2022 International Conference on Machine Learning and Knowledge Engineering (MLKE) | 978-1-6654-9567-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/MLKE55170.2022.00064

Text Mining Of English Grammar Books In 19th ~ 20th Century

Chao Cui
College of Foreign Languages
Inner Mongolia Agricultural University
Hohhot, China
e-mail: cuichao@imau.edu.cn

***Abstract*—English grammar books are important materials for English learning. In order to systematically analyze the themes and trends of English grammar books in the 19th and 20th centuries, unstructured data and structured information are correlated. In this paper, text mining, co-word analysis and other research methods were adopted to study 15 books named English Grammar in the Gutenberg Project, which lasted from 1812 to 1918 and had a total of 1863536 English words. Wordstat9.0.4, KH CODER3 and Excel were used to analyze the collected information. This paper presents the results of text mining, such as word frequency analysis, co-occurrence analysis, keyword analysis and part of speech analysis.**

Keywords-English grammar books; diachronic research; text mining

I. INTRODUCTION

The history of The development of English grammar can be divided into three stages[1] : 16th to 17th century, 18th century, 19th to 20th century.

English grammar books are systematic summaries and generalizations of the rules of the English language. For English learners, English grammar goes through all parts of language listening, speaking, reading and writing. Existing researches on English grammar mainly focus on such topics as English grammar teaching methods, explanation of grammar rules, grammar cognition and second language acquisition. To integrate the study of English grammar into the disciplines of language and literature, education, mathematics and computer science[2-4]. The text mining technology adopted in this paper was mostly used in the research fields of computer science, library and history of science and technology in the early days, while the text mining method was seldom used in the research of English linguistics. Therefore, from the perspective of academic value, the research results of this paper can enrich the theoretical content of English grammar research. From the perspective of application value, the results of this study can bring enlightenment to English grammar teaching and learning.

II. DATA COLLECTION AND COLLATION

In the stage of data collection and processing, 15 books with English Grammar as the theme were selected from the Gutenberg Project (https://www.gutenberg.org/). After deleting the Gutenberg sample chapter, the total number of English words was 1863536 and the text capacity was 11.4MB.

TABLE I. INFORMATION ON 15 ENGLISH GRAMMAR BOOKS

Year	Title
1812	A grammar of the English tongue
1829	English grammar in familiar lectures
1838	Lectures on Language as Particularly Connected with English Grammar
1845	The Comic English Grammar A New And Facetious Introduction To The English Tongue
1851	The grammar of English grammars
1895	An English grammar by Gilliam Malone basketball and james witt sewell
1896	Higer lessons in English
1896	Anglo-Saxon Grammar and Exercise Book with Inflections, Syntax, Selections for Reading, and Glossary
1899	Graded Lessons in English An Elementary English Grammar Consisting of One Hundred Practical Lessons, Carefully Graded and Adapted to the Class-Room
1904	English-Bisaya Grammar, in Twenty Eight Lessons
1905	Anglo-Saxon Primer, With Grammar, Notes, and Glossary
1906	Phrases and Names Their Origins and Meanings
1913	An advanced English grammar with exercises
1914	Practical Grammar and Composition
1918	Word study and English grammar

Data source: Collated by the author

TABLE II. TERM FREQUENCY DISTRIBUTION DESCRIPTIVE

Year	Sentences	Paragraphs	Types of Words(n)	Mean of TF	Std.Deviation of TF
1812	2161	1813	3975	2.99	13.85
1829	13352	10414	8846	7.04	60.23
1838	9347	6888	6922	6.81	49.92
1845	4698	3823	5963	3.91	27.81
1851	169147	93916	32390	20.75	314.02
1895	14290	10717	7907	7.02	52.49
1896	19121	13149	8466	8.13	62.97
1896	8928	7398	9416	5.19	43.88
1899	9866	6736	5265	6.59	41.67
1904	7192	4393	6417	4.05	22.49
1905	6248	5086	5811	8.32	147.33
1906	17282	11565	14645	5.03	32.72
1913	19635	13128	8127	8.84	70.39
1914	11961	8069	6149	6.88	53.95
1918	2785	2104	2826	4.24	22.16

Data source: Collated by the author

978-1-6654-9567-7/22/$31.00 ©2022 IEEE
DOI 10.1109/MLKE55170.2022.00064

III. WORD, PHRASES AND TOPIC ANALYSIS

A. *Total Word Frequency Results*

After excluding articles, prepositions and conjunctions in function words, the word frequency of content words is analyzed. Wordstat 9.0.4 is used to analyze the total word frequency and get the word cloud.

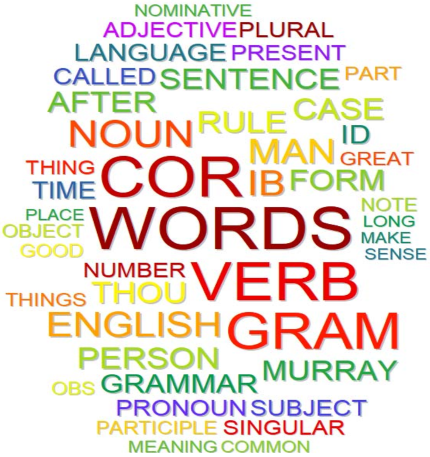


Figure 1. word cloud

TABLE III. WORD FREQUENCY

Word	Frequency	% shown	% processed	% total
Words	5244	1.56%	0.65%	0.28%
Cor	4969	1.48%	0.61%	0.26%
Verb	4752	1.41%	0.59%	0.25%
Gram	4409	1.31%	0.54%	0.23%
Noun	3212	0.95%	0.40%	0.17%
Ib	3069	0.91%	0.38%	0.16%
English	2983	0.89%	0.37%	0.16%
Man	2975	0.88%	0.37%	0.16%
Thou	2582	0.77%	0.32%	0.14%
Rule	2546	0.76%	0.31%	0.14%
Case	2529	0.75%	0.31%	0.13%
Person	2513	0.75%	0.31%	0.13%
Form	2458	0.73%	0.30%	0.13%
Sentence	2430	0.72%	0.30%	0.13%
After	2429	0.72%	0.30%	0.13%
Murray	2375	0.71%	0.29%	0.13%
Grammar	2259	0.67%	0.28%	0.12%
Id	2204	0.65%	0.27%	0.12%
Language	2100	0.62%	0.26%	0.11%
Time	2098	0.62%	0.26%	0.11%

Data source: Collated by the author

B. *Phrases Analysis*

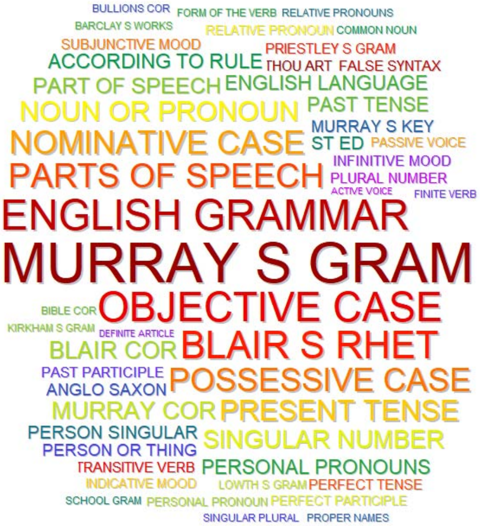


Figure 2. Phrases analysis

TABLE IV. PHRASES FREQUENCY

Phrases	Frequency
Murray’s gram	627
English grammar	470
objective case	450
Blair’s rhet	433
parts of speech	378
possessive case	366
nominative case	356
present tense	322
noun or pronoun	305
singular number	278
murray cor	276
blair cor	251
part of speech	239
past tense	223
English language	222
personal pronouns	213
according to rule	212
st ed	209
person singular	199
Murray’s key	184

Data source: Collated by the author

C. *Topics Analysis*

Extract the subject words from the text and describe the key words for each topic. Coherence (NPMI) within each topic was also calculated.

TABLE V. TOPICS ANALYSIS

Topic	Keywords	Coherence(Npmi)	Freq

English grammar st ed	mo; pp; ed; grammar; London; st; English;
	English grammar; st ed; English language;
		0.355	4377

transitive verb intransitive	transitive; intransitive; active; verbs;verb; transitive verb;transitive verbs; intransitive verbs;	0.222	3246
infinitive mood indicative;	mood;indicative; subjunctive; imperative; infinitive;infinitive mood;subjunctive mood;indicative mood;imperative mood;	0.218	1749
present tense perfect	tense;perfect; present;past; imperfect; participle; future; present tense; past tense; perfect tense;present perfect; past perfect; imperfect tense; past participle;	0.217	3983
objective case	case; objective; nominative; possessive; govern; objective case; possessive case; nominative case;	0.205	3827

Data source: Collated by the author

IV. CO-OCCURRENCES AND KEYWORD-IN-CONTEXT ANALYSIS

Co-occurrences and keyword-in-Context analysis belong to the Expert analysis mode in Wordstat.

A. *Co-Occurrences Analysis*

Options for Co-occurrence analysis include clustering using keywords/categories, Occurrence based on same Paragraph, index based on Association Strength, type based on Word co-occurrence-first order. At the same time, remove single word clusters. Sihoutte coefficient is an indicator to judge the degree of sample clustering. The average Sihoutte coefficient of the whole text is 0.710.

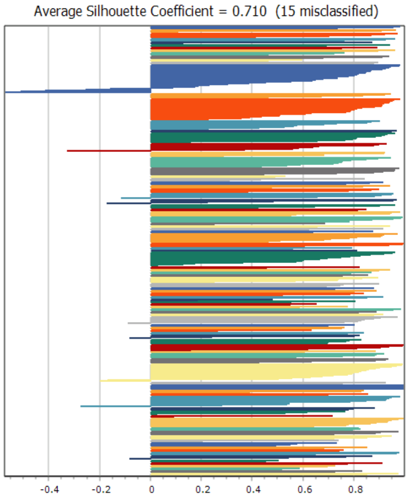


Figure 3. Sihouette coefficient chart

Dendrogram was obtained according to the agglomeration order. Because the Dendrogram graph is large, part of the content is captured. The words with high frequency were analyzed.

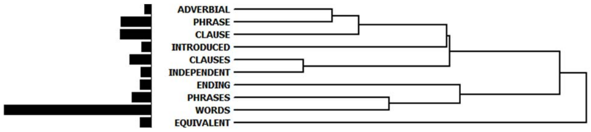


Figure 4. Partial Dendrogram graph

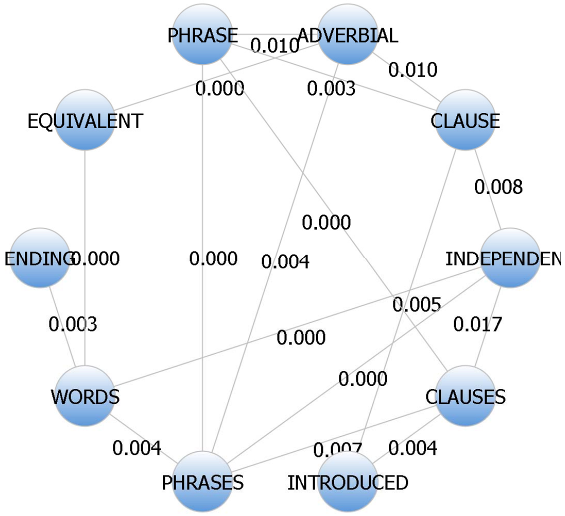
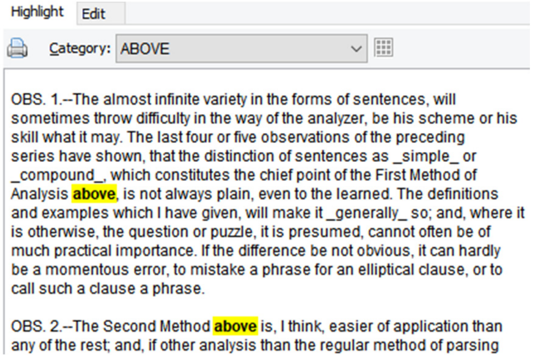


Figure 5. Link analysis

B. *Keyword in Context Analysis*

The keyword in context feature explores how each word is used in specific text. Take the word "above" as an example to show how it was used in grammar books in the 19th and 20th centuries.



Category: ABOVE

OBS. 1.--The almost infinite variety in the forms of sentences, will sometimes throw difficulty in the way of the analyzer, be his scheme or his skill what it may. The last four or five observations of the preceding series have shown, that the distinction of sentences as _simple_ or _compound_, which constitutes the chief point of the First Method of Analysis above, is not always plain, even to the learned. The definitions and examples which I have given, will make it _generally_ so; and, where it is otherwise, the question or puzzle, it is presumed, cannot often be of much practical importance. If the difference be not obvious, it can hardly be a momentous error, to mistake a phrase for an elliptical clause, or to call such a clause a phrase.

OBS. 2.--The Second Method above is, I think, easier of application than any of the rest; and, if other analysis than the regular method of parsing

Figure 6. Keyword highlight

V. PART OF SPEECH MINING AND ANALYSIS

Content words mainly include nouns, pronouns, adjectives, numerals, adverbs and verbs. This paper mainly analyzes four parts of speech in English grammar books, namely noun, adjective, adverb and verb. KH CODER 3 and Excel pivot table are used for statistical analysis.

A. Frequently Used Nouns

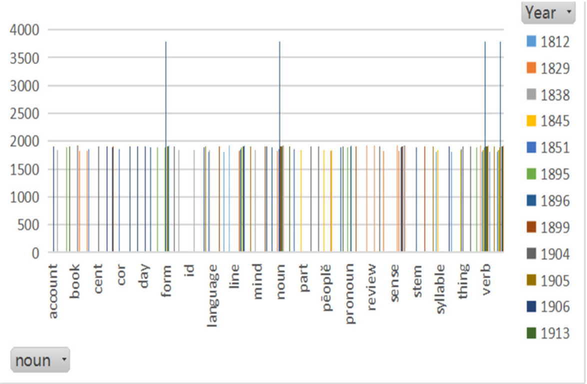


Figure 7. frequently used nouns

B. Frequently Used Adjectives

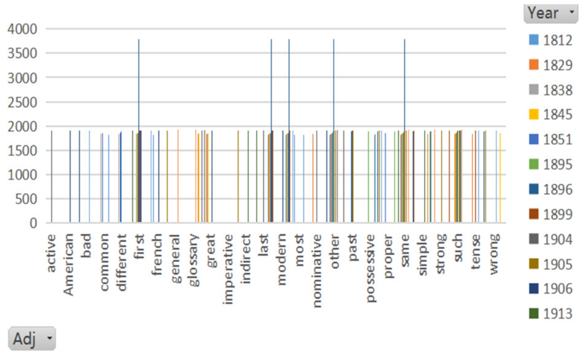


Figure 8. frequently used adjectives

C. Frequently Used Adverbs

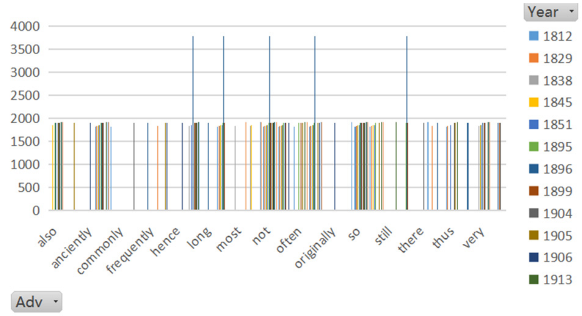


Figure 9. frequently used adverbs

D. Frequently Used Verbs

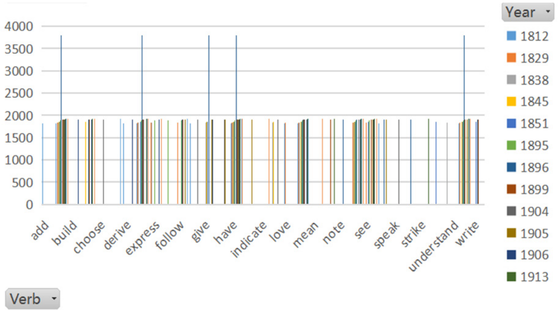


Figure 10. frequently used verbs

VI. CONCLUSIONS

In the 15 grammar books, the top 10 Words with the highest total frequency are words, cor, verb, gram, noun, ib, English, man, thou, rule. The most common English phrases are Murray’s gram, English grammar, objective case, Blair’s rhet, parts of speech, possessive case, nominative case, present tense, noun or pronoun and singular number. The most frequent topics are English grammar， transitive verb， intransitive， infinitive mood，indicative, present tense，perfect，objective case.

The average Sihoutte coefficient of the whole text is 0.710. Co-occurrence relationships between words can be analyzed by dendrogram and link analysis. The keyword query function can accurately find the specific application of a specific word.

In the part of speech analysis, the most commonly used nouns are account, book, cent, cor, day, form, id, language, line, mind, noun, part, people, pronoun, review, sense, stem, Syllable, a thing, the verb. In the part of speech analysis, the most commonly used adjectives are active, American, bad, common, different, first, French, general, Glossary, great, imperative, indirect, last, modern, most, nominative, other, past, possessive, proper, same, simple, strong, such, tense and wrong. The most common adverbs are also, anciently, commonly, frequently, hence, long, most, not, often, originally, so, still, there, thus and very. The most common verbs are add, build, choose, derive, express, follow, give, have, indicate, love, mean, note, see, speak, strike, understand and write.

REFERENCES

[1] Hu zhuanglin. “English grammar and its history of development”. Beijing, Journal of Beijing International Studies University, 2017, pp.5-16.

[2] Hardie, A. , A. Mcenery , and S. Piao . Historical Text Mining and Corpus-Based Approaches to the News books of the Common wealth. 2010.

[3] Krallinger, M. , et al. "Text Mining." Comprehensive Biomedical Physics 6.10 Supplement, 2014, pp.51-66.

[4] Meng, X. . "Text data management and analysis: a practical introduction to information retrieval and text mining. " Computing Reviews 58.4, 2017, pp.223-224.

[5] D B holat, et al. "Text Mining for Central Banks." SSRN Electronic Journal 33, 2015, pp.1-19.