DOI: 10.23977/langl.2023.060305
ISSN 2523-5869 Vol. 6 Num. 3

A Study of Synonyms Based on COCA Corpus Road and Street as Examples

Lin Hao, Xu Shujuan

College of Foreign Languages, North China University of Science and Technology, Tangshan, Hebei, 063210, China

Keywords: COCA Corpus, English Learning, Near synonym identification

Abstract: There are a large number of near-synonyms in English, differentiating these words have always been one of the major difficulties in English learning. The participation of the corpus provides a new approach to the identification of near-synonyms.Based on the Corpus of Contemporary American English (COCA), this study identifies Road and Street, a group of near-synonyms in terms of register, collocation and semantic rhyme, and the results show that the main difference between the prosody is in semantic and lexical collocation.The purpose of this paper is to help learners further improve their vocabulary recognition skills by means of the corpus.

1. Introduction

English is one of the richest languages in the world in terms of synonyms. According to statistics, the number of synonyms and near-synonyms in the English language accounts for more than 60% of the total vocabulary.[1]Vocabulary learning has always been one of the major difficulties in English learning. These synonyms create a certain degree of confusion for learners in terms of lexicality, word meaning, structure or usage. Therefore, synonym identification is a difficult aspect of vocabulary learning and teaching. The traditional way of vocabulary analysis is to consult a dictionary or to seek advice from a teacher, whose understanding of the meaning of words is sometimes subjective, while the interpretation of dictionaries aims to make readers understand the general meaning rather than specific analysis. The advancement of technology has led to the development of modern corpus technology. The corpus-based near- synonym word analysis provides a new perspective for near-sense word analysis with the help of corpus, which is truly huge and easy and fast to search the vocabularies.Road and Street is a group of vocabulary that is with high-frequency in the English syllabus of Chinese primary and secondary schools and are familiar to English learners. Therefore, this study takes them as examples and analyzes their similarities and differences in terms of register, collocation, and semantic rhyme based on the Corpus of Contemporary American English (COCA).

2. Theoretical foundation

The study combines theories related to register, collocation and semantic rhyme. In layman's terms, a domain is a conversational context in which language is used differently. The British linguist Han Liddell defines domain as "the general term for the occasion or domain in which language is used", which explains the social nature of language and provides a basis for the distinction between synonyms.[2]Collocation in linguistics refers to the habitual conjunctive use of words, which is a linear co-occurrence of relationships within a certain range.[3]Researchers in corpus linguistics have long summarized a set of statistical instruments to measure this linear co-occurrence relationship, i.e., to calculate the mutual attraction between nodal words and collocations within a certain span.[4]Semantic rhyme is an important linguistic mechanism discovered and studied by corpus linguists. It describes the collocational behavior of words, where words habitually attract words with the same or similar semantic features to form collocations with them, thus presenting a special semantic collocational atmosphere [5]. Semantic rhyme has an attitude marking function, and Stubbs classifies semantic rhymes into positive, negative and neutral rhymes according to the overall semantic features of collocated words.[5-6]

3. Study Design

3.1 Research Questions

A comparison of the usage of Road and Street in the corpus is studied.

3.2 Research Tools

The Corpus of Contemporary American English (https://www.english-corpora. org/coca/), or the COCA corpus, contains over 1 billion words (20 million words per year from 1990-201) from spoken language, fiction, popular magazines, newspapers, academic texts, television and movie subtitles, blogs, and other web pages. The COCA corpus is used as the primary research tool in this study because of the referential nature of its data, which originates from native English-speaking countries. Frequency distribution chart

3.3 Near synonym selection

The COCA corpus was searched for synonyms by entering [=road] in the List option, and the close synonyms of Road were obtained in descending order of word frequency. Figure 1 lists the top five words in terms of word frequency distribution.

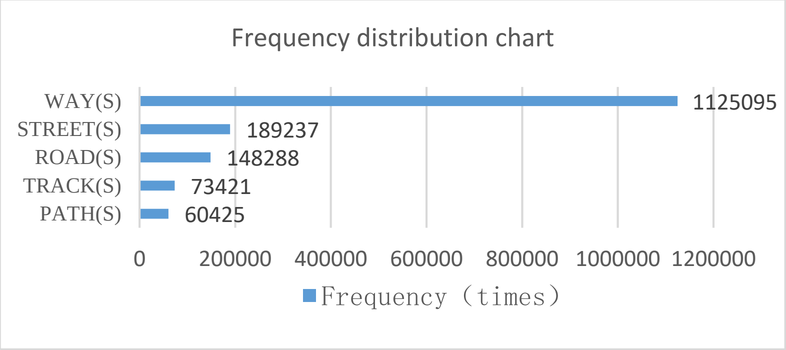

Figure 1: Frequency of near-synonyms for Road

As shown in Figure 1, the word Way far exceeds the frequency of other words, but the word has multiple meanings, and the meaning of "road" is only one of them, and the frequency of the words Street and Road are in the top and close to each other, so this group of words is worth studying.

4. COCA-based comparison analysis of ROAD and STREET

When you type "street" in the word software (take Youdao Dictionary as an example), the Chinese definition given is "street", while "road" gives the following meanings such as “lu”, “gonglu” and “malu”. From the condensed expressions, it is easy to see that there is a difference in the spatial scope of the two terms, and the spatial size is distinguished by "lu", "dao", and “jie”, which is a specific place of use. However, the differences between Chinese and English languages may cause the lack of information in the translation process. The Oxford English Dictionary.[7]The English definitions of the words Street and Road are as follows:

“Road [noun]

1) a wide way leading from one place to another, especially one with a specially prepared surface which vehicles can use.

2) a series of events or a course of action that will lead to a particular outcome.

Street[noun]

3) a public road in a city, town, or village, typically with houses and buildings on one or both sides.”

According to the English interpretation, Road not only has the basic meaning of the word itself, but also has the meaning of "a certain stage" and a proper noun. In addition, in English communication, the concept of "jie" is blurred, and the concepts of "lu" and "dao" are rarely taken into account, which is confusing for English learners to use. In view of this, the following is a comparative analysis based on data from the COCA corpus in terms of register, collocation and semantic rhyme.

4.1 Comparison of language domains

The language domain is reflected in the different style and tone of the language.[8]In the COCA corpus, the domain can be seen as a distribution of word frequencies (word frequency here refers to the number of occurrences of the retrieved word per million words).The frequency distribution of words varies from one word to another, and the search allows us to compare the distribution of words, which is useful for reference in the vocabulary usage scenario. The results of the "chart" function are shown in the following table.

Table 1: Word frequency classification of Road versus Street in different domains in the COCA corpus

		News	Fiction	Magazines	Web	BLOG	spoken language	TV Movies	Academic
Road	word frequency	32315	28740	20517	16841	14781	14121	13912	6982
	Frequency(per mil)	265.44	242.90	162.72	135.54	114.93	111.95	108.62	58.29
Street	word frequency	40017	39444	21782	18291	18531	24160	18465	8420
	Frequency(per mil)	328.70	333.36	172.75	147.21	144.08	191.54	144.17	70.29

Table 1 lists the word frequencies of the two search terms in different corpus sources, and the comparison shows that the texts with the largest distribution of both terms are news, followed by fiction, while the two terms appear least frequently in academics, indicating that they appear most frequently in news texts and literary texts, and have a certain degree of formality in their use, while the degree of academics is not very prominent. In addition, Street appears more frequently in spoken language, followed by TV, Blog and Web, indicating that the word is more often used in spoken and everyday concrete contexts. The word Road, on the other hand, appears less frequently in spoken language.

4.2 Pairing Comparison

The collocation of words has an important role for English learners. Combined with the COCA corpus, we can analyze the characteristics of collocated words by calculating the MI values of collocated words and nodal words to calculate the collocation intensity. MI value is Mutual Information Value (MI), the larger the MI value, the higher the collocation strength between the two words.[8]The results can be studied by filtering collocations through the corpus and combining them with MI values. In the COCA corpus, the search terms are entered separately through the "collocate" function, and since the selected words are all nouns, the spacing between the left and right collocations can be set to "1" when examining collocations. Considering the large number of collocations, in this study, the top 20 collocations in terms of frequency will be listed, as shown in Table 2

Table 2: Collocation of Road and Street in the COCA corpus

Road			Street		
word frequency	MI	New Words	word frequency	MI	New Words
2845	6.96	TRIP	29692	9.27	WALL
1914	8.01	DIRT	7978	8.27	JOURNAL
1655	7.00	MAP	6251	7.21	MAIN
1218	5.20	MAIN	1539	9.80	SESAME
1056	4.85	AHEAD	1278	5.78	CORNER
709	9.84	N.E	990	3.82	MARKET
593	3.87	GAMES	871	3.00	SIDE
583	6.97	RAGE	753	7.23	CORNERS
574	6.54	TRIPS	587	7.78	14TH
560	3.80	COUNTY	544	10.22	42END
529	8.01	GRAVEL	479	4.83	BRIDGE
523	9.03	PEACHTREE	451	8.57	TWO-WAY
509	7.13	SILK	446	7.49	16TH
466	4.53	HILL	444	4.57	LIGHTS
442	6.97	TOLL	425	9.15	DOWNING
433	3.00	TOWARD	421	5.59	GANG
432	4.80	CONSTRCTION	418	4.40	SIGNS
425	7.99	WINDING	417	4.75	K
424	8.28	PAVED	414	4.85	BROAD
423	6.77	MILL	411	6.15	PROTESTS

Table 2 collocations are found at the left or right end of the retrieved words, and the first 20 words formed by clicking on the collocations are road trip(s), dirt road, road map, main road, road ahead, road N.E, road games, road rage, county road, gravel road, peach tree road, silk road, hill road, road toll, road toward, road construction, winding road, paved road, mill road. According to the results, among the first twenty collocations, Road mostly precedes and qualifies the subsequent components. In terms of lexical nature, the collocations were counted as 13 nouns, 4 adjectives, 2 adverbs, and one contraction. This indicates that the word is mostly collocated with nouns. And according to the collocation results, among the 20 words with high frequency, "road rage (road rage)", "silk road (silk road)", "road games (away games) The three groups of words "road rage", "silk road", and "road games" are all new words, which are derived from the new concepts combined with the derivation of Road. In usage, the word is often collocated in front, serving to enrich the modifying or qualifying role of the postnominal trait.

Similarly, the twenty words forming collocations with Street can be combined as "Wall Street Journal (from the source corpus, the first two collocations form an exclusive collocation with the search term)", main street, sesame street, street corner(s), market street, side street, 14th street, 42 end street, bridge street, two-way street, 16th street, street lights, Downing Street, street According to the results, street was mostly qualified or modified after the first twenty collocations. From the lexical point of view, according to the statistics, there are 14 nouns, 2 adjectives and 3 number words in this collocation, in addition to one abbreviation. This indicates that the word is likewise mostly collocated with nouns. In terms of collocation, among the 20 words with high frequency, there are such words as "Wall Street Journal", "Downing Street (Downing Street, the British Prime Minister's official residence)", which are associated with news and politics, and also such words as "Wall Street Journal" and "Downing Street". There are also collocations with American cultural labels such as "Sesame street". Street is used in its original sense, combined with a specific name or number to become a place noun.

4.3 Semantic rhyme comparison

Semantic rhyme itself has the function of expressing attitudes. According to Subbs, semantic rhymes are divided into three main categories: positive semantic rhymes, neutral or mixed semantic rhymes, and negative semantic rhymes [6]. The identification of semantic rhyme types usually relies on whether the affective tone embodied in the corpus is positive, negative, or neutral. Adjectives themselves have a grammatical meaning that indicates the nature and state of things, and the semantic rhyme of retrieved words can be explored in the COCA corpus based on adjectival lexical collocations. The top 20 adjective collocations in the COCA corpus were searched by entering the retrieved words and the regular expression "[j*]", and their MI values were all greater than The results are listed below in descending order of frequency.

Table 3: Semantic collocation of Road and Street in the COCA corpus

	Adjective collocation
Road	LONG, MAIN, HIGH, OPEN, PAVED, WINDING, NEW, TOUCH, HARD, NARROW, ROCKY, TWO-LANE, LOW, DUSTY, ROUGH, EASY, OLD, BUMPY, ONLY, DIFFICULT
Street	MAIN, TWO-WAY, BUSY, ONE-WAY, HIGH, NARROW, RISIDENTIAL, QUIET, EMPTY, EASY, DARK, ARAB, SUBURBAN, CROWDED, OTHER, PUBLIC, LITTLE, TREE-LINED, WHOLE, DESERTED

As shown in the table 3, the use of the adjectives collocated with both words is objectively descriptive, without positive or negative semantic tendencies, and is a neutral semantic rhyme. Clicking randomly on the collocated words, checking the source corpus, and analyzing them with the context, the words co-occurring with both words have no obvious semantic tendency, and the collocated words also have no obvious semantic tendency in the context, so it can be concluded that both words have neutral semantic rhyme.

5. Conclusion

The corpus provides a new means for the study of lexical discrimination. The COCA corpus was used to analyze the near-synonymous words Road and Street in terms of domain, collocation and semantic rhyme, and the main differences between the two groups of words were found to be in the domain and collocation. From the perspective of domain, the two words appear most frequently in news and novels and least frequently in academic articles, and Street is mostly used in spoken language and daily life corpus, while from the perspective of collocation, Road has a broader collocation and is often used as a definite component or to derive concepts, while Street is mostly used to form specific places with its collocation. The COCA corpus relies on a large corpus to identify words, and the results are well supported by the corpus, allowing language learners to learn in a more effective way.

References

[1] He Xiaodong. English-Chinese Dictionary of English Synonyms. Beijing: The Commercial Press, 2003
[2] Halliday M A K, Hasan R, Han lide .Cohesion in English. Beijing: Foreign Language Teaching and Research Press, 2012.
[3] Firth J R. A synopsis of linguistic theory 1930-1955// Firth J R. Studies in linguistic analysis. Oxford: Philological Society, 1957.
[4] Yang Chunxia. A study of corpus-based near-synonym identification - taking suspect and doubt as examples. Journal of Southwest University of Science and Technology (Philosophy and Social Sciences Edition), 2014, 31 (5):45-49, 86.
[5] Liu Qianqian, Liu Min. A Corpus-based Investigation on News Attitudes towards the Belt and Road Initiative. English Abroad, 2018 (3):198-200.
[6] Michael Stubbs. Text and Corpus Analysis: Computer-assisted Studies of Language and Culture [J]. Functions of Language, 1996, 3(2):269-272.
[7] A. S. Hornby. Oxford Advanced Learner's English Dictionary. Oxford University press, 2014.
[8] Li Jingyi. A comparative analysis of English near-synonyms based on the COCA corpus - taking ability and capability as examples. English Abroad, 2020(15):3-4+15.