Education, Language and Sociology Research
ISSN 2690-3644 (Print) ISSN 2690-3652 (Online)
Vol. 4, No. 1, 2023
www.scholink.org/ojs/index.php/elsr

Original Paper

A Comparative Analysis of Except and Except for in the COCA

Namkil Kang[1]

[1] Far East University, South Korea

Received: February 17, 2023 Accepted: February 28, 2023 Online Published: March 9, 2023

doi:10.22158/elsr.v4n1p21 URL: http://dx.doi.org/10.22158/elsr.v4n1p21

Abstract

The ultimate goal of this paper is to show how similar except and except for are in the Corpus of Contemporary American English (COCA). A point to note is that except and except for exhibit the same pattern in six genres, whereas they show a different pattern in two genres. What this suggests is that except is 75% the same as except for in the genre analysis of the COCA. A further point to note is that except and except for exhibit the lowest similarity in the fiction genre, whereas they reveal the highest similarity in the spoken genre. The COCA clearly shows that except people is the most preferable one among Americans, followed by except death, except Mr, and except water, in that order. The COCA further shows that except for people (45 tokens) is the most preferable one for Americans, followed by except for Mr, except for cases, and except for things, in that order. Finally, this paper argues that 21.95% of 41 nouns are the collocations of both except and except for. This amounts to saying that except is 21.95% the same as except for in the analysis of 41 collocations.

Keywords

Euclidean distance, token, COCA, collocation

1. Introduction

The main purpose of this paper is to provide a comparative analysis of *except* and *except for* in the Corpus of Contemporary American English. First, we aim to compare the use of *except* and that of *except for* in the eight genres of COCA. We, by comparing them, can see how similar *except* and *except for* are in the eight genres of the COCA. Put differently, we, by analyzing the ranking of *except* and *except for* in the eight genres of the COCA, can see how similar they are. Second, we aim to measure the distance between *except* and *except for* in the eight genres of the COCA. More specifically, we, by using the Euclidean distance, see how close *except* and *except for* are. Note that the more *except* and *except for* are close, the more they exhibit a similarity. Third, we aim at comparing the collocation of *except* and that of *except for* in the COCA. By comparing the collocations of *except* and *except for* in the COCA, we can see how similar they are in the COCA. Finally, we aim to capture the degree of the

similarity between *except* and *except for* by using the software package NetMiner. By linking the collocations of *except* and *except for*, we can calculate the degree of the similarity between *except* and *except for*.

2. The COCA and Eight Genres

In what follows, we aim to inquire into the similarity between *except* and *except for* in the eight genres of the COCA. Table 1 shows the frequency of *except* and *except for* in the COCA:

Table 1. Frequency of Except and Except for

GENRE	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
Except	97,441	14,256	16,003	12,686	7,757	20,688	9,740	7,971	8,310
Except for	27,234	3,440	3,580	3,467	2,139	6,857	2,778	2,455	2,518

It is probably worthwhile noting that the overall frequency of *except* in the COCA is 97,441 tokens, whereas that of *except for* is 27,234 tokens. Put differently, the overall frequency of *except* is three times higher than that of *except for*. It seems thus reasonable to hypothesize that Americans prefer using *except* to using *except for*.

Perhaps it is worthwhile pointing out that *except* and *except for* rank first (20,688 tokens vs. 6,857 tokens) in the fiction genre. This in turn means that the types *except* and *except for* exhibit a similarity in the fiction genre. Simply put, they exhibit the same pattern in rank-one. It should be pointed out, however, that in the fiction genre, the use of *except* is by far higher (three times) than that of *except for*. We take this as meaning that American writers prefer using *except* to using *except for* in their novels.

It is particularly noteworthy that *except* and *except for* rank second (16,003 tokens vs. 3,580 tokens) in the web genre. Again, the types *except* and *except for* show the same ranking in the web genre, thereby revealing a similarity. It must be stressed, however, that in the web genre, the use of *except* (16,003 tokens) is even higher (more than four times) than that of *except for*. This in turn implies that *except* (16,003 tokens) may be preferred over *except for* (3,580 tokens) by American web designers.

It is interesting to note that *except* ranks third (14,256 tokens) in the blog genre, whereas *except for* ranks third (3,467 tokens) in the TV/movie genre. Quite interestingly, the types *except* and *except for* show a different pattern, thus exhibiting no similarity. When it comes to the blog genre, the use of *except* (14,256 tokens) is much higher (more than four times) than that of *except for* (3,440 tokens). We take this as indicating that American bloggers prefer using *except* to using *except for*. It is worth observing, on the other hand, that in the TV/movie genre, the use of *except* (12,686 tokens) is still higher (more than three times) than that of *except for* (3,467 tokens). We take this as implying that American celebs like using *except*.

It is worth mentioning that *except* ranks fourth (12,686 tokens) in the TV/movie genre, whereas *except for* ranks fourth (3,440 tokens) in the blog genre. What this suggests is that the types *except* and *except for* show a mismatch between themselves with respect to their ranking, hence exhibiting no similarity.

It is interesting to point out that *except* and *except for* rank fifth (9,740 tokens vs. 2,778 tokens) in the magazine genre. More interestingly, the types *except* and *except for* exhibit the same ranking in the magazine genre, thereby showing a high degree of similarity in the magazine genre. It must be pointed out, on the other hand, that in the magazine genre, the use of *except* (9,740 tokens) is even higher (more than three times) than that of *except for* (2,778 tokens). This in turn shows that American journalists are fond of using *except* (9,740 tokens) rather than using *except for* (2,778 tokens).

It is worth pointing out that *except* and *except for* rank sixth (8,310 tokens vs. 2,518 tokens) in the academic genre. Again, the two types reveal the same property, thus showing a similarity again. It should be noted, however, that in the academic genre, the use of *except* (8,310 tokens) is by far higher (more than three times) than that *except for* (2,518 tokens). It can thus be inferred that teachers in America prefer using *except* rather than using *except for*.

It is interesting to observe that *except* and *except for* rank seventh (7,971 tokens vs. 2,455 tokens) in the newspaper genre. Exactly the same can be said of rank-seven. That is to say, the two types show the same property in rank-seven, thus revealing the similarity between them. It is important to mention, on the other hand, that the use of *except* (7,971 tokens) is still higher (more than three times) than that of *except for* (2,455 tokens). It seems thus reasonable to assume that American journalists are keen on using *except*.

It is worth noting that *except* and *except for* rank eighth (7,757 tokens vs. 2,139 tokens) in the spoken genre. Again, the types *except* and *except for* exhibit the same ranking in the spoken genre, hence revealing a similarity in rank-eight. It is significant to note, however, that the use of *except* (7,757 tokens) is much higher (more than three times) than that of *except for* (2,139 tokens). This in turn indicates that Americans like using *except* in daily conversation. To sum up, *except* and *except for* exhibit the same pattern in six genres, whereas they show a different pattern in two genres. From all of this, it seems clear that *except* is 75% the same as *except for* in the genre analysis of the COCA.

Now attention is paid to the percentage of *except* and *except for* in each genre:

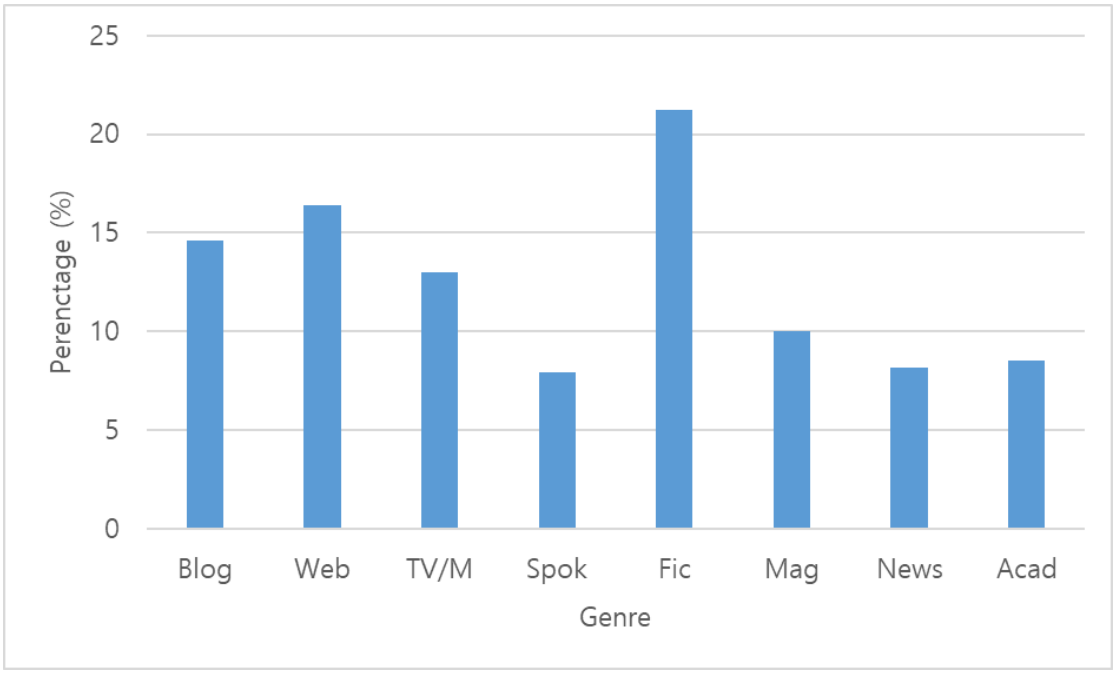

Figure 1. Percentage of Except

As exemplified in Figure 1, the fiction genre is the most influenced by *except*, followed by the web genre, the blog genre, the TV/movie genre, the magazine genre, the academic genre, the newspaper genre, and the spoken genre, in that order.

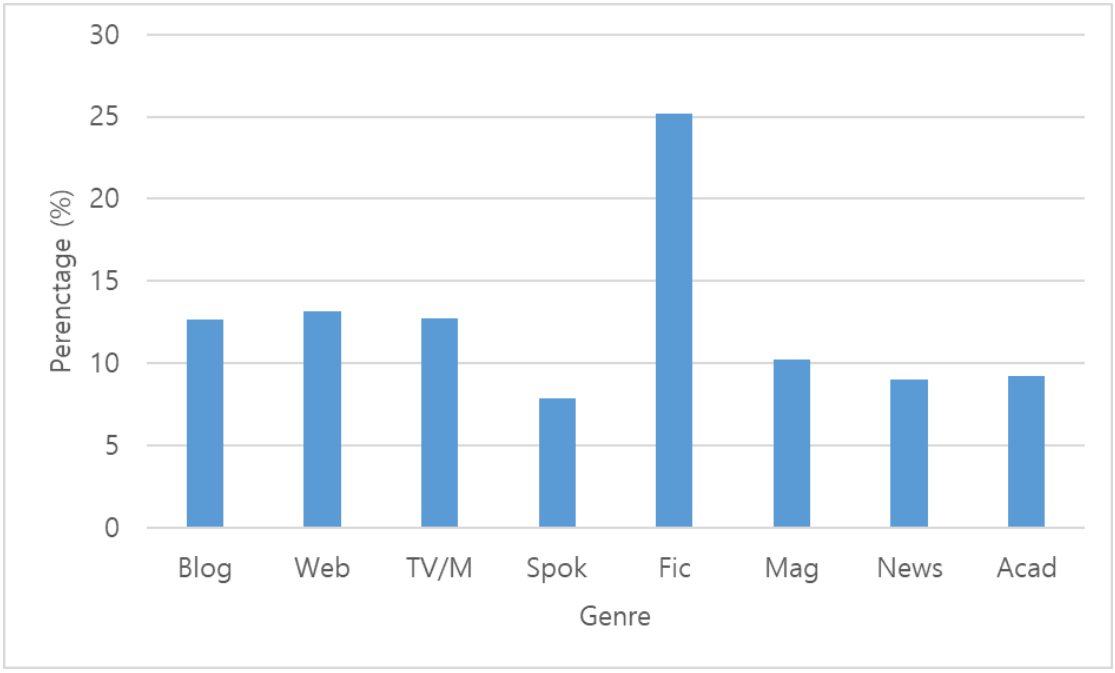

Figure 2. Percentage of Except for

As illustrated in Figure 2, the fiction genre is the most influenced by *except for*, followed by the web genre, the TV/movie genre, the blog genre, the magazine genre, the academic genre, the newspaper genre, and the spoken genre, in descending order.

3. The Euclidean Distance

In the following, we aim at investigating the similarity between *except* and *except for* in the eight genres of the COCA. Note, to begin with, that the Euclidean distance provides the similarity index between two types (*except* vs. *except for*) in each genre. It indicates that the more the distance between two types is close, the more they exhibit a similarity. Now let us define the Euclidean distance as follows:

(1) Euclidean distance

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Now attention is paid to the distance between *except* and *except for* in each genre:

Table 2. Euclidean Distance between Except and Except for

GENRE	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
Percentage of except	14.63	16.42	13.01	7.96	21.23	9.99	8.18	8.53
Percentage of except for	12.63	13.14	12.73	7.85	25.17	10.2	9.01	9.24
Euclidean distance	2	3.28	0.28	0.05	3.94	0.21	0.83	0.71

It is important to note that *except* is the furthest from *except for* in the fiction genre. To be more specific, the Euclidean distance between *except* and *except for* is 3.94, which is the highest among eight genres. This in turn suggests that the types *except* and *except for* reveal the lowest similarity. It is significant to note, on the other hand, that *except* is the nearest to *except for* in the spoken genre. More specifically, in the spoken genre, the Euclidean distance between *except* and *except for* is 0.05, which is the lowest among eight genres. This in turn implies that the types *except* and *except for* exhibit the highest similarity in the spoken genre. As exemplified in Table 2, the spoken genre reveals the highest similarity and the magazine genre, the TV/move genre, the academic genre, the newspaper genre, the

blog genre, the web genre, and the fiction genre follow. Thus, it can be concluded that *except* and *except for* show the highest similarity in the spoken genre.

4. The Collocations of Except and Except for in the COCA

In what follows, we aim to examine the collocations of *except* and *except for* in the COCA. Also, we compare the collocation of *except* and that of *except for* in the COCA. In both cases, our list was cut off in the top 25:

Table 3. Collocation of Except in the COCA

Number	Collocation	Frequency
1	Except people	72
2	Except death	46
3	Except Mr	41
4	Except water	34
5	Except Christmas	33
6	Except oil	27
7	Except time	27
8	Except food	24
9	Except money	22
10	Except sleep	21
11	Except play	20
12	Except wait	20
13	Except sex	19
14	Except treason	19
15	Except watch	19
16	Except work	19
17	Except men	18
18	Except talk	18
19	Except family	17
20	Except lag	16
21	Except Islam	15
22	Except thanksgiving	15
23	Except things	15
24	Except apple	14
25	Except chicken	14

It is important to mention that the expression *except people* has the highest frequency (72 tokens) in the COCA. This in turn means that the collocation *people* along with *except* is the most preferred one for Americans. It is interesting to consider the expression *except death*. It is worthwhile noting that the expression *except death* ranks second (46 tokens) in the COCA. This in turn implies that it is the second most widely used one (46 tokens). It is interesting to note that the expression *except Mr* ranks third (41 tokens) in the COCA. It is worth observing, on the other hand, that the expression *except water* ranks fourth (34 tokens) in the COCA. This in turn indicates that it is the fourth most occurred one (34 tokens). It seems thus reasonable to hypothesize that *except people* is the most preferable one among Americans, followed by *except death*, *except Mr*, and *except water*, in that order. It should also be pointed out that *except time* ranks sixth (27 tokens) in the COCA, whereas *except money* ranks ninth (22 tokens). To sum up, the expression *except people* is the most preferable one for Americans (72 tokens).

Now attention is paid to the collocation of *except for*:

Table 4. Collocation of Except for in the COCA

Number	Collocation	Frequency
1	Except for people	45
2	Except for Mr	22
3	Except for cases	20
4	Except for things	18
5	Except for emergency	16
6	Except for Mrs	16
7	Except for emergencies	15
8	Except for defense	14
9	Except for work	14
10	Except for food	12
11	Except for money	12
12	Except for age	11
13	Except for children	11
14	Except for school	11
15	Except for family	10
16	Except for parts	10
17	Except for aunt	9
18	Except for college	9
19	Except for compliance	9
20	Except for holidays	9

21	Except for right	9
22	Except for water	9
23	Except for gender	8
24	Except for sex	8
25	Except for provisions	8

It is significant to note that the expression *except for people* has the highest frequency (45 tokens). This in turn shows that *except for people* is the most preferred (45 tokens) by Americans. It should be noted, on the other hand, that *except Mr* ranks third (41 tokens) in the COCA, whereas *except for Mr* ranks second (22 tokens). This provides confirming evidence that *except* and *except for* exhibit a similar property with respect to the use of their collocation. It is interesting to observe the expression *except for cases*. The expression *except for Mr* is followed by the expression *except for cases* and the latter ranks third (20 tokens) in the COCA. It seems thus safe to assume that the expression *except for cases* is the third most widely used one (20 tokens) in the COCA. It should also be emphasized that *except for things* ranks fourth (18 tokens) in the COCA. It seems thus reasonable to hypothesize that *except for people* (45 tokens) is the most preferable one for Americans, followed by *except for Mr*, *except for cases*, and *except for things*, in that order. Additionally, it must be noted that *except work* ranks thirteenth (19 tokens) in the COCA, whereas *except for work* ranks eighth (14 tokens). To sum up, *except people* and *except for people* have the highest frequency, respectively, which in turn implies that they are the most preferable ones for Americans.

Now let us turn our attention to the visualization of the collocations of *except* and *except for*. We attempt to capture the collocations of *except* and *except for* in terms of the software package NetMiner. By linking their collocations, we can see how similar *except* and *except for* are. As exemplified in Figure 3, 16 nouns are linked to *except* and *except for*, respectively. Most importantly, only 9 nouns are linked to both *except* and *except for*. Note that in both cases, our list was cut off in the top 25:

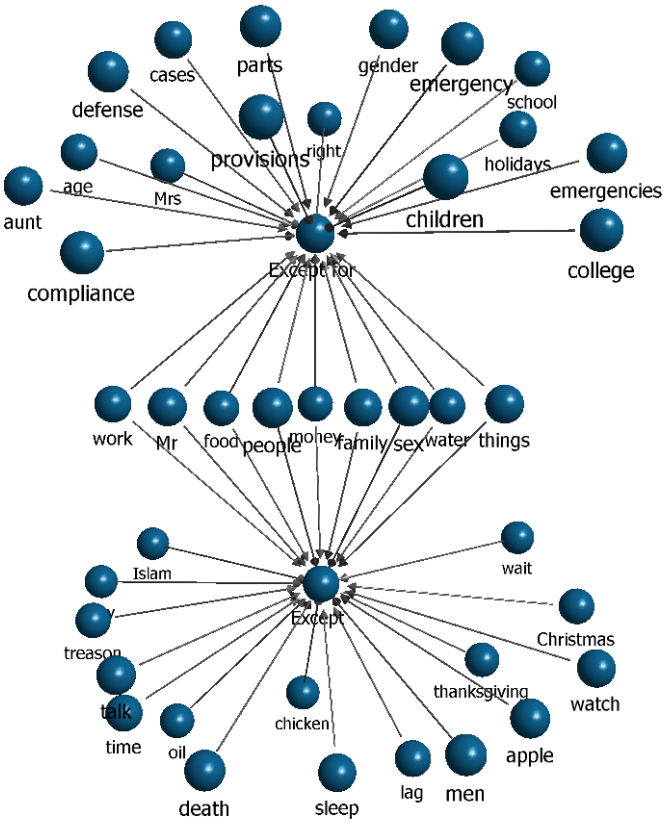

Figure 3. Visualization of the Collocations of Except and Except for

It must be emphasized that only 9 of 41 nouns are linked to both *except* and *except for*. These nouns are the collocations of both *except* and *except for*. The nine nouns linked to both *except* and *except for* are *work*, *Mr*, *food*, *people*, *money*, *family*, *sex*, *water*, and *things*. From all of this, it seems evident that 21.95% of 41 nouns are the collocations of both *except* and *except for*. It seems thus reasonable to conclude that *except* is 21.95% the same as *except for* in the analysis of 41 collocations. For the visualization of synonyms and keywords, see Kang (2022a, 2022b, 2022c, 2022d, 2023a, 2023b).

5. Conclusion

To sum up, we have shown how similar *except* and *except for* are in the COCA. In section 2, we have argued that *except* and *except for* exhibit the same pattern in six genres, whereas they show a different pattern in two genres. It seems thus clear that *except* is 75% the same as *except for* in the genre analysis of the COCA. In section 3, we have further argued that *except* and *except for* exhibit the lowest similarity in the fiction genre, whereas they show the highest similarity in the spoken genre. In section 4, we have contended that *except people* is the most preferable one among Americans, followed by *except death*, *except Mr*, and *except water*, in that order. We have also contended that *except for people* (45 tokens) is the most preferable one for Americans, followed by *except for Mr*, *except for cases*, and *except for things*, in that order. Finally, we have shown that 21.95% of 41 nouns are the collocations of

both *except* and *except for*. This in turn implies that *except* is 21.95% the same as *except for* in the analysis of 41 collocations.

References

Kang, N. (2022a). A Comparative Analysis of Search for and Look for in Four Corpora. *Advances in Social Sciences Research Journal*, *9*(3), 168-178.

Kang, N. (2022b). A Comparative Analysis of Impressed by and Impressed with in Two Corpora. *Theory and Practice in Language Studies*, *12*(5), 819-827.

Kang, N. (2022c). On Speak to and Talk to: A Corpora-based Analysis. *Theory and Practice in Language Studies, 12*(7), 1262-1270.

Kang, N. (2022d). On Speak with and Talk with: A Corpora-based Analysis. *International Journal of Social Science and Human Research*, *5*(8), 3354-3360.

Kang, N. (2023a). K-Pop in BBC News: A Big Data Analysis. *Advances in Social Sciences Research Journal*, *10*(2), 156-169.

Kang, N. (2023b). K-Dramas in Google: A NetMiner Analysis. *Transaction on Engineering and Computing Sciences*, *11*(1), 193-216.