International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Corpus Based Study on the Colligation of *The Time Machine*

Yuxin Zhu
Foreign Language School, Jinan University, GuangZhou, China
Corresponding Author: Yuxin Zhu, **E-mail**: 853793483@qq.com

| ABSTRACT

The past 50 years have witnessed huge progress and great evolution in Corpus Linguistics. Corpus Linguistics and colligation have their solid philosophy foundations. Colligation is believed to be the co-occurrence of the node words and abstract grammatical categories; thus, it can be an abstract reflection of certain writing habits in language use. The Time Machine is a famous science fiction written by George Herbert Wells. Through the study of colligation in this science fiction, the author tends to use different writing techniques and typical collocation. The result shows that the attributive clause is widely adopted in the text, followed by the object clause. The flaws of the software CLAWS Tagger are also discovered through the study.

| KEYWORDS

Corpus linguistics, The Time Machine, Colligation, Collocation

| ARTICLE DOI: 10.32996/ijllt.2022.5.2.8

1. Introduction

The past 50 years have witnessed huge progress and great evolution in Corpus Linguistics ever since the 1950s. The relevant research fields of corpus linguistics are also being expanded. Up to today, Corpus Linguistics has been widely applied to various fields, the most apparent of which is Applied Linguistics (in its narrow sense, namely, second language acquisition), sociolinguistics, translation study and comparative translation study, etc. More and more corpora are being built, widely used and fully revised, most of which are written corpora. The most typical and renowned ones are BNC, COCA, etc. Different types of corpus provide convincing linguistics evidence for different purposes and usages of the study. The research method of corpus linguistics emphasizes inductive and empirical research.

Relevant techniques and skills such as computer software, probability, statistics, etc., are being excavated and exploited by many corpus linguists as well. For example, AntConc is mainly used as a concordancer to retrieve information as user’s needs; Sketch Engine is a rather powerful linguistics tool equipped with various functions other than evidence retrieving; moreover, there are LancsBox, W-matrix, and all manner of tools. Other than these large and acknowledged corpus tools, there are also some small and self-developed ones that have been created by various institutions, such as Colligator 1.0, Collocator 2.0, developed by BFSU, providing an automatic algorithm for log-value and log-likelihood value, etc.

Written by George Herbert Wells in 1895, the science fiction *The Time Machine* gives an account of the story of a time traveller who invents the time machine himself and uses it to travel through time back and forth, during which a lot of intrigued things happen. Influenced by Wells’ thought, this sci-fi also takes evolution and class antagonism as two central subjects, which may or may not shed light on this paper in the study of the conflict and acute language used in the novel.

This paper seeks to investigate the abstract language system of *The Time Machine* through colligation by means of relevant corpus linguistics skills and corpus tools, AntConc, CLAWS Tagger and Sketch Engine.

2 Literature Review

Corpus linguistics is a corpus-based science that analyzes, counts, and studies many natural and real corpora. In recent years, corpus linguistics has developed rapidly and has been widely used. Corpus Linguistics has its deep and solid philosophical origin. The famous British linguist R.H. Robins (1997) once pointed out that the opposition between empiricism and rationalism will run through the whole history of linguistics in different forms. The philosophical basis of corpus language is empiricism (Zhang, 2011).

In the 1950s, Wittgenstein (1953) put forward his famous linguistics viewpoint suggesting that the meaning of a word lies in its usage. The famous philosophical topic 'the meaning of a word is its use in the language' (Wittgenstein,1953) indicates that meaning is not a concrete entity but the nature and function of language. This proposition made a huge and great twist in the philosophy of language. What can be concluded from Wittgenstein is that using the language is an obvious action of endowing language meaning. It was truly forward-thinking for Wittgenstein to emphasize the importance of context back in the 1950s. In corpus linguistics, real and practical linguistics evidence can be provided, which, to some extent, caters to Wittgenstein's view of language use because a word can have multiple usages under different contexts. Ever since then, Linguistics, especially Pragmatics in the later period, has been deeply influenced by it.

Later, the founder of the London School, Firth, put forward his contextual theory of meaning, which stress the non-detachability of context in an even more profound manner (Firth, 1957). Thus, to truly understand the meaning, what we need to do is to base our study on the analysis of context and its function; what's more, Firth (1957:19) also clarifies that a complex context consists of phonetics, grammar, words and semantics, all of which can be exactly highlighted in corpus through ways of annotation and tagging.

In Firth's work in 1957, he assumes that all statements of meaning are actually statements of their 'contextual relations', which indicates that understanding the meaning of a word is more constructive for us to inspect it by collocation. He (2013) then picks up this standpoint and crystallizes that there are actually 2 concepts included in his statement; one, the concept of collocation, referring to the repeated co-occurrence of certain words (for example, the word 'dark' usually comes with 'night'); two, the concept of colligation, denoting to the habitual and grammatical co-occurrence between parts of speech (for example, adjectives usually show up with nouns and nouns verbs).

The concept of colligation is believed to be a noun raised by J.R. Firth by most scholars, yet back in Firth's thesis of 1957, he himself pointed out that it was H. F. Simon who, for the first time, brought forward the concept of colligation back in the 1950s and categorized it into two kinds, i.e. major colligation and minor colligation, with the former denoting a whole sentence, the latter indicating different language structures in the sentence (H.F.Simon,1953). As Simon (1953:25) once noted, 'The term *colligation* is applied to describe the syntactic juxtaposition of two or more categories， and the derivative 'colligability' has been coined to denote the *possibilities de combinaison* of any given category.'

Moreover, colligation is said to be the mutual relation of syntax between grammatical categories (Firth,1957). The concept was further developed by J.R. Firth, who believes that as for the research for discourse meaning, colligation should be interpreted in the perspective of syntax(Firth,1957:30). Both Simon and Firth are representatives of the London School; therefore, the concept of colligation was given birth to from the research of situational meaning by London school, restricted to the scope of syntax category (Pang,2014). In later research, both the connotation and denotation of colligation has surpassed the original definition set by Firth and other scholars that started the study of the colligation, but its central idea remains almost the same.

Sinclair (1996) believes that words or phrases don't just show up randomly; thus, he develops the concept of colligation by clarifying the roles played by semantic preference and semantic prosody. He defines colligation as the co-occurrence of the node words and abstract grammatical categories. In his work *The Search for Units of Meaning*, he states that phrase is the carrier of meaning and proposed his 3C2S method in corpus analysis, that is, central word-collocation-colligation- semantic preference- semantic prosody. He breaks the limit of the original definition of colligation and expands it to the combination of subject-predicate. Hoey (1997) expounds that colligation is about the grammatical combination and the priority in word-positioning. He points out that some words are more likely to be at the beginning of a passage while others in the ending.

In today's linguistics research, especially corpus linguistics research, colligation has become a rather popular field towards which linguists are paying more and more attention. In 2006, Baker officially listed colligation as an item into A Glossary of Corpus Linguistics. McEnery (2006) makes his revision suggesting that the concept of colligation is usually expanded to the combination or connection between grammatical words and other lexical words, for such combination does not require the part of speech tagging and can be easily observed in the raw corpus.

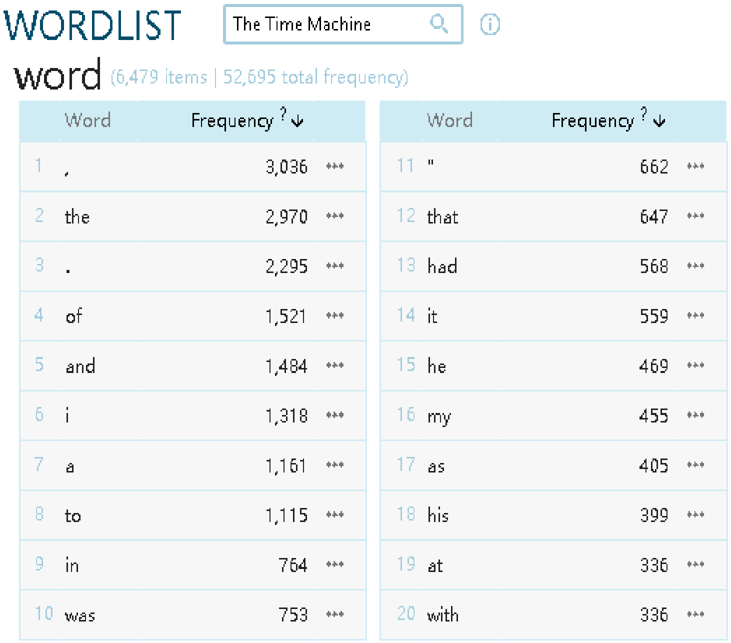
From Pang (2014), the research of colligation aboard is about its combination with collocation on the basis of tagging, and that of domestic is mostly concerned with a certain kind of word or topic. Regardless of different focal points, what is apparent in both study abroad and home is that the research of colligation is usually linked to applied linguistics, mostly in language acquisition.

Firth (1957) once stated that colligation is not as salient as collocation to be noted by most people, while both of them are grammatical concepts. Thus it is necessary for us to dig out the commonalities among languages through some looks into its abstract colligation (Pang,2014). Colligation is believed to be the abstract combination of chunk collocation (or, as quoted from Xu and Xiong, the combination of '词语搭配'). The study of it is even more helpful to investigate the relationship between sentence structures between different languages. It makes a huge breakthrough from the original restriction of word, phrase, sentence or paraphrase, thus shedding light on a more abstract facet of language research (Xu, Xiong, 2009). Pu Jianzhong (2003) used three collocations of 'discover', namely, discover n / discover that / discover wh-, as an example to illustrate that colligation can be used to represent the grammatical characteristics of a word or a class of words. What can be concluded is that colligation is a level higher than collocation to present the co-occurrence between word and grammar. The relationship between collocation and colligation is also worth noting. Cong (2011) believes that the collocation of a certain word is the demonstration of its colligation, for example, colligation *art. + adj.+ N.* has reflected in its collocation *a major reason*. Most of the research towards collocation are discussed within the frame of its colligation. So far, we have every reason to believe that it is urgently and practically needed to study colligation from every aspect.

The empiricism-nature of corpus linguistics has been previously mentioned. It is necessary for us to bear in mind that corpus linguistics is developing, and the philosophy of language are also experiencing a lot of great changes. After the development of empiricism in the middle ages, especially up to modern times, empiricism carries some color of rationalism (Chen,2003). In the 20th century, the coexistence and integration of empiricism and rationalism were becoming more and more obvious (Chen,2003). The influence of rationalism and the combination of the two contribute to a new research method and model in corpus linguistics, that is, to combine the deductive method of rationalism with the inductive method of empiricism and to combine positivism, statistics and the rational thinking of researcher; not only should they pay attention to experience, but also combine the reasonable side of rationalism, adopt scientific rationalism and learn from experimental research methods (Zhang,2010). Zhang suggested a new research model: empirical induction- a prior deduction- proposing a theory- testifying theory- revising theory- proposing a new one. Such a new model is time and energy-consuming in language research to some extent; however, to cater to the need for philosophy development, it is truly worth the effort to integrate rationalism with empiricism. As a response, today, the building of basic corpus has been strengthened, and it has been applied to more different fields and disciplines such as translation studies, sociolinguistics, psycholinguistics, computer linguistics, etc.; moreover, corpus tools have also been developed. This paper, using statistics and probability to investigate the colligation of *The Time Machine* is, exactly combining rationalism with empiricism.

3 Discussion

In this paper, in order to investigate the colligation of The Time Machine, CLAWS Tagger, Sketch Engine, AntConc and Colligator 2.0 will be the tools that are used. First of all, the text is uploaded into Sketch Engine as a new corpus and then we used the wordlist function in the dashboard to see the most frequently used words.



WORDLIST The Time Machine

word (6,479 items | 52,695 total frequency)

	Word	Frequency ? ↓		Word	Frequency ? ↓
1	,	3,036	11	"	662
2	the	2,970	12	that	647
3	.	2,295	13	had	568
4	of	1,521	14	it	559
5	and	1,484	15	he	469
6	i	1,318	16	my	455
7	a	1,161	17	as	405
8	to	1,115	18	his	399
9	in	764	19	at	336
10	was	753	20	with	336

Picture 1: Wordlist of The Time Machine

In English, the word THAT carries many different roles in grammar, as is shown in its parts of speech and functions. THAT is believed to be the most frequently used word after which a clause is embedded. To some extent, THAT contributes to the recursiveness of language. In English, THAT can be a determiner, a pronoun, conjunction or an adverb, depending on the request of the sentence and grammar. As is shown in picture 1, the most frequently used words are grammatical words such as articles, conjunctions, prepositions, etc. What is worth noting is that the word THAT ranks 12 in the whole wordlist, with the frequency of 647 times. Thus, the word *that* is chosen as the item that this paper looks into.

We use CLAWS Tagger to automatically tag the parts of speech in the whole text. CLAWS Tagger is exploited to automatically annotate the text, applying the UCREL CLAWS7 Tagset (C7) as its basis. Apart from the punctuation tags, C7 tagging system contains over 160 tags as C6. In the C7 tagset, THAT is categorized into 2 different tags. CST refers to THAT as a conjunction, and DD1 stands for singular determiner including THAT. To cater to the need of this paper, CST will be adopted to see the subordinate clauses of The Time Machine. After tagging, the text is uploaded and opened in AntConc to see the patterns. In the raw corpus, THAT hits 647 times regardless of its part of speech.

E.g.:

1. We hope **that** you enjoy this sample… (object clause)
2. …lights in the lilies of silver caught the bubbles **that** flashed and passed in our glasses. (appositive clause)
3. …there was **that** luxurious after-dinner atmosphere when… (determiner)
4. I shall have to controvert one or two ideas **that** are almost universally accepted. (attributive clause)
5. Is not **that** rather a large thing to expect us to begin upon? (determiner)

When retrieving THAT as a conjunction in the tagged text, there are totally 384 times hit in the text, suggesting that there are 384 clauses led by THAT. As is shown below,:

E.g.:

1. We_ PPIS2 hope_ VV0 **that_CST** you_ PPY enjoy_VV0 this_DD1 sample_NN1…
2. …lights_NN2 in_ II the_ AT lilies_NN2 of_ IO silver_NN1 caught_ VVD the_ AT bubbles_NN2 **that_CST** flashed_ VVD, and_ CC passed_ VVD in_ II our_ APPGE glasses_NN2 ._.
3. I_PPIS1 shall_ VM have_ VHI to_ TO controvert_ VVI one_MC1 or_ CC two_ MC ideas_NN2 **that_ CST** are_ VBR almost_ RR universally_ RR accepted_ VVN ._.
4. You_ PPY know_VV0 of_RR21 course_RR22 **that_CST** a_AT1 mathematical_ JJ line_NN1…
5. Can_VV0 a_AT1 cube_NN1 **that_ CST** does_ VDZ not_ XX last_ VVI for_ IF any_ DD time_NNT1 at_RR21 all_RR22 ,_, have_VH0 a_AT1 real_ JJ existence_NN1 ?_?

…

Here, we take NN1 and NN2 as in the same tag NN, and VV0 and others referring to the lexical verbs VV when calculating its frequency. After analyzing, it is found that the following colligations are most frequently used in the text (shown in the table together with its collocation):

Colligation	collocation
VV+**CST** (106') VVD+RT+**CST** VVO+PPY+**CST** (6) VVO+DD1+**CST** VVD+PPX1+**CST** (3) VVG+RGT+RR+**CST** (2) VVD+RR+RR+**CST** (2)	we hope **that** you knew intuitively **that/** found afterwards **that** tell you **that** stand another **that** told myself **that** dreaming most disagreeably **that** I perceive clearly enough **that**
NN+**CST** (166') DB+**CST** NN1+PN1+**CST** (3)	the bubbles **that** flashed; a cube **that** does not I think of all **that** I ..adventure, one **that**
RR+**CST** (4)	you know, of course, **that** a mathematical line
JJ+**CST** (16)	it is very remarkable **that** this
DD1+**CST**	for this **that** follows-unless
PPIO1+**CST** (6*) PR+**CST** (2)	It appears incredible to me **that** It may seem strange, perhaps, **that**
CC+**CST** (10) CCB+**CST** (2) CSA+**CST**	..and **that**... ..but **that** ... ..as **that**...
(emphatic sentence) RT+**CST** (2) RL+**CST** RR+**CST**	 It was at 10 o'clock today **that** the. It was here **that** It was from her too **that**
VB+**CST** (8)	The fact is **that**
RR+**CST** (2) RR+RT+**CST** (2)	even **that** there was even then **that**
II+**CST** (5)	From **that,** I could

Table 2: The Colligation of 'THAT' in the text

In this paper, we put the tagged text into the AntConc and then analyze it. After calculating and eliminating those that are mistaken-tagged examples, the result is demonstrated in table 2.

Among 384 results hit by 'THAT' (apart from some mis-tagged one eliminate), the most frequently used clause led by THAT is an attributive clause, with the hit of 166 times.

THAT can be used as a word to lead an object clause; such writing technique is exploited in the novel. The object clause led by 'THAT' ranks the second in the result of frequency, hitting 121 times on the whole. What's worth noticing is that the author uses a variety of expressions when embedding the object clause. The most obvious is that the author inserts other components between verb-object grammatical structure. Inserting adverbs, indirect objects, pronouns, nominal adverbs of time, etc., are all choices made by the author (Examples are given in table 2, extracted from the original text). As can be seen in the collocation column, **'v+ adv+ that'** is the habitual and special expression of the author. What's more, it is also shown within the result that the 'VVD+PP+CST' colligation is highlighted; that is to say, the structure of **'v+ sb.+ sth.'** are exploited (Here, PP stands for both PPX and PPXI).

Also, the author fancies emphatic sentences led by THAT in his writing. The author seems to emphasize the time, person and place by adopting the structure **'It is/was adv. that'.** The structure of coordinating conjunction with the THAT clause is also another characteristic shown in the result.

4. Conclusion

From the result of THAT, it can be seen that the attributive clause is widely adopted in the text, followed by the object clause. The use of THAT in leading a subject clause is also highlighted.

CLAWS Tagger is a rather useful and effective tool for tagging the text, but there are still some problems that are worth noting and solving. When looking into the pattern of THAT, there is such a sentence in the raw corpus :

'And now I must be explicit, for this ***that*** *follows—unless his explanation is to be accepted****—****is an absolutely unaccountable thing..'*

In the automatically tagged text, it is tagged as '*for_ IF this_DD1 that_ CST* ***followsunless_*** *JJ his_ APPGE*'. The problem lies in that the Tagging system seems to be stuck with the breaking of the sentence, for it cannot recognize the hyphen and puts two words 'follows' and 'unless' together as one, which will definitely cause errors when concluding the pattern of the text. Also, in the original sentence 'At that the Time Traveller laughed cheerfully', THAT should be the object of LAUGHED by manual analysis, and somehow it is tagged as a conjunction by the tagging system.

Another problem I found here is that the tagset of the CLAWS Tagger is too meticulous in suiting the different needs of the study. For example, the pronoun YOU and the reflexive pronoun YOURSELF are 2 different tags in the tagset. However, in this study, both YOU and YOURSELF can be seen as the same structure 'tell sb. that'. Breaking them into 2 different categories will probably increase the workload of calculating the statistics.

To draw a conclusion, a more intelligent tagging system is a tough question and urgently need to be solved in order to investigate colligation further. Although so far the corpus tool has been greatly developed, there is still room for its intelligence to be improved.

Statements and Declarations

This research received no external funding. The author declares no conflict of interest.

References

[1] Firth, J. (1957). A synopsis of linguistic theory,1930 – 1955 [A]. In J. Firth et al. Studies in Linguistic Analysis[C]. Oxford: Blackwell.

[2] Hoey, M. (1997). From concordance to text structure: New uses for computer corpora [A]. In Melia, J. & Lewandoska, B. (eds.). Proceedings of PALC 97 [C]. Lodz: Lodz University Press.

[3] Simon, H. F. (1953). Two Substantival Complexes in Standard Chinese[J]. Bulletin of the School of Oriental and African Studies, University of London.

[4] Sinclair, J. (1996). *The Search for Units of Meaning* [M]. Oxford: Oxford University Press.

[5] McEnery, T., X. Richard & T. Yukio. (2006). *Corpus-Based Language Studies* [M]. London: Routledge.

[6] Wittgenstein, L. (1953). *Philosophical Investigations (3rd edition)* [M]. New York: Macmillan.

[7] 陈勇.论经验主义和理性主义之争——关于西方语言学研究中的认识论[J].外语学刊,(2003) (03):57-62.DOI:10.16263/j.cnki.23-1071/h.2003.03.012.

[8] 丛丽君.基于语料库的搭配和类联接研究——以回指概念外壳名词为例[J].广州大学学报(社会科学版),(2011) ,10(06):80-84.

[9] 何安平.从"意义即使用"哲学观到语料库的"意义单位"探究. 外语与外语教学(03),44-48. doi:10.13458/j.cnki.flatt.003915.

[10] 濮建忠.英语词汇教学中的类联接、搭配及词块[J].外语教学与研究,2003(06):438-445+481.

[11] 罗宾斯, 德宝, 建明, & 明亮. (1997). *简明语言学史: A short history of linguistics*. 中国社会科学出版社.

[12] 庞双子.类联接概念的演变及其与语料库语言学研究范式的内在联系[J].外语电化教学,(2014) (02):11-16.

[13] 许家金,熊文新.基于学习者英语语料的类联接研究概念、方法及例析[J].外语电化教学,(2009) (03):18-23.

[14] 张红燕,金晶,逄锦凤.论语料库语言学的哲学基础[J].湖北民族学院学报(哲学社会科学版),(2011) ,29(01):154-156.