A Corpus-Based Study of Semantic Categorizations of Attracted Adjectives to the *it BE ADJ clause* Construction

Jiaojiao Wang¹,² and Jiangping Zhou²,³

**Abstract**

By implementing such corpus approaches as collexeme analysis and hierarchical cluster analysis, this paper investigates the semantic categorizations of adjectives that are extremely significantly attracted to the *it BE ADJ clause* construction in English. The findings demonstrate that these adjectives could be at least subdivided into seven different semantic clusters that denote (1) importance, necessity, and possibility; (2) appropriateness, reasonability, and unreality; (3) impracticability and irrelevance; (4) undeniability and axiomaticity; (5) obviousness; (6) dubiety, desirability, and ease; (7) improbability and anomalousness.

**Keywords**

attracted adjectives, *it BE ADJ clause* construction, semantic categorizations, collexeme analysis, hierarchical cluster analysis

**Introduction**

The *it BE ADJ clause* construction is a grammatical pattern that comprises a projecting clause and a projected clause in the form of hypotactic clause complex, wherein the projecting clause is realized by a matrix clause composed of an anticipatory *it*, a copular *be*, and an adjective, and the projected clause is realized by an infinitive clause, an -*ing* clause, or a *that*-clause. This construction has been termed as “*it*-extraposition” in previous studies (Kaltenböck, 2003, 2005; Quirk et al., 1985; Zhang, 2017; inter alia). Quirk et al. (1985) defined it as a syntactic process of shifting the clause (e.g., finite clause, infinitive, and -*ing*-clause) in the subject position to the right position of the superordinate predicate while replacing it with an anticipatory *it*. Illustrating examples are demonstrated in (1a–c) (Kaltenböck, 2005, p. 120, original emphasis).

(1) a. It is surprising that John went to Paris.
   b. It will be imperative to find a job in Paris.
   c. It is fun living in Paris.

The non-extraposition clause of (1a) is that *John went to Paris* is surprising, and (1a) is the result of moving the finite clause *John went to Paris* from the subject position to the superordinate predicate position by adding an anticipatory *it* in the original position. In the same vein, non-extraposition clauses in (1b) and (1c) are to *find a job in Paris will be imperative* and *living in Paris is fun* respectively. The reasons that we do not term the same grammatical pattern as “*it*-extraposition” but as *it BE ADJ clause* construction are twofold. First, the former foregrounds the movement of the finite clause, infinitive, or -*ing*-clause, whereas the latter foregrounds the adjective in the construction; second, the former emphasizes little on the choices of adjectives while discussing *it*-extraposition clauses, whereas the latter caters to our topic in this research that which adjectives are significantly attracted to the *it BE ADJ clause* construction.

Previous researches that closely associated with the present study are the exploration of *it’s ADJ to V* construction, based on the data sourced from British National Corpus or BNC, conducted by Hilpert (2014). He investigated the most attracted adjectives in this construction by using a collexeme analysis (for detailed information about this analysis, cf. Stefanowitsch, 2020; Stefanowitsch & Gries, 2003; section 3.3) and the interdependencies between adjectives and verbs by using a covarying collexeme analysis (cf. Hsiao & Mahastuti, 2020; Stefanowitsch & Gries, 2005). However, his collexeme analysis of the *it’s ADJ to V* construction (his covarying collexeme analysis of this construction concerning the interdependencies is not considered in this research) could be...
Further expanded in terms of two respects. On the one hand, the corpus he adopted is BNC which is typical of British English. We can also explore the same linguistic phenomenon by looking into the Corpus of US Supreme Court Opinions or COUSCO which is characteristic of American English to testify whether the findings surveyed in BNC are universal of dialectal Englishes. On the other hand, Hilpert (2014) delimited the *it’s ADJ to V* construction within the present tense of the copula verb, that is, *is*, and the infinitive clause as the projected clause. His research is obviously oblivious of the modification of a modal auxiliary to the copula verb (e.g., it could be ADJ to V), the modification of an adverb to the adjective (e.g., it’s extremely ADJ to V), and the interpolation of a prepositional phrase (e.g., it’s ADJ for you to V). His research, furthermore, did not consider other projected clauses such as *it’s ADJ-ing-clause* and *it’s ADJ that-clause*. Therefore, we in this research intend to examine not only these variants of the *it BE ADJ clause* construction, but also the linguistic phenomenon in American English. Pertaining to the way he clustered these attracted adjectives into different groups, he more or less divided them subjectively by ignoring the collocates they accompany, while we in this research uses a hierarchical clustering analysis or HCA to agglomerate the adjectives that occur in the similar linguistic context so as to cluster these attracted adjectives as objectively as possible. Accordingly, relevant research questions are proposed.

1. What are the significantly attracted adjectives in the *it BE ADJ clause* construction in COUSCO?
2. What are the semantic categorizations of these attracted adjectives in the *it BE ADJ clause* construction in COUSCO?

This research is outlined as follows. Section 2 delineates the *it BE ADJ clause* construction and its previous collexeme analysis. Section 3 profiles the corpus we use, the way that data are collected, and the methods that we adopted to facilitate our analysis. Section 4 is to identify what the significantly attracted adjectives are in the *it BE ADJ clause* construction, and section 5 is to cluster these attracted adjectives and analyze what specific meanings this construction denote. Section 6 is the summary of this research.

**Theoretical Framework**

*The it BE ADJ clause Construction in English*

The *it BE ADJ clause* construction is composed of two parts which form a clause complex. The first three elements in this construction constitute the matrix clause, and the final element forms the subordinate clause (cf. Zhang, 2017). Besides, each element in the four slots is filled by certain lexical items. In other words, the first slot in the construction could only be filled by the anticipatory *it*; the copula verb *BE* in the second slot could be either in the present tense (i.e., *is*) or in the past tense (i.e., *was*); the third slot is filled by an adjective; and the final slot is filled by a clause in Hallidayan sense, that is, any finite or non-finite clauses (cf. Halliday, 1985, 1994; Halliday & Matthiessen, 2004, 2014; Hao, 2020; He, 2019, 2020; inter alia). This construction is typically exemplified by the following clause which is sourced from Hilpert (2014, p. 393).

\[(2) \text{ It is hard to be a corpus linguist.} \]

This typical construction has some variations, which are shown in the following examples (wherein examples in (3) and (4) are sourced from COUSCO, and examples in (5) are adapted from example (1c)). Examples (3) demonstrate variants of *it BE ADJ clause* construction that the clause in the fourth slot is realized by an infinitive; examples (4) by a finite clause; and examples (5) by an *-ing-clause*. Each variant from (3) to (5) is either realized typically (i.e., (a)), or the copula verb is modified by a modal auxiliary (i.e., (b)), or the adjective is modified by an adverb (i.e., (c)).

\[(3) \text{ it BE ADJ to-clause} \]

\[a. \text{ On the occasion of this purchase, I told them that it was impossible to tell what the quality of the madder was unless I examined it.} \]

\[b. \text{ It would be difficult to disturb a claim thus sanctioned by time, however unfounded it might have been in its origin.} \]

\[c. \text{ It is entirely reasonable to limit the award of attorney’s fees to those parties who, in order to obtain relief, found it necessary to file a complaint in court.} \]

\[(4) \text{ it BE ADJ that-clause} \]

\[a. \text{ From the establishment of some facts, it is possible that others may be presumed, and less than positive testimony may establish facts.} \]

\[b. \text{ It should be obvious that the powers exercised by territorial courts tell us nothing about the nature of an entity, like the Tax Court, which administers the general laws of the Nation.} \]

\[c. \text{ It is equally true that the state may invest local bodies called into existence for purposes of local administration with authority in some appropriate way to safeguard the public health and the public safety.} \]

\[(5) \text{ it BE ADJ -ing-clause} \]

\[a. \text{ It was fun living in Paris.} \]

\[b. \text{ It would be fun living in Paris.} \]

\[c. \text{ It is quite fun living in Paris.} \]

Previous researches were conducted by focusing on one of the elements in the four slots. Kaltenböck (2003) examined the semantic status of *it* in the first slot of the construction and
argued against the claim that it at issue is a meaningless, semantically empty dummy element because this view failed to consider the actual use of it in the context. There were also researches which were conducted to explore the copula BE (Herriman, 2000). Herriman claimed that narrative texts favor the past tense while expository texts prefer the present tense. Collins (1994) and Herriman (2000) focused on the communicative and functional factors of the extraposed element respectively. Collins argued that the to-clause and that-clause may be freely extraposed if there are no grammatical factors (e.g., a matrix predicate containing a subordinate clause or an identified complement) impeding the extraposition; the -ing-clause, which is more highly nominalized, extraposes less freely (Collins, 1994). Herriman investigated the three functions that the construction performs, that is, ideational, interpersonal, and textual functions, and variations of the three functions in different text types such as fiction, reportage, and so on (Herriman, 2000). With respect to the adjectives in the third slot, they are considered by Collins (1994) and Hilpert (2014). Collins categorized adjectives in this construction semantically into five different types: emotional and rational judgment (e.g., fascinating, true, and clear), deontic conditions (e.g., necessary, desirable, and better), potentiality (e.g., possible and impossible), ease/difficulty (e.g., easy, difficult, and hard), and usuality (e.g., customary, usual, and common).

According to Collins (1994), the first three types occurred with both to-clause and that-clause in this construction, and the last two types occurred only with to-clause, while the -ing-clause is very rarely identified. However, Hilpert (2014) also categorized these adjectives semantically based not on their raw and/or normalized frequencies, but on their association strengths with the it's ADJ to V construction. Hilpert's study is further scrutinized in the following subsection.

**Collexeme Analysis of the it's ADJ to V Construction**

Hilpert (2014) conducted a collexeme analysis on the attracted adjectives in the it's ADJ to V construction, which is only one variant of the it BE ADJ clause construction. The implementation of collexeme analysis is to identify which lexical items in a given slot are most strongly attracted to a grammatical construction, that is, the preferred adjectives of the it's ADJ to V construction in Hilpert's research. According to his research findings, the top 10 attracted adjectives to this construction are difficult, easy, essential, hard, important, impossible, interesting, necessary, possible, and advisable. One may wonder that, by instinct, these adjectives are closely associated with the construction. Indeed, our instinct is confirmed by the observed frequencies retrieved from BNC.

However, our instinct will also mislead us that a certain adjective with a higher observed frequency is more attracted to the construction than that with a lower frequency. For instance, possible (2,434) is more frequent than difficult (1,949), easy (1,139), essential (338), hard (979), important (1,844), impossible (995), interesting (453), and necessary (1,367), but its collostructional strength with the it's ADJ to V construction is the least of all (cf. Table 3 in Hilpert, 2014). Therefore, this additional effort of a collexeme analysis is actually warranted.

In addition to the investigation of the ranking of the attracted adjectives based on their collostructional strength, analysts should also make sense of these data qualitatively. In Hilpert's research (2014), he interpreted these data to assess the constructional semantics, that is, the investigation of groups of semantically related items that are attracted to the construction. Specifically, he identified these attracted adjectives to make references to different scales: ease (difficult, easy, and hard), possibility (possible and impossible), importance (important, necessary, and essential), and advisability (advisable, better, best, and wise). However, his categorization of these attracted adjectives is problematic in at least two respects. On the one hand, the categorization is subjectively oriented because he does not consider uses of these adjectives in their linguistic context, for instance, the collocation with other adverbs or nouns in the same clause. On the other hand, polysemous elements such as hard, as Hilpert comments, will weaken their collostructional strength. Precisely, the adjective at issue hard denotes both the sense of “difficult” and that of “solid” and the less frequently used sense will weaken the association strength between the adjective hard and the it's ADJ to V construction. The analysts’ subjective categorization unavoidably ignores the demarcation of the two different senses. Nevertheless, this research, while considering the clustering of these semantically related adjectives, examines not only the adverbs that modify the adjectives but also the nouns that are modified by the adjectives such as a possible plan, and that are complemented by adjectives such as the plan is advisable. By so doing, the clustering of these attracted adjectives is conducted as objectively as possible.

Considering the various variations of the it BE ADJ clause construction and the categorization of attracted adjectives in this construction, we will first, in the following sections, construct relevant search queries to retrieve these variant constructions, and then compute and rank the most attracted adjectives in this construction, and finally cluster these attracted adjectives that are semantically related in terms of their collocated items in the same clause in the linguistic context.

**Methodology**

**Corpus**

Corpus of US Supreme Court Opinions or COUSCO (https://www.english-corpora.org/scotus/), which was released in March 2017, contains approximately 130 million words in 32,000 Supreme Court decisions from the 1790s to the 2010s. Texts in this corpus were taken from FindLaw.com and Justin,
and compared against the information from Cornell University to make sure that there were no missing texts.

**Data Collection**

In order to retrieve occurrences of the *it BE ADJ clause* construction in the corpus as exhaustively as possible, we constructed corresponding search queries or SQs, which are exemplified and expounded as follows.

SQ 1: it [vb*] [j*] to that [v?g*]
SQ 2: it [vb*] [j*] to that [v?g*]
SQ 3: it [vb*] [j*] for [nn*] [pp*] to that [v?g*]
SQ 4: it [vb*] [j*] for [mn*] [pp*] to that [v?g*]
SQ 5: it [vm*] [vb*] [j*] to that [v?g*]
SQ 6: it [vm*] [vb*] [j*] to that [v?g*]
SQ 7: it [vm*] [vb*] [j*] for [nn*] [pp*] to that [v?g*]
SQ 8: it [vm*] [vb*] [j*] for [nn*] [pp*] to that [v?g*]

These search queries are generally classified into two types. One type is from SQ1 to SQ4 in that the *BE verb* in these queries is not modified by any modal elements and thus delimits itself within *is* and *was*. The other type is from SQ5 to SQ8 because these queries are all modified by a modal auxiliary. Accordingly, SQ1 could be expounded as a construction in a sequence of *it*, *a modal auxiliary*, the *copula be*, an adjective, and a clause. SQ2 is different from SQ1 in that the adjective in SQ2 is modified by an element in any form (the most frequent element is such adverbs as *extremely*, *quite*, etc.); SQ3 differs from SQ1 in that it is interpolated by a *for-headed prepositional phrase* between the adjective and the clause; SQ4 contains both new elements compared with SQ1, that is, it is not only the case that the adjective is modified by an element in any form, but also the case that a *for-headed prepositional phrase* is interpolated between the adjective and the clause. These search queries are exemplified by (6a–d) respectively, in which examples (6a and 6b) are rewritten from examples (3a and 3c) and (6c and 6d) are sourced from the corpus.

(6) SQ1 to SQ4

a. On the occasion of this purchase, I told them that *it was impossible to* tell what the quality of the madder was unless I examined it. (CUSCO, 68 U.S. 359)
b. *It is entirely reasonable to* limit the award of attorney’s fees to those parties who, in order to obtain relief, found it necessary to file a complaint in court. (CUSCO, 479 U.S. 6)
c. This certainly is a decision that *it was competent for Congress to* make the revival of an act depend upon the proclamation of the President, showing the ascertainment by him of the fact that the edicts of certain nations had been so revoked or

modimodified [sic] that they did not violate the neutral commerce of the United States. (CUSCO, 204 U.S. 364)
d. *It is similarly reasonable for Congress to* have given the States primary responsibility for supervising and ensuring compliance among state sex offenders and to have subjected such offenders to federal criminal liability only when, after SORNA’s enactment, they use the channels of interstate commerce in evading a State’s reach. (CUSCO, 08-1301)

Concerning the other type of SQs, SQ5 reads as a construction in a sequence of *it*, a *modal auxiliary*, the *copula be*, an adjective, and a clause. SQs 6 to 8 differ from SQ5 in that they either contain an element in front of the adjective as in SQ6, or a *for-headed prepositional phrase* between the adjective and the clause as in SQ7, or both as in SQ8. Consider examples (7a–d) that are employed to exemplify the four SQs respectively, in which (7a) is rewritten from (3b) and (7b–d) are taken from the corpus of CUSCO.

(7) SQ5 to SQ8

a. *It would be difficult to* disturb a claim thus sanctioned by time, however unfounded it might have been in its origin.
b. *It would be very difficult to* reconcile a rule allowing the fate of a defendant to turn on the vagaries of particular jurors’ emotional sensitivities with our longstanding recognition that, above all, capital sentencing must be reliable, accurate, and nonarbitrary. (CUSCO, 506 U.S. 461)
c. *It would be anomalous for Congress to* have painstakingly described the Attorney General’s limited authority to deregister a single physician or schedule a single drug, but to have given him, just by implication, authority to declare an entire class of activity outside the course of professional practice and therefore a criminal violation of the CSA. (CUSCO, C546 U.S. 243)
d. Our answer to this is that *it would be utterly impossible for Congress to* define the numerous practices which constitute unfair competition and which are against good morals in trade, for we are beginning to realize that there is a standard of morals in trade or that there ought to be. (CUSCO, 291 U.S. 304)

The variation, *it BE ADJ -ing-clause*, is rather unfavorably used by language users (Collins, 1994), and it is further corroborated by this corpus because this pattern is very rarely attested. Therefore, we do not consider this variation further in the following analysis. By combining all the attested adjectives in the *it BE ADJ clause* construction, the final raw
methods are shown in Table 1, which only tabulates token frequencies of the top 10 adjectives and the total raw frequencies of all types of adjectives.

### Methods

This paper generally adopts a collexeme analysis and a hierarchical clustering analysis. The former is used to identify attracted adjectives in the *it BE ADJ clause* construction and the latter is used to group semantically these attracted adjectives for the purpose of examining the ways that legal opinions unfold by employing this construction.

Collexeme analysis (Gries, 2019; Stefanowitsch & Gries, 2003) is one of the three family members in collostructional analysis, the other two being distinctive collexeme analysis (Gries & Stefanowitsch, 2004) and co-varying collexeme analysis (Stefanowitsch & Gries, 2005). It refers to the analysis of potential lexical items that could occur in the same slot of a grammatical construction (Hilpert, 2014; Stefanowitsch & Gries, 2003). For instance, all potential nouns that could cooccur with the grammatical construction *N waiting to happen* are called collexemes. In order to implement this method efficiently, its operationalization and computation, exemplified by the adjective *necessary*, are expounded with the help of Table 2. In this contingency table, raw frequency of cooccurrences of *necessary* and *it BE ADJ clause* construction (7,235), the row total number of raw frequency of all adjectives occurred in the construction (70,546), the column total number of raw frequency of the adjective *necessary* in all constructions (62,930), and raw occurrences of all adjectives that occurred in the corpus (7,553,881) could be attested directly from the corpus, and the other numbers in this table could obtained by subtraction.

We first computed the expected frequencies (i.e., frequencies of occurrence we would expect if *necessary* and the *it BE ADJ clause* construction are statistically independent) of *necessary* in Table 2, and then measured its association strength with the construction by implementing the log-likelihood ratio test or $G^2$ (Desagulier, 2017). For a contingency table with $i$ rows and $j$ columns, formulas for computing expected frequencies (equation (1)) and $G^2$ (equation (2)) are defined as follows, respectively.

$$E_{ij} = \frac{O_iO_j}{N}.$$  \hspace{1cm} (1)

\[ E_{ij} \] denotes the expected frequency of cell $i, j$; $O_i$ stands for the sum total of the observed frequency in the $i$th row, $O_j$ stands for the sum total of the observed frequency in the $j$th column, and $N$ stands for the sum total of the contingency table. It could be exemplified by computing the expected frequency of *necessary* in the construction, that is, $E_{11}=R_1C_1/N=70,546 \times 62,930/7,553,881=587.7$. By implementing this formula in Table 2, we obtained their corresponding expected frequencies: $E_{11}=587.7$, $E_{12}=69,958.3$, $E_{21}=62,317.3$, and $E_{22}=7,418,017.7$.

$$G^2 = 2 \sum_i \sum_j O_{ij} \log \left( \frac{O_{ij}}{E_{ij}} \right).$$  \hspace{1cm} (2)

In this formula, $O_y$ stands for the observed frequency at row $i$ and column $j$, and $E_y$ stands for the expected frequency at row $i$ and column $j$. With respect to the adjective *necessary* at issue, its $G^2$ equals to 24,426 (cf. equation (3)).

\[
G^2 = 2 \left[ \frac{7235 \times \log 7235}{587.7} \right] + \left[ \frac{63311 \times \log 63311}{69958.3} \right] + \left[ \frac{55695 \times \log 55695}{62317.3} \right] + \left[ \frac{7424640 \times \log 7424640}{7418017.7} \right] = 24426. \hspace{1cm} (3)
\]

A $G^2$ score of 3.8415 is significant at the level of $p < .05$, and a score of 10.8276 is significant at the level of $p < .001$. The magnitude of the $G^2$ score that we obtain (24,426), doubtless, shows the significance of the association between *necessary* and the *it BE ADJ clause* construction. In other words, the adjective *necessary* is extremely significantly attracted by the construction.
Table 3. Observed Frequencies of Adjectives in Both the Corpus and the Construction.

| Adjectives | Frequency in corpus | Frequency in construction |
|------------|---------------------|--------------------------|
| True       | 34,029              | 7,703                    |
| Necessary  | 62,928              | 7,235                    |
| Clear      | 39,914              | 6,138                    |
| Difficult  | 12,198              | 3,958                    |
| Unnecessary| 11,167              | 2,675                    |
| Impossible | 7,709               | 2,573                    |
| Sufficient | 35,490              | 2,172                    |
| Apparent   | 9,801               | 1,995                    |
| Wholesome  | 8                   | 1                        |
| Zealous    | 102                 | 1                        |
| Total      | 2,887,676           | 70,546                   |

Hierarchical clustering analysis is an exploratory data analysis method, which draws on a number of various algorithms for sorting different objects into groups in such a way that the similarity of two objects in the same group is maximal and the similarity that belongs to different groups is minimal (Divjak & Fieller, 2014). In other words, it can be used to identify structures in data and it does this without explaining why that structure exists. Hierarchical clustering analysis therefore is not a regular statistical test based on probability theory; on the contrary, it is a data analytic technique that put different objects into clusters according to well-defined similarity rules. Hierarchical clustering analysis could be completed by the two functions hclust or pvclust in R language.

Identification of Attracted Adjectives to the it BE ADJ clause Construction

Before carrying out the collocxeme analysis to identify the adjectives that are most significantly attracted to the it BE ADJ clause construction, we first arranged the contingency table that contains all adjectives in the construction in the format as shown in Table 3, and then input it into R by performing the function of coll.analysis. The results are shown in Table 4. The first column in Table 4 presents the adjectives that are significantly attracted or repelled by the it BE ADJ clause construction, omitting those adjectives that lies somewhere in the middle. The second and the third columns demonstrate the observed and expected frequencies respectively. The fourth column shows the relation between the adjective and the construction, being either attraction or repulsion. The last column tells us the association strength between the adjective and the construction quantitatively by implementing the $G^2$.

Table 4 shows that the top attracted adjectives to this construction are true, necessary, clear, difficult, impossible, unnecessary, apparent, obvious, evident, possible, and so on. Compared with the findings drawn by Hilpert (2014), both researches identified necessary, difficult, impossible, and possible as the most attracted adjectives to the construction among the top 10 list. However, if extended members among the list are considered, the attracted adjectives that are identified by the two studies differ greatly (there are even six different adjectives among the top 10 list with respect to the attraction to the construction). The possible reasons are threefold. First, the two studies analyzed the data based on different regional Englishes. Hilpert retrieved data from the corpus of BNC which is typical of British English whereas this research collected data from the corpus of CUSCO which is characteristic of American English. Second, the adjectives at issue that are considered by the two researches are considerably different. Hilpert only considered the it’s ADJ to V construction while this research considered not only the it’s ADJ to V construction, but also the other two variations of the it BE ADJ clause construction, particularly the it BE ADJ to-clause construction and the it BE ADJ that-clause construction. In addition, concerning the it’s ADJ to V construction itself, this research also focuses on cases that were ignored by Hilpert. Specifically, we included either the case that the copula BE is modified by a modal auxiliary (cf. example 7a), or the adjective is pre-modified by another element (cf. example 7b), or a for-headed prepositional phrase is interpolated between the adjective and the clause (cf. example 7c), or all factors cooccur in this construction (cf. example 7d). Third, genres or text types included in the two corpora are different. BNC covers such genres as fiction, magazine, newspaper, and academic, whereas CUSCO includes solely legal opinions. It is possible that different adjectives will be attracted by the construction in different genres.

An interesting phenomenon in Table 4 relates to the repulsion. In Hilpert’s research, he did not list the repelled adjectives, and thus we do not know which adjectives are repelled by the construction. However, concerning the adjectives in question in this research, adjectives that are most significantly repelled by the construction include the ones that denote juristic meanings such as constitutional, judicial, criminal, and legal. The reason might be that it is still in the inceptive phase that language users associate these juristic adjectives with the construction while constructing legal opinions. In order to examine whether this phenomenon is testified by other genres, particularly in Hilpert’s research, we retrieved the construction with the four adjectives in BNC, and only identified criminal (three hits) and legal (five hits) in this corpus. This repulsion of adjectives criminal and legal by this construction further supports the assumption that these adjectives are still in their inceptive phases. Additionally, pertaining to the association between such adjectives as constitutional and judicial and the construction, the fact that no cases are identified in BNC provides further evidence that the association is in the inceptive phase. Although we cannot predict whether this association will be a general feature in English or not, we could argue that the
association between constitutional and judicial and the construction is a feature in the genre of English legal opinions.

With respect to the constructional semantics, Hilpert agglomerated these significantly attracted adjectives into such groups as ease, possibility, importance, and advisability. It is obvious that his classification is based on the conceptual meaning of these attracted adjectives, and does not consider the linguistic contextual factors such as the cooccurring lexical items. This is the major task in the following section.

Categorization of Adjectives in it BE ADJ clause Construction

In this section, we first cluster adjectives that are significantly attracted to the it BE ADJ clause construction, and then discuss their meanings in the linguistic context.

Cluster of Significantly Attracted Adjectives

In order to categorize the attracted adjectives by this construction as objectively as possible in terms of their semantic meanings, we implemented a hierarchical clustering analysis. In other words, we use the method to facilitate our groupings of these adjectives that are used in similar linguistic contexts. The underlying assumption is that adjectives could either be modified by adverbs or function as the predicate in the clause. For practical purpose, we selected the top fifty most significantly attracted adjectives to the construction to further investigate their linguistic contexts. There are exactly 289 adjectives that are significantly attracted by the construction. Theoretically, hierarchical clustering analysis could group them in different clusters, but practically, the fact that so many clusters that are presented in one picture makes analysts very difficult to distinguish different clusters. What is worse, some adjectives are overlapped with each other because of the narrow space. Specifically, the top fifty adjectives that are most attracted by the it BE ADJ clause construction are further conducted by examining the total number of lexical types and corresponding frequencies of adverbs that precede them immediately and/or nouns that follow them immediately or precede them within 2-gram. They are exemplified by (8a–c) respectively.

(8) a. The second possible effect is that a reasonably probable entrant has been excluded from the market and a measure of horizontal competition has been lost. (CUSCO, 386 U.S. 568)

b. And certainly no inconsistency results from permitting both rights to be enforced in their respectively appropriate forums. (CUSCO, 415 U.S. 36)

c. The question whether the exclusionary rule’s remedy is appropriate in a particular context has long been regarded as an issue separate from the question whether the Fourth Amendment rights of the party seeking to invoke the rule were violated by police conduct. (CUSCO, 514 U.S. 1)

In retrieving adverbs that immediately precede these adjectives, we set the specific adjective as the net word and their adverbs as the collocates. These hits are grouped by lemmas and sorted by relevance, setting the minimum mutual information (or MI) into three which is generally regarded as the threshold of meaningful association (cf. Church & Hanks, 1990). While retrieving nouns in collocation with these adjectives, the grouping and sorting measures are further employed. The only difference is that we delimited the position of nouns within two words that precede or one word that follows the adjectives. By so doing, hundreds of lexical types of nouns are retrieved. Reducing the data set into a manageable size, we further constrained nouns into those with five co-occurrences with the net words. In this way, a contingency table is further constructed with 1,266 columns (names of collocates) and 50 rows (names of adjectives). We input the txt-filed table into the R Table 4.

| Adjectives | Observed | Expected | Relation | Collostructural strength |
|------------|----------|----------|----------|--------------------------|
| True       | 7,703    | 317.8    | Attraction | 36,906.81                |
| Necessary  | 7,235    | 587.69   | Attraction | 24,426.19                |
| Clear      | 6,138    | 372.76   | Attraction | 24,234.93                |
| Difficult  | 3,958    | 113.92   | Attraction | 21,991.82                |
| Impossible | 2,573    | 71.99    | Attraction | 14,418.75                |
| Unnecessary| 2,675    | 104.29   | Attraction | 12,962.63                |
| Apparent   | 1,819    | 91.53    | Attraction | 7,790.83                 |
| Obvious    | 1,652    | 84.68    | Attraction | 7,007.59                 |
| Constitutional | 25 | 596.76 | Repulsion | 994.75                   |
| Judicial   | 2        | 487.25   | Repulsion | 956.44                   |
| Criminal   | 2        | 408.85   | Repulsion | 798.61                   |
| Legal      | 18       | 449.59   | Repulsion | 753.91                   |
language and performs the function of `hclust`, and the output cluster dendrogram is presented in Figure 1. The dendrogram clustered these attracted adjectives into different semantic groups, which are highlighted in red and will be discussed respectively in the following sub-section.

**Semantic Categorization of Significantly Attracted Adjectives**

By implementing a hierarchical clustering analysis to adjectives in *it* BE ADJ clause construction, it is shown that at least seven semantic groups are included, which will be briefly discussed based on the underlying reasons that they are clustered and the adjective(s) that seem(s) to be anomalous (but actually plausible) to this grouping.

**Group 1: **Adjectives expressing importance, necessity, and possibility.** The first group of adjectives that are clustered together include difficult, important, essential, necessary, possible, likely, clear, and plain. Except for adjectives difficult, clear, and plain, the others by and large denote modal meanings. They are used in the *it* BE ADJ clause construction as explicit objective modal expressions to realize interpersonal metaphors of modality in Hallidayan sense (Halliday, 1985, 1994). According to Halliday, interpersonal metaphors of modality refer to the fact that explicit objective modal expressions (cf. example 9a which is rewritten from 4a) are used as if they were implicit objective modal expressions (cf. example 9b). Considering legal opinions, these modal adjectives in this construction are used by juristic experts to entertain some other voices as shown in (9a) and/or to express the obligation of performing an activity that a proposition or state of affairs denotes (see example 10). The attitude that the proposition others may be presumed, and less than positive testimony may establish facts expresses in (9a) is “heteroglossic” (Bakhtin, 1981). In other words, the writer, that is, the juristic expert, also invites some other different opinions toward this proposition in order not to fully take the responsibility. In example (10), the action of purchasing the land and paying the money for it without knowledge of this previous deed is an obligation.

(9) a. From the establishment of some facts, *it is possible* that others may be presumed, and less than positive testimony may establish facts.

(10) *It is necessary* that he should have purchased the land and paid the money for it without knowledge of this previous deed. (CUSCO, 75 U.S. 27)

We further analyzed the reasons why adjectives difficult, clear, and plain are clustered in this group. A scrutinization of the collocates with the adjectives in this group demonstrates that the collocated words between these modal adjectives and difficult are such adverbs of degree as equally and sufficiently, and nouns as task. The collocated words between modal adjectives and clear and plain are generally adverbs such as equally, reasonably, and sufficiently, and nouns such as implication and inconsistency. In other words, the

![Cluster Dendrogram](Image)
underlying reason that the three adjectives are incorporated into group 1 is that they, together with these modal adjectives, are modified by similar adverbs like equally, reasonably, and sufficiently, and preceded or followed by nouns such as task, implication, and inconsistency. With respect to these adjectives in the it be ADJ clause construction, they are, in varying degrees, employed by the writer of legal opinions to express the necessity or importance of performing a certain activity. Consider examples (3b) and (11a and 11b). In (3b), the writer of legal opinions uses the construction to implicate that it is of importance or necessity not to disturb a claim thus sanctioned by time, however unfounded it might have been in its origin. In the same vein, it is clear that in (11a) and it is plain that in (11b) also implicate the importance of performing the activity that the proposition denotes.

(11)  a. It is clear that the present transaction does not fall within the prohibition of dealing or trading in the preceding part of the same article. (CUSCO, 34 U.S. 378)
    b. It is plain that the paper offered, is not the best evidence of which the nature of the case admits. (CUSCO, 2 U.S. 230)

Group 2: Adjectives expressing appropriateness, reasonability, unreality, etc. Adjectives in group 2 incorporate conceivable, arguable, proper, appropriate, sufficient, reasonable, significant, unnecessary, unreasonable, unrealistic, competent, immaterial, true, and open. This type of semantic clustering is confirmed by such adverbs as altogether, certainly, constitutionally, entirely, equally, hardly, legally, and reasonably, and nouns as interpretation, intrusion, and possibility. Writers of legal opinions use these adjectives in the construction to express whether the performance of an activity denoted by the proposition is appropriate, reasonable, or unrealistic. Adjectives true and open seem to be anomalous members in this group. We once again turned to the two adjectives in the construction and found that they can also express the appropriateness and/or the reasonability of the activity that the proposition denotes. Consider examples (12a and 12b). Construction with the adjective true in (12a) and the adjective open in (12b) implicate that the activities expressed by the propositions are appropriate and/or reasonable.

(12)  a. It is true that the acts, doings, and declarations of individual members of the corporation, unsanctioned by the body, are not binding upon it. (CUSCO, 25 U.S. 64)
    b. The Commission’s order was presumptively valid, as the state court held, but it was open to attack in this action under the state statute. (CUSCO, 304 U.S. 224)

Group 3: Adjectives expressing impracticability and irrelevance. Adjectives significantly attracted to the construction in group 3 include impossible, impracticable, probable, inappropriate, and irrelevant. Adjectives in this group express generally negative meanings of practicability and relevance of carrying out the proposition that the subordinate clause denotes. Semantically, they are clustered on the ground that they are modified by three types of adverbs (no single noun is identified to be proximal to these adjectives): adverbs expressing the meaning of high degrees such as completely, entirely, totally, utterly, and wholly, the ones expressing the meaning of obviousness such as clearly, manifestly, and obviously, and finally the one expressing the meaning of legality like legally. The final type of adverbs is plausible and typical of the text type in question because the ideational meaning of legal opinions certainly will be associated closely with the meaning of legality. The adjectives that seem to be anomalous to this group are probable and impossible, which, in terms of the modal meaning of probability, should have been clustered into the first group. However, the reason that they are clustered in group 3, besides the linguistic contextual factors mentioned above, is that they denote the modal meaning of high and medium values, which could be also employed to express the meaning of impracticality. This is instantiated by examples (13a and 13b). Both examples in (13a and 13b) could be paraphrased as to read the testimony of the principal surveyor is impracticable and that the supreme court of the territory yielded to these contentions is impracticable respectively.

(13)  a. It is impossible to read the testimony of the principal surveyor. (CUSCO, 20 U.S. 122)
    b. It is probable that the supreme court of the territory yielded to these contentions. (CUSCO, 215 U.S. 398)

Group 4: Adjectives expressing undeniability and axiomaticity. Adjectives in group 4 include idle, undeniable, noteworthy, inconceivable, undisputed, and axiomatic. What renders the agglomeration of these adjectives is that they are usually modified by such adverbs as almost and equally. The preceding and/or following nouns do not seem to be decisive factors to cluster these adjectives. With respect to the construction in question, it mainly attracts these adjectives to articulate the fact that propositions or statements in legal opinions are so undeniable and/or axiomatic that the authority of laws is foregrounded.

Group 5: Adjectives expressing obviousness. Adjectives in group 5 incorporate evident, manifest, apparent, and obvious. From the vantage point of sense relations, the four adjectives are more or less synonyms; from the perspective of linguistic contextual factors, they are mostly modified by such adverbs as immediately, particularly, perfectly, plainly, and sufficiently, and such nouns as anger, inconsistency, injustice, intent, intention, and purpose. These adjectives are attracted by the construction to express the obviousness of the state of affairs of the proposition.
Group 6: Adjectives expressing dubiety, desirability, and ease. Adjectives in group 6 include doubtful, repugnant, hard, easy, interesting, unlikely, and desirable. References to their collocated words tell us that adverbs are mostly equally, especially, extremely, highly, and particularly. Once again, there is not a single noun that is preferred simultaneously by most adjectives in this group.

Group 7: Adjectives Expressing Improbability and Anomalousness. Adjectives that are significantly attracted to the construction in the last group incorporate improbable, ironic, surprising, absurd, anomalous, and strange. The underlying reason for the clustering of these adjectives is that they are modified by such adverbs as altogether, equally, especially, indeed, particularly, and truly, and such nouns as conclusion and consequence. Concerning these adjectives in this construction, they are able to posit the proposition in a negative semantic prosody (cf. Wei, 2002). Pertaining to the superficially mismatching adjective, surprising seems to be an outlier, but instances from the corpus again prove that it is not the case. This is exemplified by (14). The adjective surprising in the construction in example (14) actually denotes a negative semantic prosody, and it is therefore clustered, together with other adjectives expressing negative meanings, in this group.

(14) Because the facts of this case are so unusual, it is surprising that the Court considers it appropriate to grant certiorari and address the merits. (CUSCO, 454 U.S. 14)

Accordingly, the it BE ADJ clause construction extremely significantly attracts adjectives that express the seven types of meanings clustered above. It could also be identified that negative adjectives (cf. groups 3, 4, and 7) are preferably employed by the writer of legal opinions for the purpose to persuade people to behave properly and not to violate the plausibility that the proposition expresses. In addition, adjectives that express modal meaning are scattered in all types of groups with the dominance of group 1 for the purpose of entertaining different voices concerning the proposition in the construction.

Comparison With Hilpert’s Semantic Categorization

With respect to the predicative complement in the third slot of the it BE ADJ clause construction, we do not compare the findings drawn from this research with those from Collins’ (1994) in that Collins considered not only adjectives but also nominal and prepositional phrases as the complement. While compared with Hilpert’s (2014) semantic categorization which was based on different scales, we identified some similarities and differences. The two researches are similar in that groups 1 and 6 in this research clustered adjectives which are also considered by Hilpert’s categorizations of “importance” and “ease,” respectively. The differences are threefold. First, we identified seven groups of adjectives that are significantly attracted by the construction while Hilpert identified only four types of different scales. The possible reason might be that Hilpert considered only adjectives in the it’s ADJ to V construction whereas we considered adjectives in the it BE ADJ clause construction which covers not only the variation of to-clause construction but also that of that-clause construction. In addition, we also considered such cases as an element occurring immediately before the adjectives, a modal auxiliary modifying the copula BE, and a for-headed prepositional phrase interpolating between the adjective and the clause. Second, adjectives that are clustered in a single group are more than those in Hilpert’s categorizations. The underlying reason is that the extremely significantly attracted adjectives to the construction considered in this research are more than those in Hilpert’s. We analyzed the top fifty most attracted adjectives while Hilpert only considered the top 10 adjectives. Third, adjectives in Hilpert’s type of “possibility” are separately considered in groups 1 and 3 in this research; and those in his type of “advisability” are not identified in this research. This might be caused by the fact that Hilpert categorized the attracted adjectives based on their conceptual meanings whereas we clustered these adjectives in terms of their cooccurring adverbs and nouns in the linguistic context. Specifically, his research is more or less subjectively oriented while this one is generally objectively conducted; in other words, this research is at least more objectively carried out than Hilpert’s. This might also be caused by the different choices of genres or text types; Hilpert’s semantic subdivision of attracted adjectives is based on the data obtained from different genres in BNC while we obtained data solely from legal opinions in CUSCO. In addition, genres in BNC are typical of British English while the genre of legal opinions in CUSCO is typical of American English.

Conclusion

This paper centers on the predicate adjectives in the it BE ADJ clause construction. With respect to the American writers’ preference in legal opinions, the collexeme analysis reveals that they favor such adjectives in the top list as true, necessary, clear, etc. in the construction. Semantically, linguistic contextual factors such as preceding and following nouns and/or adverbs objectively cluster the most attracted adjectives into seven different groups, which differs greatly from Hilpert’s classification in that he identified only four types in terms of different scales bases on their conceptual meanings. This research is significant in three respects. Methodologically, it objectively clustered the adjectives that are significantly attracted to the construction by implementing a hierarchical cluster analysis. Theoretically, it extended
the present semantic categorization of these attracted adjectives in the construction into seven different groups. Practically, it will provide writers of legal opinions with a number of potential choices in terms of the seven different semantic categorizations while constructing their suggestions to others.

Although this research is conducted in a more fine-grained way than Hilpert’s, the findings need to be further testified because we considered only the genre of legal opinions in American English. Therefore, we suggest future studies consider more genres in different regionalized Englishes.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Jiangping Zhou https://orcid.org/0000-0003-3305-2651

References
Bakhtin, M. M. (1981). The dialogic imagination (translated by Emerson & M Holquist.). University of Texas Press.
Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22–29.
Collins, P. (1994). Extraposition in English. *Functions of Language, 1*(1), 7–24.
Desagulier, G. (2017). *Corpus linguistics and statistics with R: Introduction to quantitative methods in linguistics*. Springer Nature.
Divjak, D., & Fieller, N. (2014). Cluster analysis: Finding structure in linguistic data. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 405–441). John Benjamins.
Gries, S. (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.
Gries, S., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on ’alternations’. *International Journal of Corpus Linguistics, 9*, 97–129.
Halliday, M. A. K. (1985). *An introduction to functional grammar* (1st ed.). Edward Arnold.
Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). Edward Arnold.
Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). Hodder Arnold.
Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday’s introduction to functional grammar* (4th ed.). Routledge.
Hao, J. (2020). Nominalizations in scientific English: A tristratal perspective. *Functions of language, 27*(2), 143–173.
He, Q. (2019). *A corpus-based approach to clause combining in English from the systemic functional perspective*. Springer.
He, Q. (2020). A corpus-based study of interpersonal metaphors of modality in English. *Studia Neophilologica, 92*, 1–22.
Herriman, J. (2000). The functions of extraposition in English texts. *Functions of Language, 7*(2), 203–230.
Hilpert, M. (2014). Collostructional analysis: Measuring associations between constructions and lexical elements. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 391–404). John Benjamins.
Hsiao, H. S., & Mahastuti, L. (2020). A collostructional analysis of ditransitive constructions in Mandarin. In J. Hong, Y. Zhang, & P. Liu (eds.), *Chinese lexical semantics, CLSW 2019* (pp. 37–51). Lecture Notes in Computer Science. Springer.
Kaltenböck, G. (2003). On the syntactic and semantic status of anticipatory it. *English Language and Linguistics, 7*(2), 235–255.
Kaltenböck, G. (2005). *It*-extraposition in English: A functional view. *International Journal of Corpus Linguistics, 10*(2), 119–159.
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.
Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language science Press.
Stefanowitsch, A., & Gries, S. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209–243.
Stefanowitsch, A., & Gries, S. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory, 1*(1), 1–43.
Wei, N. (2002). An approach to the study of semantic prosody. *Foreign Language Teaching and Research, 34*(4), 300–307.
Zhang, Z. B. (2017). *A new English grammar coursebook*. Shanghai Foreign Language Education Press.