English Language Teaching and Linguistics Studies
ISSN 2640-9836 (Print) ISSN 2640-9844 (Online)
Vol. 6, No. 5, 2024
www.scholink.org/ojs/index.php/eltls

Original Paper

A Corpus-based Behavioral Profile Study of Synonymous Expressions Meaning “to Start”

Kexin Shi[1]

[1] School of Foreign Studies, University of Science and Technology Beijing, Beijing, China

Received: July 29, 2024 Accepted: September 07, 2024 Online Published: September 29, 2024
doi:10.22158/eltls.v6n5p180 URL: http://dx.doi.org/10.22158/eltls.v6n5p180

Abstract

Synonymy is an extremely significant component in language learning, and the usage of synonymy is rather flexible and difficult to distinguish. In this research, the author employs a corpus-based behavioral profile approach to study the usage of a set of English synonymous expressions meaning “to start”, namely, start, begin, commence, launch, set out. The author first searches in American Contemporary English Corpus to explore the overall usage frequency and specific genre of each synonymous expression, then randomly selects 500 sentences, and makes manual annotation with 12 ID tag and 46 ID tag levels. Later, hierarchical clustering analysis and correspondence analysis have been employed to examine the semantic differences and usage structures of this set of expressions.

The results shows that: ① The overall usage frequency varies greatly—from the highest to the lowest are start, begin, launch, set out, commence. ② They are used in various genres. Start is mainly used in informal styles including blogs, while the other four are mainly used in formal styles, including fictions and magazines. ③ Begin and set out have the greatest semantic similarity; start, begin, and set out form one semantic branch, and commence and launch form another, and there is a huge difference between these two branches. ④ There are great differences between the five expressions and their morphological, syntactic and semantic changes.

Quantitative analysis method has been used to reveal the different semantic features and usage of a set of synonymous expressions, which is conducive to both promote the study of synonymous relations, and verify the effectiveness and applicability of the behavioral profile approach in semantic study.

Keywords

Synonymy, synonymous expressions; semantics, BP analysis, “to start”

1. Introduction

Synonym was directly originated from late Latin word “synonymum”, and from Greek word "synonymon", referring the the word that has the same sense as another, or denotes the same idea. Synonymy is an important, but also intricate linguistic phenomenon (Divjak & Gries, 2006; Taylor, 2003). While synonyms always basically express the same meanings and ideas, they often do so in different forms and/or different contexts, and/or from different perspectives. Therefore, slight differences exist within sets of synonymous expressions. Synonymy is a demanding and significant lexical category for the reason that it plays an essential role to express shades of meaning to help us convey precisely desired ideas for exchanges and in the meanwhile, avoid unwanted mistakes (Edmonds & Hirst, 2002; Hatch & Brown, 1995).

Recent years has witnessed the rapid development of corpus-based study in linguistics, providing a new approach to the scientific and statistical research on semantics. Promoted by Hanks (1996), corpus-based behavioral profile approach (BP analysis for short) has been recognized as a quantitative method to reveal the lexical usage and lexical characteristics of polysemy, antonym and synonymy for particular, in terms of features in morphology, syntax, semantics, and other aspects. Wu and Liu (2020) systematically introduced BP analysis in China, including the theoretical basis, general steps, its application in different fields of semantic study, its advantages, disadvantages and future development. It has been proved a mainstream method in quantitative research on lexical semantics in the foreign countries, however, it has not been widely used in China.

Based on the previous mentioned backgrounds, the author intends to use BP analysis to figure out the semantic distribution of a set of English synonymous expressions meaning "to start" (*start*, *begin*, *launch*, *commence, set out*). In the meanwhile, the author employs two statistical methods, namely, hierarchical cluster analysis and correspondence analysis, to precisely visualize their general usage situation, semantic relationship, and semantic distributions. This research should be reckoned as a useful pattern to delve into the semantic study and promote the corpus-based quantitative study as well.

2. Literature Review

In this section, significant papers and thesis pertaining to synonymous words, synonymous verbs, and expressions meaning "to start" will be presented and illustrated, so as to give a firm foundation for further research and analysis.

2.1 Research on Synonymous Words

Synonyms account for a large proportion in English, one of the most widely used expressions in the world. According to statistics, the number of synonyms in English takes a proportion for more than 60% of the total vocabulary. They form a huge system with similarities in morphology, semantics, structure or pragmatics. Correctly learning and using synonyms is important to improve English writing, reading, conversation and other skills.

A basic approach to analyze synonymy has been put forward, in which three methods have been introduced, namely, from the perspectives of semantics, stylistics, and corpus-based linguistics. This paper is a brief summary of the previous studies, sorting out the reference materials, and setting a firm foundation for future researches. However, traditional methods for studying synonymy mostly rely on intuition and personal experience, and adopt introspective qualitative methods to analyze the meaning in detail.

Along with the development of technology and the appearance of new tools, the research on synonymous expressions has been booming in recent decades. Corpus-based research on synonymy is flourishing. Zhang and Liu (2005) have proposed a general corpus-based research approach, within which, synonyms can be differentiated with reference to: ① their distributions among different registers; ② their significant collocates, and the MI value and Z score between synonyms and their collocates; ③ their collocational behaviors and semantic prosodies with regard to certain colligational frameworks. Besides, many scholars have began to make the best use of corpora to analyze synonymous words, for instance, Lu (2010) has figured the differences in collocation and prosody in English texts written by Chinese English learners, Yang (2007) has used Sketch Engine to explore the grammatical and structural composition of synonyms, and so on.

2.2 Research on Synonymous Verbs

Based on the former research framework, research on synonymous verbs has been conducted smoothly. In China, Zhang and Liu (2006) have researched on the usage of English *happen*, *occur*, and Chinese "发生" (*fasheng*). Through the statistics of their word frequency distribution in many registers in the corpus, they have observed the linguistic features such as different collocation relations, class connections and semantic prosody in different search lines to clarify the differences between them. This method can play an effective complementary role to the traditional vocabulary discrimination.

The author of this paper has adopted some of the parameters from Zhang and Liu's research, and furthermore, add some more parameters (ID tag levels in the latter section) into research, so as to present a more comprehensive description of the semantic features of sets of synonymous expressions.

2.3 Research on Expressions Meaning "to Start"

There exists a great bunch of synonymous verbs and verb phrases meaning "to start", and in this research, based on the previous published papers, the author has decided to research five expressions, namely, *start*, *begin*, *launch*, *commence*, and *set out.*

In foreign countries, this type of research begins at an earlier time, and many scholars have presented their abundant research achievements. How certain words have been translated into another language has been verified as a good way to examine the similarities and differences of sets of synonymous words. Nida (1964) believes that language is a multi-functional communicator tool. Translation is communication, that is, it is a way to exchange information and message. However, communication is not unilateral and if there is no response from the other party, there is no actual communication. The ultimate goal is the author' response to the translation and the reader's response to the original texts are basically the same. Therefore, analyzing semantic distance in original and translated texts is a practical method, based on the commonsense understanding of the experiential world. Susan and Thomas (2010) figure out the translation options—the way lexical constructions containing *begin* and *start* have been rendered into Norwegian, a language which contains the cognate verbs *begynne* and *starte,* in an English-Norwegian Parallel Translation Corpora, so as to discover the extent and nature of the similarities and differences between these two words and constructions. Vandevoorde, Lefever, Plevoets, and De Sutter (2017) have conducted a research in the semantic difference among inchoative verbs, which sometimes are called inceptive verbs, showing a process of beginning or becoming in Dutch. They have collected the inchoativity in both original Dutch and the translated Dutch, and then used the method of semantic mirror to analyze them.

In China, there have also emerged some research outcomes. Pang and Yang (2012), taking synonymous sets *begin*, *start* and *commence* as an example, examine how differently synonymous words are primed from the perspective of Hoey's Lexical Priming Model. The two scholars mainly concentrate on the frequency, morphological changes, lexical collocation, grammatical structure, transitivity processes, and semantic meaning (positive/negative) of these three synonymous words, thus presenting the normal usage of them. However, frequency and percentage are not clear enough to make a dynamic and comprehensive description. Wang (2016) has chosen a different perspective to analyze synonymous verbs. He, based on native speakers' corpus LOB and non-native speakers' corpus CLEC, has investigated the usage of English verbs *begin* and *start* used by Chinese EFL learners. Through the calculation of the overall usage frequency and grammatical structures of these two verbs, the researcher has generated the following findings: Chinese EFL learners overuse "begin + to infinitive"; compared with native speakers, Chinese students underuse both *begin* and *start* as an intransitive verb. Furthermore, by using different Corpora—SWECCL and COCA, Wang (2016) has also picked up this education-oriented corpora to delve into the research of *begin* and *start.*

However, based on these previous published papers, it can be seen that there is a lack of relevant researches on expressions meaning "to start". In addition, the research method, research parameters, analyzing methods, and tools used are all expected to be enriched, towards a more scientific and comprehensive analysis.

3. Corpus-based BP Approach to the Study of Synonymy

Hanks (1996) first uses the concept of Behavioral Profile (BP) to describe the usage pattern of vocabulary, while Firth, is the first to advocate the emphasize on the usage mode of vocabulary in the 1950s. Firth (1957) has an assertion that "you shall know a word by the company it keeps". Later, Halliday and Sinclair (Krishnamurthy, 2006) further develops the study of lexical semantics. The research of them has deeply influenced modern lexical semantics research and encouraged the usage of empirical methods to establish the close relationship between lexical meaning and its use patterns, which has laid the foundation for the corpus-based approach to lexical behavioral profile. Gries (2010) has made a detailed summary and description of the lexical behavioral profile approach and proposed four steps when using it. It is not only descriptive, but also explanative in the interpretation of theoretical implications found in the research, which conforms to the theory of cognitive linguistics. Its inner logic is the usage-based method of language acquisition, representation and processing.

Thanks to the advancements in corpus linguistics, there have been increasingly more corpus-based behavioral profile (BP) studies on synonymous verbs (Beraz & Grice, 2009; Divjak, 2006; Divjak & Gries, 2006), synonymous adjectives (Gries & Otani, 2010; Liu, 2010), synonymous adverbs (Liu & Espino, 2012; Wu, Liu, & Liu, 2021), synonymous construction (Klavan, 2014), and so on. All of these researches have proved the effectiveness and practicality of using BP analysis in semantic research.

4. Methodology

This part is about methodology, in which three research questions and two main research methods will be exhibited. In the meanwhile, the concrete research design, and the four steps in conducting BP analysis in this research, will be shown one by one.

4.1 Research Questions

Through conducting this research, the author wants to answer the following three questions related to synonymous expressions meaning “to start”:

① What is the general usage of these synonymous expressions?

② What is the semantic distance of these five expressions?

③ To what extent, do there exists exclusive semantic features to each of these expressions?

4.2 Research Methods

In this research, both quantitative and qualitative analysis methods have been employed by the author. In terms of quantitative method, the author has first randomly downloaded sentences for each synonymous expression from Corpus of Contemporary American English (COCA for short), and built a small-sized corpus based on the former language materials, which becomes the direct research contents. In terms of qualitative method, the author has given detailed analysis on the same and different semantic features and provided the underlying explanations with examples.

4.3 Research Design

This research is conducted by using the corpus-based BP approach, and Gries (2010) has summarized the basic vocabulary behavior features and proposed four steps when using this method.

Step 1: Data collection. Inputting the capital form of each expression and choosing its verb form in both the section of “List” and “Chart” in Frequency in COCA (*START*, *BEGIN*, *LAUNCH*, *COMMENCE*, *SET OUT*). And then randomly download 100 sentences for each synonymous expression, 500 sentences in total. In this process, extracted sentences that are irrelevant and unsuitable to this research have been omitted. Finally, 500 sentences compose as a small-sized corpus, which will be annotated and analyzed.

Step 2: Manual annotation. It is one of the most important step to make manual annotation according to the expressions’ features and properties in usage. Atkins (1987) has claimed that these properties are referred to as ID tags and include, but are not limited to, the morphological, syntactic, semantic, and other characteristics. In this research, the author has confirmed three basic types of ID tags (see Table

1).

Table 1. Manual Annotation of Type of ID Tag, ID Tag, and ID Tag Levels

Type of ID Tag	ID Tag	ID Tag Levels
morphological	tense	present, past, future
	aspect	simple, perfective, progressive
	mode	indicative, imperative, subjunctive
	voice	active, passive
	negation	affirmative, negative
	person	first, second, third
syntactic	sentence type	declarative, exclamative, imperative, interrogative
	clause type	main, subordinate/dependent
	subject	pronoun, noun/noun phrase, sentence, zero
	object	pronoun, noun/noun phrase, verb/preposition phrase, sentence, zero
semantic	prosody	positive, negative, neutral
	following components	to do, doing, noun/noun phrase, collocation, clause, zero, pre-with, pre-time, pre-place, pre-by, pre-in, pre-for

Verb is characteristically the grammatical center of a predicate and expresses an act, occurrence, or mode of being, etc. And most of the changes are represented in verb, therefore, in terms of morphological tag, tense, aspect, mode, voice, negation, person have been analyzed. Then, since verb plays a significant role in constructing a sentence, the author has also delved into the research of sentence type, clause type, subject, and object. Besides, prosody (positive, negative, and neutral) and following components have been detected. All in all, 12 ID tags, 46 ID tag levels have been confirmed The overall frequency of each ID tag level of different words has been placed at the very last of this paper (see Appendix 1).

Step 3: Create a co-occurrence frequency table. Based on the manual annotation and the calculation of frequency of occurring of each ID tag level, the author makes a co-occurrence frequency table (see Table 2).

Table 2. Co-occurrence Frequency (in part)

Type of ID tag	ID Tag	ID Tag Levels	start	begin	launch	commence	set out
Morphological	tense	present	0.540	0.490	0.650	0.680	0.600

		past	0.450	0.490	0.320	0.290	0.400
		future	0.010	0.020	0.030	0.030	0.000
	aspect	simple	0.920	0.840	0.970	0.820	0.820
		perfective	0.020	0.070	0.020	0.070	0.120
		progressive	0.060	0.090	0.010	0.110	0.060
	mode	indicative	0.880	0.980	1.000	0.910	0.980
		imperative	0.090	0.020	0.000	0.090	0.020
		subjunctive	0.030	0.000	0.000	0.000	0.000
	voice	active	0.940	0.990	0.890	0.920	0.980
		passive	0.060	0.010	0.110	0.080	0.020
	negation	affirmative	1.000	0.950	1.000	1.000	0.930
		negative	0.000	0.050	0.000	0.000	0.070
	person	first	0.300	0.130	0.080	0.090	0.230
		second	0.120	0.070	0.010	0.110	0.080
		third	0.580	0.800	0.910	0.800	0.690
...	...	...	...	...	...	...	...

Step 4: Conduct statistical analysis. At last, two statistical analysis have been conducted, so as to present scientific and visualized graphs to make further analysis. Hierarchical clustering analysis is helpful to present the overall similarities and differences among verbs; while correspondence analysis can benefit researchers to figure out the co-occurrence between each verb and all semantic characteristics.

5. Results and Discussion

Section 5 is the main part of this paper, in which general usage, similarities and differences in semantic distribution of these five synonymous expressions, and discussion from the perspective of frame semantics will be presented in a detailed way.

5.1 General Usage

To understand the overall usage of these five synonymous expressions, it is needed to know their usage frequency in the first stage. The author has retrieved and recorded the frequency of each expression in COCA (see Table 3). From Table 3, it can be seen that the usage frequency varies from one word to another, showing great disparity in number.

Table 3. The Frequency of Five Synonymous Expressions

Expression	START	BEGIN	LAUNCH	COMMENCE	SET OUT
Frequency	643473	421901	69389	4677	14326

From the highest to lowest frequency are *start*, *begin*, *launch*, *set out*, *commence*. *Start* and *begin* are the two most frequently used words; while, *commence* is the least used one. All forms of *start* have been used in daily life for more than 100 times than the word *commence*.

Turning from the overall frequency of all expressions, we come to the specific context in which they are used, based on the eight genres COCA has clarified. From Table 4, we can see that the two colored parts of START are located differently when compared to those of the other four words. It has been used for 25053 times, accounting for 3.89% of all START in academic genre, which is the lowest percentage among all 8 genres, and it has been used for 97362 times, accounting for 15.13% of all START in blog genre, which is the highest percentage. Therefore, we can make a brief conclusion that START tends to be an informal representation employed in oral and daily English.

Besides, BEGIN, LAUNCH, COMMENCE, and SET OUT have the lowest usage frequency and percentage either in TV/Media genre or in spoken English, and own the highest frequency and percentage in the genre of website, fiction, and magazine, all of which can be reckoned as regular and solemn styles in English. Therefore, these four synonymous expressions are more likely to be formal usages when people want to convey the meaning of "to start".

Table 4. Genre Distribution of Five Synonymous Expressions

Genre	START		BEGIN		LAUNCH		COMMENCE		SET OUT	
	Freq	Per	Freq	Per	Freq	Per	Freq	Per	Freq	Per
BLOG	97362	15.13%	36224	8.59%	10711	15.43%	471	10.08%	1389	9.70%
WEB	85358	13.27%	48322	11.45%	9994	14.40%	1013	21.67%	2093	14.61%
TV/M	87693	13.63%	18155	4.30%	3701	5.33%	691	14.78%	563	3.93%
SPOK	95331	14.82%	58206	13.80%	7284	10.50%	148	3.17%	1416	9.88%
FIC	83509	12.98%	72122	17.09%	4145	5.97%	703	15.04%	2239	15.63%
MAG	82395	12.81%	69883	16.56%	16797	24.20%	496	10.61%	2951	20.60%
NEWS	86739	13.48%	65336	15.49%	11580	16.69%	292	6.25%	1713	11.96%
ACAD	25053	3.89%	53644	12.72%	5184	7.47%	860	18.40%	1962	13.70%
ALL	64344	100%	42189	100%	6939	100%	4674	100%	1432	100%

0	2	6	6

“Freq” is short for “Frequency”, and “Per” is short for “Percentage”. Painted in light green represents the lowest frequency and percentage of one expression, while painted in light red represents the highest.

5.2 Similarities and Differences in Semantic Distribution

A hierarchical cluster dendrogram has been made by clustering all semantic features of these five synonymous expressions (see Figure 1). It can reveal the general similarities and differences among clustered members, but cannot present the specific characteristics that lead to such differences and similarities.

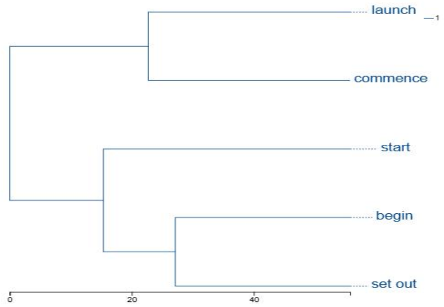

Figure 1. Hierarchical Cluster Dendrogram

There are two main fundamental principles in interpreting the hierarchical clustering graph: ① The similarity between the members that are clustered at an earliest time is higher than that of the later clustered ones, and there is more similarity in semantic; ② If the distance between the clustered groups is far, their semantic distance is relatively far, and they are more distinctive in meaning.

We can analyse Figure 1 from the right side to the left side in two aspects. For one thing, the earlier the words are clustered in a sub-branch, the closer their semantic distance is. In this research, *begin* and *set out* are firstly combined into a single branch, indicating that they have a similar semantic field and semantic features when compared with the other synonymous words. Then *begin* and *set out* are clustered with the word *start*, presenting the semantic distance among them three is comparatively close. Likewise, *begin* and *set out* are clustered before *launch* and *commence*, showing that the semantic similarity between *begin* and *set out* is higher than that of *launch* and *commence.*

For another, the whole semantic field of “to start” can be divided into two main branches, which are clustered by *start*, *begin*, *set out*, and clustered by *launch*, *commence.* The lexical components within one branch are more closely related to each than those clustered in another branch. Therefore, *start*, *begin*, *set out* are closely connected to each other, and *launch, commence* are closely interrelated to each other; while, this two branches have a relatively far distance.

5.3 Semantic Features and Semantic Distribution

Hierarchical clustering analysis has grouped the five synonymous expressions that represent the meaning of "to start" according to the semantic similarity. The result is that the usage patterns of *begin* and *set out* are the most similar, and the semantics of *launch* and *commence* are different from the other three. In order to present the specific differences in the usage of these five expressions, the author further adopts corresponding analysis method. It is a multi-variable exploratory spatial reduction technology analysis (Glynn, 2014), which is similar to cluster analysis in many aspects, but it can exhibit the degree and type of relevance among all factors in a two-dimensional planar plot. Secondly, it can visualize the distance between various language factors, which is a very intuitive and effective method (Glynn, 2014).

The author uses SPSS 26.0 to make a correspondence analysis (see Figure 2). Five different synonymous expressions (*start, begin, launch, commence, set out*) have been named as 50, 60, 70, 80, 90, which are presented in blue; while the 46 ID tag levels have named from 1 to 46, presented in color of pink.

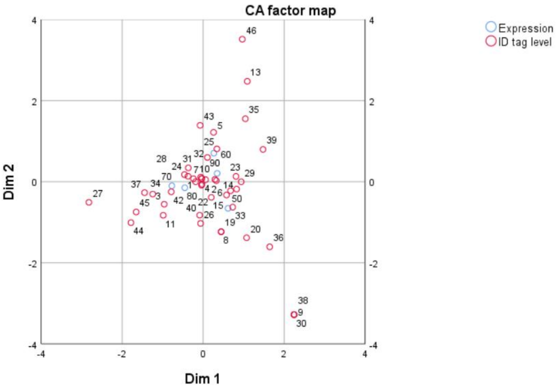

Figure 2. Correspondence Analysis

According to Figure 2, it can be clearly seen that these five expressions are distributed in three different quadrants, of which *launch* and *commence* are located in the lower left quadrant, with the closest distance, indicating that the similarity between them is the greatest; while *start* is located in the lower right quadrant, and *begin* and *set out* are distributed in the upper right quadrant. Compared with *launch* and *commence, start* is located in a place more close to *begin* and *set out*, that is to say, they are relatively near. This result can also verify the conclusion proven by hierarchical clustering analysis in Figure 1.

Specifically, the factors that are significantly related to *start* mainly include first person, second person, subordinate/dependent, and negative prosody. *Begin* and *set out* are located in the same quadrant, and the most significant semantic features are the aspect of progressive, active voice, sentence as subject, and positive prosody. *Launch* and *commence* are located in the same quadrant, and the most significant semantic features are the tense of present, noun/noun phrase as subject, following time-related components. However, the semantic features in the middle, such as the tense of past, indicative, neutral prosody, followed by with-structure, are all related to the use of these five synonyms, indicating that these features are least affected by other features and are the most commonly used ones. The following is a detailed analysis of the results for each semantic feature.

6. Discussion

In this section, the author will analyze some of the most significant ID tag and ID tag levels that are related to the five synonymous expressions. Besides, examples will be given, so as to give a clear description and analysis.

6.1 Morphological Aspects

Most of the six ID tags in morphological aspects (number 1-16) are located in the center of Figure 2. The tense of present, past, future, the mode of indicative, imperative, subjunctive are used in a balanced way in the five synonymous expression. However, voice and negation are greatly different. Passive voice is located close to the words *commence* and *launch*, which means that when *commence* and *launch* are used, people are more likely to used it in a passive way.

Example 1: The last year of high school before they would all be launched into adulthood where a whole new horizon awaited them.

Example 2: Control measures were launched to halt rabies spread in skunks and limit the potential for human exposures.

From Example 1 and 2, we can see that passive voice has been used, in order to objectively describe the event happened. This is a way to emphasize the process, and weaken the existence of entity that initiates the action. The subject "students" in Example 1 has been replaced by pronoun *they*, while the true agent has been omitted. And the true agent in Example 2 has not be presented, which may be not important to the description of the events.

In addition, negative is located in the upper right quadrant, near to the words *begin* and *set out*, which is a symbol that they are used at the same time with a high frequency. Expressions, including *not*, *no*, *never ever*, are employed. *Not begin* and *not set out* are always used to express denying and negation.

6.2 Syntactic Aspects

ID tag levels within the scope of syntactic (number 17-31) are mostly located in the center in Figure 2, and they are placed more concentrated than those of morphological type. However, sentence as subject and zero subject are two protruding semantic features. In the sentences which contain the words *begin* and/or *set out*, the subject can be a form of a sentence/clause, while, in the sentences which contain the words *start*, *commence* and/or *launch*, there maybe exist no subject. This is, to some extent, can be interpreted as more imperative sentences are started with the words *start*, *commence* and *launch*, so as to express the meaning of pushing and encouraging someone to do something, and/or commanding the start of something.

Example 3: The participants at the Ronneby conference adopted the first Baltic Sea Declaration, which sets out a number of principles and priority actions necessary to enhance the Baltic environment.

In Example 3, an attributive clause "which sets out a number of..." has been used, and before *set out* is the word *which*. In this paper, the author has clarified this form of expression as sentence/clause as subject. With the frequent usage of it, we can make a brief conclusion that *begin* and/or set out can sometimes be used in dependent/subordinate clause.

Example 4: Soon after the conclusion of the late war, there commenced a memorable change in the treatment of these colonies.

Commence has been used in an inverted sentence without a subject, rather with *there* in the front. It is similar to the passive voice, in which the actual agent do not appear, showing an objective expression and meaning.

6.3 Semantic Aspects

In terms of semantic aspects, both prosody and components following the synonyms are used differently compared with each other.

Firstly, it is about the prosody. Semantic prosody, also known as semantic penetration, is a corpus-related term coined by Sinclair (1991) by borrowing the word once used by Firth. Through observations of a large number of language materials in actual use from the corpora, it is found that there are certain semantic rules in the collocation of words—some words have been habitually combined with other words that have the same or similar semantic characteristics. Because these words are collocated frequently in the text, they have a special semantic feature attached to the context, resulting in a certain fixed semantic atmosphere in the whole language environment, which is semantic prosody (Louw, 1993). Semantic prosody can be divided into three categories: positive prosody, neutral prosody, and negative prosody (Stubbs, 1995), which is the division in this research.

The words with negative semantic meaning are basically used together with node words, and the context will be infected with negative semantic atmosphere. *Start* and *begin* are more likely to be combined with negative expressions. For example,

Example 5: The room's twenty-plus occupants cheered as the crowding began to diminish.

Example 6: He began to doubt his calculations and wondered whether by some chance he had predicted the wrong....

Example 7: The fiesta begins with the shooting off of a rocket from the balcony of the Town Hall and

In these three examples, we can see the words having negative and frustrating meaning have been used after *start* and *begin*, including, *diminish, doubt, shoot* and so on. These words are in nature likely to arouse people's upsetting emotion and feelings.

On the contrary, in positive semantic prosody, node words are used to collocate with words with positive semantic characteristics, which makes the whole context positive and relaxing. *Commence* and *set out* are two expressions often related with applauding words. For example,

Example 8: He set out to develop his own study of distance running, gleaning training programs from magazines.

Example 9: It set out to revolutionize the Indian education, and to help hundreds of millions of its citizens....

Example 10: This commences with the full acknowledgment of his belief that God had made important revelations in past....

Words appear with *commence* and *set out* tend to be appealing, supportive, and encouraging, enabling the readers feel relaxed. In the meanwhile, *commence* is always used with solemn and somber events.

7. Conclusion

In the former sections, the research steps, statistical data, and semantic analysis have been demonstrated. On such a basis, conclusions of this research will be made in this section, embodying major findings and suggestions for further relevant study.

In this research, the author adopts the corpus-based behavioral profile approach, combines the hierarchical cluster analysis with the correspondence analysis, and makes visual presentations and systematic interpretation of the semantic characteristics of English synonymous expressions *start*, *begin*, *launch*, *commence*, *set out*, all meaning “to start”. Compared to the former traditional research methods, this research functions better in semantic study.

Several academic findings have been discovered by conducting this research. Generally speaking, the usage frequency of these five synonymous expressions is greatly distinctive. From the highest to lowest frequency are *start*, *begin*, *launch*, *set out*, *commence*. Besides, they are used in various genre and style. The word *start* has appeared mostly in the genre of blog, which is an informal genre, and least used in academic genre. Compared to this, the other four expressions tend to appear in, to some extent, formal genre, expressing the solemness of them. Specifically speaking, according to the hierarchical clustering analysis, *begin* and *set out* are most similar to each other, and then they are combined with *start* to create a semantic branch; while, *commence* and *launch* are clustered as another independent branch. In addition, different ID tag levels have been used when it comes to different expressions meaning “to start”, and the interrelation is regular and can be detected. This quantitative analysis will help learners better distinguish this set of synonymous expressions, and also provide reference for lexicography.

Furthermore, suggestions for further study are as follows. Firstly, language materials used in this research is still in a small quantity, so scholars can broaden the self-built corpus in the future. Secondly, based on the former studies, we can find that lexical verbs have been more studied on than abstract verbs, which creates a huge research gap for scholars.

References

Atkins, B. T. S. (1987). Semantic ID tags: corpus evidence for dictionary senses. *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary*, 17-36.

Berez, A., & S. Gries. (2009). *In defense of corpus-based methods: A behavioral profile analysis of* *polysemous 'get' in English*. Proceedings of the 24th Northwest Linguistics Conference.

Divjak, D. (2006). Ways of intending: Delineating and structuring near synonyms. In S. Th. Gries, & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis* (pp. 19-56). Berlin/New York: Mouton de Gruyter.

Divjak, D., & S. Th. Gries. (2006) Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistics Theory*, (1), 23-60.

Edmonds, P., & Hirst, G. (2002). Near synonyms and lexical choice. *Computational Linguistics*, *28*(2), 105-144.

Firth, J. (1957). *Papers in Linguistics*. New York: OUP.

Glynn, D. (2014). Correspondence analysis: Exploring data and identifying patterns. In Glynn, D. & Robinson, J. (Eds.), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy* (pp. 444-485). Amsterdam/Philadelphia: John Benjamins.

Gries, S. Th., & N. Otani. (2010). Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal*, (1), 121-150.

Hanks, P. (1996). Contextual dependency and lexical sets. *International Journal of Corpus Linguistics*, *1*(1), 75-98.

Hatch, E., & Brown, S. (1995). *Vocabulary, Semantics, and Language Education*. Cambridge: Cambridge University Press.

Klavan, J. A. (2014). Multifactorial corpus analysis of grammatical metonymy. In D. Glynn., & J. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy* (pp. 253-278). Amsterdam/Philadelphia: John Benjamins.

Krishnamurthy, R. (2006). Collocations. In Brown, K. (Ed.), *Encyclopedia of Language & Linguistics*. Boston.

Liu, D. (2010). Is it a chief, main, major, primary or principal concern? A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics*, (1), 56-87.

Liu, D., & M. Espino. (2012). Actually, Genuinely, Really, Truly: A corpus-based behavioral profile study of near-synonymous adverb. *International Journal of Corpus Linguistics*, (2), 198-228.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Baker, M., Francis, G., & Tognini-Bonelli, E. (Eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins.

Lu, J. (2010). A corpus-based study on collocational behavior and semantic prosody of near synonyms in Chinese learner English. *Modern Foreign Languages*, *33*(03), 276-286+329-330.

Nida, E. (1964). *Toward a Science of Translating*. Leiden: Brill.

Pang, Y., & Yang, Y. C. (2012). A corpus-based study of lexical priming of synonymy: taking begin,start and commence as examples. *Foreign Languages and Their Teaching*, (6), 21-25.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: OUP.

Stubbs, M. (1995). Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of Language*, *2*, 23-55.

Susan, N., & Thomas, E. (2010). The (near-)synonyms begin and start: Evidence from translation corpora. In Helsinki (Ed.), *Re-thinking synonymy: semantic sameness and similarity in languages and their description-Book of Abstracts*.

Taylor, J. R. (2003). Near synonyms as co-extensive categories: high and tall revisited. *Language Sciences*, *25*, 263-284.

Vandevoorde, L. et al. (2017). A corpus-based study of semantic differences in translation: The case of inchoativity in Dutch. *Target-international Journal of Translation Studies*, *29*, 388-415.

Wang, L. L. (2016). A corpus-based contrastive study of begin and start. *Journal of Changzhi University*, *33*(01), 56-58.

Wang, R. (2016). A corpus-based comparison study of synonym by Chinese EFL learners: A case study of begin and start. *Value Engineering*, *35*(19), 170-173.

Wu, S. Q., & Liu, D. L. (2020). Quantitative corpus methods for lexical semantic studies: Behavioral analysis. *English Studies*, (01), 153-164.

Wu, S. Q., Liu, D. L., & Liu, Q. (2021). A corpus-based behavioral profile study of synonymous adverbs indicating confirmation: A case study of 的确 dique, 确实 queshi, 实在 shizai and 着实 zhuoshi. *Foreign Language Education*, *42*(05), 19-25.

Yang, J. Z. (2007). Corpus collocation extraction and English synonym discrimination. *Technology Enhanced Foreign Language Education*, (04), 41-46.

Zhang, J. D., & Liu, P. (2005). Corpus-based approaches to the differentiation of English synonyms. *Foreign Languages Bimonthly*, (06), 53-56+100.

Zhang, J. D., & Liu, P. (2006). A corpus-based study of the differences between the three synonyms: happen, occur & “fasheng” (发生). *Foreign Languages Research*, (05), 19-22+80.

Appendix 1

Type of ID tag	ID tag	ID tag levels	start	begin	launch	commence	set out
morphological	tense	present	54	49	65	68	60
		past	45	49	32	29	40
		future	1	2	3	3	0
	aspect	simple	92	84	97	82	82
		perfective	2	7	2	7	12
		progressive	6	9	1	11	6
	mode	indicative	88	98	100	91	98
		imperative	9	2	0	9	2
		subjunctive	3	0	0	0	0
	voice	active	94	99	89	92	98

		passive	6	1	11	8	2
	negation	affirmative	100	95	100	100	93
		negative	0	5	0	0	7
	person	first	30	13	8	9	23
		second	12	7	1	11	8
		third	58	80	91	80	69
Syntactic	sentence type	declarative	88	96	99	91	98
		exclamative	0	0	0	0	0
		imperative	9	2	0	9	2
		interrogative	3	2	1	0	0
	clause type	main	67	81	77	82	80
		subordinate/dependent	33	19	23	18	20
	subject	pronoun	50	29	18	11	49
		noun/noun phrase	26	63	67	70	44
		sentence	1	0	1	0	2
		zero	23	8	14	19	5
	object	pronoun	0	0	3	0	0
		noun/noun phrase	14	9	57	36	15
		verb/preposition phrase	70	75	18	19	50
		sentence	1	0	0	0	0
		zero	15	16	22	45	35
semantics	prosody	positive	11	13	7	23	26
		negative	51	32	20	11	19
		neutural	38	55	73	66	55
	following components	to do	12	42	1	6	49
		doing	35	19	0	4	1
		noun/noun phrase	11	12	72	38	14
		collocation	7	0	0	0	0
		clause	2	0	0	0	3
		zero	18	10	5	30	5
		pre-with	8	8	0	8	5
		pre-time	3	7	9	10	2
		pre-place	3	2	9	1	16
		pre-by	1	0	4	1	0
		pre-in	0	0	0	2	0
		pre-for	0	0	0	0	5