International Journal of Humanities Social Science and Management (IJHSSM)
Volume 2, Issue 4, Sep.-Oct. 2022, pp: 39-44 www.ijhssm.org

On said to and supposed to: A Corpora-based Analysis

Namkil Kang
Far East University, South Korea

Date of Submission: 25-08-2022 Date of Acceptance: 07-09-2022

The main goal of this paper is to provide a comparative analysis of *said to* and *supposed to* in the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC).With respect to the BNC, it is interesting to note that *said to* and *supposed to* reveal different rankings in five genres, whereas they show the same ranking in the other two genres. This in turn shows that *said to* may be 28.57% the same as *supposed to* in seven genres. With respect to the COCA, it is worth pointing out that *said to* and *supposed to* show the same ranking in the TV/movie genre, whereas they show different rankings in the other seven genres. This in turn suggests that *said to* may be 12.5% the same as *supposed to* in eight genres. More interestingly, the COCA clearly shows that the expression *said to tell* may be the most preferred (216 tokens) by Americans, followed by *said to come*, and *said to make.* The COCA further shows that the expression *supposed to mean* may be the most preferred (2,306 tokens) by Americans, followed by *supposed to go*, and *supposed to get*. Talking about the COCA, 37.5% of thirty two verbs are the collocations of both *said to* and *supposed to*. Finally, the BNC shows that the expression *said to represent* may be the most preferred (31 tokens) by the British, followed by *said to exist*, and *said to give*. The BNC further shows that the expression *supposed to go* may be the most preferred (65 tokens) by the British, followed by *supposed to mean* (*supposed to take*), and *supposed to say*. When it comes to the BNC, 27.51% of thirty five verbs are the collocations of both *said to* and *supposed to*. It can thus be inferred that *said to* and *supposed to* may be low similarity synonyms in the COCA and BNC.

Keywords: supposed to, said to, COCA, BNC, type, token

I. INTRODUCTION

The main purpose of this paper is to compare *said to* and *supposed to* in the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC). As Murphy (2016, 2019) points out, "we can use *supposed to* in the same way as *said to*" (Murphy 2019: 90):

(1) I want to see that film. It is supposed to be good. (= people say it is good)
(Murphy 2019: 90)

As Murphy (2019) points out, *said to* and *supposed to* are used in the same way. The main goal of this paper is to show which type is preferred by Americans and the British and to differentiate between *said to* and *supposed to* in the COCA and the BNC. First, we aim to consider the frequency of *said to* in the COCA and that of *said to* in the BNC. On the other hand, we aim to examine the frequency of *supposed to* in the COCA and that of *supposed to* in the BNC. Second, we examine how similar they are in the COCA and the BNC. Third, we aim to explore a collocation relationship of *said to* and *supposed to* in the COCA and the BNC. In terms of the collocations of *said to* and *supposed to*, we can see how similar two types are in the COCA and BNC. This paper is organized as follows. In section 2, we argue that *said to* and *supposed to* reveal different rankings in five genres, but they show the same ranking in the other two genres. The BNC clearly shows that *said to* may be 28.57% the same as *supposed to* in seven genres. In section 3, we maintain that *said to* and *supposed to* show the same ranking in the TV/movie genre, but they show different rankings in the other seven genres. The COCA clearly indicates that *said to* may be 12.5% the same as *supposed to* in eight genres. In section 4, we show that *said to tell* may be the most preferred (216 tokens) by Americans, followed by *said to come*, and *said to make.* We also show that *supposed to mean* may be the most preferred (2,306 tokens) by Americans, followed by *supposed to go*, *supposed to get*, and *supposed to know*. The COCA shows that 37.5% of thirty two verbs are the collocations of both *said to* and *supposed to*. Finally, we contend that the expression *said to represent* may be the most preferred (31 tokens) by the British, followed by *said to exist*,

International Journal of Humanities Social Science and Management (IJHSSM)
Volume 2, Issue 4, Sep.-Oct. 2022, pp: 39-44 www.ijhssm.org

and *said to give*. We also contend that *supposed to go* may be the most preferred (65 tokens) by the British, followed by *supposed to mean* (*supposed to take*), and *supposed to say*. The BNC clearly indicates that 27.51% of thirty five verbs are the collocations of both *said to* and *supposed to*. We thus conclude that *said to* and *supposed to* may be low similarity synonyms in the COCAand BNC.

II. BRITISH NATIONAL CORPUS

In what follows, we aim to consider how similar *said to* and *supposed to* are in the BNC. Table 1 shows the frequency of *said to* and *supposed to* in the BNC:

Table 1 Frequency of said to and supposed to in the BNC

GENRE	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACAD	MISC
Said to	1,005	41	89	74	60	258	320	163
Supposed to	1,867	469	503	116	126	222	191	240

An immediate question is "Which type is the preferable one among the British?" Table 1 clearly shows that *supposed to* may be preferred over *said to* by the British. More specifically, the overall frequency of *said to* is 1,005 tokens, whereas that of *supposed to* is 1,867 tokens. This in turn suggests that the British prefer using *supposed to* (1,867 tokens) rather than using *said to* (1,005 tokens).

It is important to note that *said to* ranks first (320 tokens) in the academic genre, whereas *supposed to* ranks first (503 tokens) in the fiction genre, thus exhibiting a different property in rank-one. It should be pointed out that *said to* may be preferable to *supposed to* in the academic genre. As can be seen from Table 1, the frequency of *said to* (320 tokens) is still higher than that of *supposed to* (191 tokens) in the academic genre. This in turn implies that the British like using *said to* rather than using *supposed to* in the academic field. It is worth pointing out, on the other hand, that in the fiction genre, the frequency of *supposed to* (503 tokens) is five times higher than that of *said to* (89 tokens). This in turn suggests that British writers are fond of using *supposed to* rather than using *said to* in their novels.

It is worth observing that *said to* ranks second (258 tokens) in the non-academic genre, whereas *supposed to* ranks second (469 tokens) in the spoken genre. This indicates that *said to* and *supposed to* exhibit a different property in rank-two, hence showing no similarity in rank-two. It should be noted, however, that the type *said to* (258 tokens) is favored over the type *supposed to* (222 tokens) in the academic genre. This in turn shows that the British like using *said to* rather than using *supposed to* in the non-academic field. It is worthwhile pointing out, on the other hand, that in the spoken genre, the frequency of *supposed to* (469 tokens) is eleven times higher than that of *said to* (41 tokens). This in turn indicates that the British are fond of using *supposed to* rather than using *said to* in daily conversation.

It is interesting to point out that *said to* and *supposed to* rank third (163 tokens vs. 240 tokens) in the miscellaneous genre. Quite interestingly, *said to* and *supposed to* show the same ranking in the same genre, thereby showing a high similarity between them. It must be noted, however, that the frequency of *supposed to* (240 tokens) is far higher than that of *said to* (163 tokens) in the miscellaneous genre. This in turn shows that *supposed to* may be preferable to *said to* in the mixed genre.

It is worthwhile noting that *said to* ranks fourth (89 tokens) in the fiction genre, whereas *supposed to* ranks fourth (222 tokens) in the non-academic genre. More importantly, *said to* and *supposed to* show a different ranking in different genres, thus showing a low similarity.

It is interesting to note that *said to* ranks fifth (74 tokens) in the magazine genre, whereas *supposed to* ranks fifth (191 tokens) in the academic genre. Simply put, *said to* and *supposed to* reveal a different property in rank-five. It must be pointed out that in the magazine genre, the frequency of *supposed to* (116 tokens) is even higher than that of *said to* (74 tokens). We take this fact as implying that British journalists are fond of using *supposed to* (116 tokens) rather than using *said to* (74 tokens) in their magazines.

It is worthwhile pointing out that *said to* and *supposed to* rank sixth (60 tokens vs. 126 tokens) in the newspaper genre. Quite interestingly, *said to* and *supposed to* show the same ranking in the same genre, thereby showing a high similarity. It must be noted, however, that *supposed to* (126 tokens) is preferred over *said to* (60 tokens) by British journalists. This can be derived from the fact that in the newspaper genre, *supposed to* (126 tokens) is used more widely than *said to* (60 tokens).

Finally, it is worth noting that *said to* ranks seventh (41 tokens) in the spoken genre, whereas *supposed to* ranks seventh (116 tokens) in the magazine genre. That is to say, *said to* and *supposed*

International Journal of Humanities Social Science and Management (IJHSSM)
Volume 2, Issue 4, Sep.-Oct. 2022, pp: 39-44 www.ijhssm.org

to reveal a different ranking in different genres. To sum up, *said to* and *supposed to* reveal different rankings in five genres, whereas they show the same ranking in the other two genres. From all of this, it is evident that *said to* is 28.57% the same as *supposed to* in seven genres.

III. CORPUS OF CONTEMPORARY AMERICAN ENGLISH

In the following, we consider how similar *said to* and *supposed to*are in the eight genres of the COCA. Also, we aim to compare the frequency of two types in the COCA and that of two types in the BNC. Table 2 shows the frequency of *said to* and *supposed to* in the COCA:

Table 2 Frequency of said to and supposed to in the COCA

GENRE	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
Said to	5,741	542	834	1,157	318	825	671	441	953
Supposed to	47,721	4,669	4,169	17,395	5,967	7,411	3,268	3,459	1,383

An important question is "Which type is the preferable one among Americans?" Table 2 clearly shows that *supposed to* may be favored over *said to* in America. To be more specific, the overall frequency of *said to* is 5,741 tokens, whereas that of *supposed to* is 47,721 tokens. This in turn implies that Americans prefer using *supposed to* rather than using *said to*. As observed earlier, the British prefer using *supposed to* to using *said to*. It seems thus reasonable to assume that Americans and the British like using *supposed to* rather than using *said to*, thereby having a commonality.

It is significant to note that *said to* and *supposed to* rank first (1,157 tokens vs. 17,395 tokens) in the TV/movie genre. Note that two types exhibit the same ranking in the same genre, hence revealing a high similarity. It must be pointed out, however, that the frequency of *supposed to* (17,395 tokens) is fifteen times higher than that of *said to* (1,157 tokens). We take this as suggesting that American celebs are fond of using *supposed to* rather than using *said to*.

It is noteworthy that *said to* ranks second (953 tokens) in the academic genre, whereas *supposed to* ranks second (7,411 tokens) in the fiction genre. Quite interestingly, *said to* and *supposed to* do not show the same ranking in rank-two, thus showing a low similarity between them. It is interesting to note that in the academic genre, the frequency of *supposed to* (1,383 tokens) is much higher than that of *said to* (953 tokens). This in turn indicates that *supposed to* may be preferable to *said to* in the academic field. More importantly, *said to* (320 tokens) is favored over *supposed to* (191 tokens) in the academic field of the BNC. We take this as implying that the British like using *said to* rather than using *supposed to* in the academic field. It seems thus safe to assume that Americans and the British exhibit a different property with respect to the use of *said to* and *supposed to* in the academic field. It is worthwhile pointing out, on the other hand, that the frequency of *supposed to* (7,411 tokens) is nearly nine times higher than that of *said to* (825 tokens) in the fiction genre. This in turn indicates that American writers are fond of using *supposed to* rather than using *said to* in their novels. Exactly the same can be said of the BNC. In the fiction genre of the BNC, the frequency of *supposed to* (503 tokens) is five times higher than that of *said to* (89 tokens). This in turn implies that just as in the case of American writers, British writers prefer to use *supposed to* rather than use *said to*, thereby having a commonality.

It is worthwhile noting that *said to* ranks third (834 tokens) in the web genre, whereas *supposed to* ranks third (5,967 tokens) in the spoken genre. Again, *said to* and *supposed to* show a different property with respect to rank-three, hence revealing no similarity between them. When it comes to the web genre, *supposed to* (4,169 tokens) is used more widely than *said to* (834 tokens). In a word, *supposed to* may be favored over *said to* in the web genre. Talking about the spoken genre, *supposed to* (5,967 tokens) is preferred over *said to* (318 tokens) by Americans. This can be derived from the fact that the frequency of *supposed to* (5,967 tokens) is eighteen times higher than that of *said to* (318 tokens). The same can be said about the BNC. Just as in the case of Americans, the British prefer using *supposed to* (469 tokens) to using *said to* (41 tokens) in daily conversation.

It is worthwhile observing that *said to* ranks fourth (825 tokens) in the fiction genre, whereas *supposed to* ranks fourth (4,669 tokens) in the blog genre. As expected, *said to* and *supposed to* exhibit a different property with regard to rank-four, thus showing no similarity between them. With respect to the blog genre, it is interesting to point out that *supposed to* (4,669 tokens) may be preferable to *said to* (542 tokens).

Noteworthy is that *said to* ranks fifth (671 tokens) in the magazine genre, whereas *supposed to* ranks fifth (4,169 tokens) in the web genre. Again, two

International Journal of Humanities Social Science and Management (IJHSSM)
Volume 2, Issue 4, Sep.-Oct. 2022, pp: 39-44 www.ijhssm.org

types do not show the same ranking in rank-five, thus revealing no similarity. It should be noted, however, that in the magazine genre, the frequency of *supposed to* (3,268 tokens) is nearly five times higher than that of *said to* (671 tokens). This in turn suggests that American journalists like using *supposed to* in their magazines. The same applies to the BNC. As in the case of the COCA, the BNC clearly shows that *supposed to* (116 tokens) is favored over *said to* (74 tokens) in the magazine genre.

It is interesting to note that *said to* ranks sixth (542 tokens) in the blog genre, whereas *supposed to* ranks sixth (3,459 tokens) in the newspaper genre. Again, *said to* and *supposed to* show a different ranking in rank-six, hence showing a low similarity between them. Note that in the newspaper genre, *supposed to* (3,459 tokens) is favored over *said to* (441 tokens). Most importantly, American English and British English have a commonality with respect to the use of *said to* and *supposed to* in the newspaper genre. Talking about the newspaper genre, *supposed to* (126 tokens) is preferred over *said to* (60 tokens) by British journalists.

It is interesting to point out that *said to* ranks seventh (441 tokens) in the newspaper genre, whereas *supposed to* ranks seventh (3,268 tokens) in the magazine genre. Thus, *said to* and *supposed to* reveal no similarity in rank-seven.

Finally, *said to* ranks eighth (318 tokens) in the spoken genre, whereas *supposed to* ranks eighth (1,383 tokens) in the academic genre. Again, *said to* and *supposed to* exhibit no similarity in rank-eight. To sum up, *said to* and *supposed to* show the same ranking in the TV/movie genre, whereas they show a different ranking in the other seven genres. We thus conclude that *said to*may be 12.5% the same as *supposed to* in the eight genres of the COCA.

IV. A COLLOCATION ANALYSIS OF SAID TO AND SUPPOSEDTO IN THE COCA AND BNC

In this section, we aim to compare the collocations of *said to* and *supposed to* in the COCA and the BNC. Table 3 shows the collocations of *said to* and *supposed to* in the COCA:

Table 3 Collocations of said to and supposed to in the COCA

Number	Collocations of said to	Frequency	Collocations of supposed to	Frequency
1	said to tell	216	supposed to mean	2,306
2	said to come	150	supposed to go	2,187
3	said to make	131	supposed to get	1,838
4	said to go	119	supposed to know	1,604
5	said to call	116	supposed to say	1,355
6	said to take	114	supposed to make	1,322
7	said to get	111	supposed to take	1,187
8	said to give	103	supposed to meet	973
9	said to bring	88	supposed to happen	959
10	said to represent	81	supposed to work	838
11	said to say	81	supposed to come	809
12	said to contain	78	supposed to help	783
13	said to exist	72	supposed to tell	729
14	said to meet	71	supposed to look	686
15	said to wait	71	supposed to give	602
16	said to keep	66	supposed to keep	526
17	said to stay	62	supposed to believe	499
18	said to look	54	supposed to protect	486
19	said to possess	51	supposed to talk	486
20	said to include	50	supposed to feel	459
21	said to use	48	supposed to see	448
22	said to put	46	supposed to call	377

An important question is "Which collocation is the most widely used one with *said to*?" Table 3 clearly indicates that the verb *tell* is the most widely used (216 tokens) with *said to*. This in turn suggests that the expression *said to tell* is the most preferable one (216 tokens) among Americans. As illustrated in Table 3, *said to tell* is the most preferred (216 tokens) by Americans, followed by *said to come*,

International Journal of Humanities Social Science and Management (IJHSSM)
Volume 2, Issue 4, Sep.-Oct. 2022, pp: 39-44 www.ijhssm.org

said to make, *said to go*, *said to call*, *said to take*, and *said to get*, in that order. Now an immediate question is "Which collocation is the most frequently used one with *supposed to*?" Table 3 clearly shows that the verb *mean* is the most frequently used (2,306 tokens) with *supposed to*. This in turn implies that *supposed to mean* is the most preferable one (2,306 tokens) for Americans. As indicated in Table 3, *supposed to mean* is the most preferred (2,306 tokens) by Americans, followed by *supposed to go*, *supposed to get*, *supposed to know*, *supposed to say*, *supposed to make*, *supposed to take*, and *supposed to meet*, in descending order. Most importantly, twelve of thirty two verbs are the collocations of both *said to* and *supposed to*. The twelve verbs are *go*, *get*, *say*, *make*, *take*, *meet*, *come*, *tell*, *look*, *give*, *keep*, and *call*. From all of this, it is evident that 37.5% of thirty two verbs are the collocations of both *said to* and *supposed to*. We thus conclude that *said to* and *supposed to* may be low similarity synonyms.

Now attention is paid to the collocations of *said to* and *supposed to* in the BNC:

Table 4 Collocations of said to and supposed to in the BNC

Number	Collocations of said to	Frequency	Collocations of supposed to	Frequency
1	said to represent	31	supposed to go	65
2	said to exist	27	supposed to mean	59
3	said to give	22	supposed to take	59
4	said to come	15	supposed to say	56
5	said to constitute	14	supposed to know	50
6	said to provide	14	supposed to get	47
7	said to tell	14	supposed to make	47
8	said to take	13	supposed to come	39
9	said to help	12	supposed to give	33
10	said to include	12	supposed to help	24
11	said to contain	11	supposed to keep	24
12	said to bring	11	supposed to tell	24
13	said to involve	11	supposed to look	23
14	said to lie	11	supposed to put	23
15	said to make	11	supposed to happen	22
16	said to indicate	10	supposed to feel	21
17	said to occur	10	supposed to represent	20
18	said to possess	10	supposed to meet	18
19	said to reflect	10	supposed to provide	18
20	said to resemble	10	supposed to see	18
21	said to show	10	supposed to start	18
22	said to go	9	supposed to work	17

An important question is "Which collocation is the most widely used with *said to* in the BNC?" Table 4 clearly shows that the verb *represent* is the most widely used (31 tokens) with *said to*. As exemplified in Table 4, *said to represent* may be the most preferred (31 tokens) by the British, followed by *said to exist*, *said to give*, *said to come*, *said to constitute* (*said to provide*, *said to tell*), and *said to take*, in that order. On the other hand, as can be seen from Table 4, *supposed to go* is the most widely used (65 tokens) in the BNC. It must be pointed out that *supposed to go* may be the most preferred (65 tokens) by the British, followed by *supposed to mean* (*supposed to take*), *supposed to say*, *supposed to know*, *supposed to get* (*supposed to make*), and *supposed to come*, in descending order. Most importantly, nine verbs are the collocations of both *said to* and *supposed to* in the BNC. Nine of thirty five verbs are *go*, *take*, *make*, *come*, *give*, *help*, *tell*, *represent*, and *provide*. That is to say, 27.51% of thirty five verbs are the collocations of both *said to* and *supposed to*. We thus conclude that *said to* and *supposed to* may be low similarity synonyms in the BNC.

V. CONCLUSION

To sum up, we have provided a comparative analysis of *said to* and *supposed to*in two corpora (the COCA and the BNC). In section 2, we have argued that *said to* and *supposed to* reveal different rankings in five genres, whereas they show the same ranking in

the other two genres. The BNC clearly shows that *said to* may be 28.57% the same as *supposed to* in the seven genres of the BNC. In section 3, we have maintained that *said to* and *supposed to* show the same ranking in the TV/movie genre, whereas they show different rankings in the other seven genres. The COCA clearly indicates that *said to* may be 12.5% the same as *supposed to* in the eight genres of the COCA. In section 4, we have shown that *said to tell* may be the most preferred (216 tokens) by Americans, followed by *said to come*, *said to make*, *said to go*, and *said to call*. We have also shown that the expression *supposed to mean* may be the most preferred (2,306 tokens) by Americans, followed by *supposed to go*, *supposed to get*, *supposed to know*, and *supposed to say*. Quite interestingly, the COCA shows that 37.5% of thirty two verbs are the collocations of both *said to* and *supposed to*. This in turn implies that *said to* and *supposed to* may be low similarity synonyms. Finally, we have contended that the expression *said to represent* may be the most preferred (31 tokens) by the British, followed by *said to exist*, *said to give*, and *said to*. We have also contended that *supposed to go* may be the most preferred (65 tokens) by the British, followed by *supposed to mean* (*supposed to take*), *supposed to say*, and *supposed to know*. The BNC clearly shows that 27.51% of thirty five verbs are the collocations of both *said to* and *supposed to*. It can thus be concluded that *said to* and *supposed to* may be low similarity synonyms in the BNC.

REFERENCES

[1]. British National Corpus (BNC). 1, August 2022. Online https://corpus.byu.edu/bnc
[2]. Corpus of Contemporary American English (COCA). 1, August 2022. Online https://corpus.byu.edu/coca
[3]. Murphy, R. (2016). *Grammar in Use*. Cambridge University Press.
[4]. Murphy, R. (2019). *English Grammar in Use*. Cambridge University Press.