More interestingly, *fill in* and *fill out* show no similarity in rank-three. With respect to the TV/movie genre, it is worth pointing out that the frequency of *fill out* (708 tokens) is much higher than that of *fill in* (574 tokens). This in turn suggests that American celebrities prefer using *fill out* (708 tokens) rather than using *fill in* (574 tokens).

It is interesting to point out that *fill in* ranks fourth (574 tokens) in the TV/movie genre, whereas *fill out* ranks fourth (640 tokens) in the newspaper genre. Again, there is no similarity between *fill in* and *fill out* in rank-four. It should be pointed out, however, that the frequency of *fill out* (640 tokens) is higher than that of *fill in* (500 tokens) in the newspaper genre. From this, it is clear that American journalists prefer using *fill out* (640 tokens) to using *fill in* (500 tokens) in their newspapers.

It is interesting to note that *fill in* ranks fifth (505 tokens) in the spoken genre, whereas *fill out* ranks fifth (583 tokens) in the magazine genre. Again, there is no similarity between *fill in* and *fill out* in rank-five. It should be noted, however, that the frequency of *fill in* (505 tokens) is slightly higher than that of *fill out* (496 tokens) in the spoken genre. From this, it can be inferred that Americans prefer using *fill in* (505 tokens) rather than using *fill out* (496 tokens) in daily conversation.

Noteworthy is that *fill in* ranks sixth (500 tokens) in the newspaper genre, whereas *fill out* ranks sixth (496 tokens) in the spoken genre. Again, *fill in* does not show the same pattern as *fill out* in rank-six, thus revealing no similarity in rank-six.

It is worth noting that *fill in* ranks seventh (486 tokens) in the academic genre, whereas *fill out* ranks seventh (347 tokens) in the fiction genre. Again, there is no similarity between *fill in* and *fill out* in rank-seven. With respect to the academic genre, it should be noted that the frequency of *fill in* (486 tokens) is much higher than that of *fill out* (290 tokens). We take this as indicating that *fill in* is preferable to *fill out* in the academic genre. It is worthwhile noting, on the other hand, that the frequency of *fill out* (347 tokens) is lower than that of *fill in* (451 tokens) in the fiction genre. From this, it is clear that American writers prefer using *fill in* (451 tokens) to using *fill out* (347 tokens) in their novels.

Finally, *fill in* ranks eighth (451 tokens) in the fiction genre, whereas *fill out* ranks eighth (290 tokens) in the academic genre. Again, there is no similarity between *fill in* and *fill out* in rank-eight. To sum up, there is a high similarity between *fill in* and *fill out* in the blog genre, but there is no similarity between them in the other genres (7 genres). This in turn indicates that *fill in* is 12.5% the same as *fill out*.

Now attention is paid to the percentage of the use of *fill in* and *fill out* in eight genres:

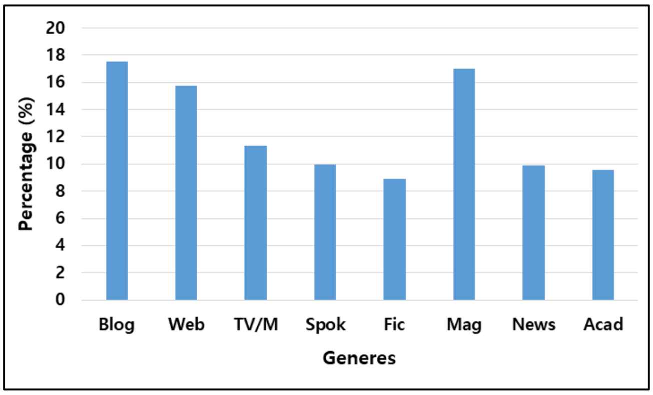

Fig 2: Percentage of the use of fill in in eight genres

As indicated in Figure 2, the blog genre is the most influenced by *fill in*, followed by the magazine genre, the web genre, the TV/movie genre, the spoken genre, the newspaper genre, the academic genre, and the fiction genre, in that order.

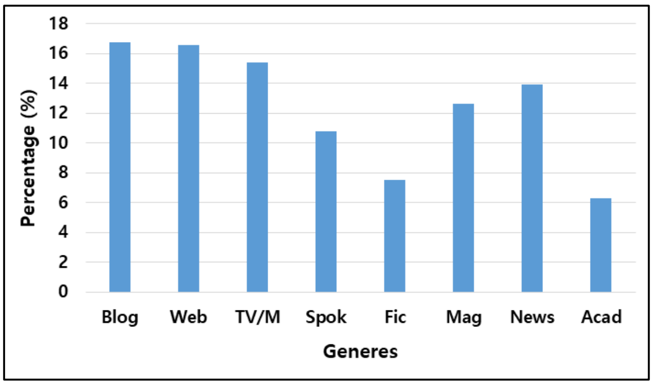

Fig 3: Percentage of the use of fill out in eight genres

As exemplified in Figure 3, the blog genre is the most influenced by *fill out*, followed by the web genre, the TV/movie genre, the newspaper genre, the magazine genre, the spoken genre, the fiction genre, and the academic genre, in descending order.

Now attention is paid to the distance between *fill in* and *fill out* in eight genres. Note that the Euclidean distance provides an index of how much alike *fill in* and *fill out* are in eight genres. We define the Euclidean distance as follows:

(1) The Euclidean distance

$$\mathrm{d(p,q)} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

The following table shows the distance between *fill in* and *fill out* in eight genres:

Table 3: Euclidean distance between fill in and fill out in eight genres

Genre	Blog	Web	TV/M	Spok	Fic	Mag	News	Acad
Percentage of fill in	17.51	15.76	11.33	9.97	8.90	17.02	9.87	9.59
Percentage of fill out	16.76	16.61	15.39	10.78	7.54	12.67	13.91	6.30
Euclidean distance	0.75	0.85	4.06	0.81	1.36	4.35	4.04	3.29

Quite interestingly, *fill in* is the furthest from *fill out* in the magazine genre. More specifically, the Euclidean distance between *fill in* and *fill out* in the magazine genre is 4.35, which is the highest. This in turn implies that *fill in* and *fill out* show a low similarity in the magazine genre. More importantly, *fill in* is the nearest to *fill out* in the blog genre. To be more specific, the Euclidean distance between *fill in* and *fill out* in the blog genre is 0.75, which is the lowest. This in turn suggests that *fill in* and *fill out* show a high similarity in the blog genre. We thus conclude that *fill in* is the nearest to *fill out* in the blog genre.

4. A collocation analysis of fill in and fill out in the COCA

In what follows, we provide a collocation analysis of *fill in* and *fill out* in the COCA. Table 4 shows the collocation of *fill in* in the top 28:

Table 4: Collocation of fill in in the COCA

Number	Collocation of fill in	Frequency
1	fill in gaps	61
2	fill in details	26
3	fill in blanks	11
4	fill in forms	11
5	fill in holes	8
6	fill in spaces	7
7	fill in data	5
8	fill in information	5
9	fill in wetlands	5
10	fill in words	5
11	fill in potholes	4
12	fill in part	4
13	fill in lips	4
14	fill in cracks	4
15	fill in bubbles	4
16	fill in areas	3
17	fill in space	3
18	fill in time	3
19	fill in voids	3
20	fill in worksheets	3
21	fill in wrinkles	3
22	fill in adjective	2
23	fill in banks	2
24	fill in bubble	2
25	fill in flash	2
26	fill in innings	2
27	fill in captchas	2
28	fill in order	2

An important question is “Which expression is the most preferred by Americans?” Table 4 clearly indicates that *fill in gaps* is the most frequently used one (61 tokens) in America. This in turn suggests that *fill in gaps* is the most preferable one (61 tokens) for Americans. As alluded to in Table 4, *fill in gaps* is the most preferred by Americans (61 tokens), followed by *fill in details*, *fill in blanks* (*fill in forms*), *fill in holes*, *fill in spaces*, and *fill in data* (*fill in information*), in that order. It is interesting to note that the everyday expressions *fill in blanks* and *fill in forms* rank third (11 tokens vs. 11 tokens) in the COCA. Quite interestingly, the expressions *fill in data* and *fill in information* rank seventh (5 tokens vs. 5 tokens) in the COCA. It is worth pointing out, on the other hand, that *fill in details* ranks second (26 tokens) in the COCA. Additionally, it should be noted that the everyday expression *fill in spaces* ranks sixth (7 tokens) in the COCA. We thus conclude that *fill in gaps* is the most preferable one (61 tokens) among Americans.

The following table shows the collocation of *fill out* in the top 28:

Table 5: Collocation of fill out in the COCA

Number	Collocation of fill out	Frequency
1	fill out forms	96
2	fill out paperwork	38
3	fill out applications	34
4	fill out questionnaires	24
5	fill out form	24
6	fill out surveys	14
7	fill out tax	11
8	fill out job	10
9	fill out papers	8
10	fill out college	6
11	fill out IRS	5
12	fill out adoption	5
13	fill out diaries	5
14	fill out reports	5
15	fill out grant	4
16	fill out expense	4
17	fill out application	4
18	fill out ballots	4
19	fill out cards	4
20	fill out absentee	3
21	fill out credit	3
22	fill out paper	3
23	fill out delivery	2
24	fill out documents	2
25	fill out evaluation	2
26	fill out health	2
27	fill out immunization	2
28	fill out insurance	2

An immediate question is “Which expression is the most preferred by Americans?” Table 5 clearly shows that *fill out forms* is the most frequently used one (96 tokens) in America. This in turn implies that *fill out forms* is the most preferable one (96 tokens) for Americans. As illustrated in Table 5, *fill out forms* is the most preferred (96 tokens) by Americans, followed by *fill out paperwork*, *fill out applications*, *fill out questionnaires*, *fill out form*, *fill out surveys*, and *fill out tax*, in descending order. It is significant to note that *fill in gaps* and *fill out forms* are the most preferable ones (61 tokens vs. 96 tokens) for Americans. It would be worthwhile mentioning, on the other hand, that *fill in forms* ranks third (11 tokens) in the COCA, whereas *fill out forms* ranks first (96 tokens). Quite interestingly, the expression *fill out questionnaires* ranks fourth (24 tokens) in the COCA. We thus conclude that *fill in gaps* and *fill out forms* are the most preferable ones (61 tokens vs. 96 tokens) among Americans.

Now attention is paid to the visualization of the collocations of *fill in* and *fill out* in the COCA:

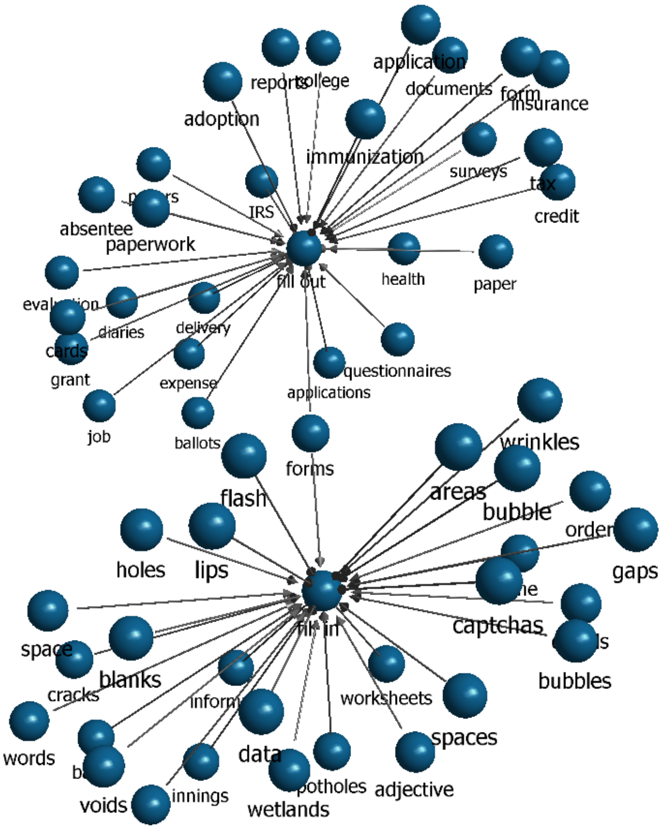

Fig 4: Visualization of fill in and fill out in the COCA

As exemplified in Figure 4, twenty seven nouns are linked to *fill in* and *fill out*, respectively. This indicates that these twenty seven nouns are the collocations of *fill in* and *fill out*, respectively. Most importantly, the noun *forms* is linked to both *fill in* and *fill out*, which indicates that it is the collocation of both *fill in* and *fill out*. From this, it is evident that 1.81% of fifty five nouns are the collocation of both *fill in* and *fill out*. This in turn suggests that *fill in* and *fill out* are low similarity synonyms.

5. Conclusion

To sum up, we have compared *fill in* and *fill out* in the MC and the COCA. In section 2, we have argued that the movie writers of six countries preferred using *fill out* rather than using *fill in*. We have further argued that *fill in* was always preferred over *fill out* by the movie writers of six countries from the 1930s to the 1970s, whereas *fill out* was preferred over *fill in* by those of six countries from the 1980s to the 2010s. In section 3, we have shown that there is a high similarity between *fill in* and *fill out* in the blog genre, but there is no similarity between them in the other genres (7 genres). This in turn suggests that *fill in* is 12.5% the same as *fill out*. We have further shown that *fill in* is the nearest to *fill out* in the blog genre. In section 4, we have maintained that *fill in gaps* and *fill out forms* are the most preferable ones (61 tokens vs. 96 tokens) for Americans. We have also contended that 1.81% of fifty five nouns are the collocation of both *fill in* and *fill out*. This in turn implies that *fill in* and *fill out* are low similarity synonyms.

References

1. Corpus of Contemporary American English (COCA). 2, March 2022. Online https://corpus.byu.edu/coca
2. Movie Corpus (MC). 2, March 2022. Online https://english-corpora.org /movies/
3. Murphy R. Grammar in Use. Cambridge University Press, 2016.
4. Murphy R. English Grammar in Use. Cambridge University Press, 2019.