Computational Analysis of Lexical Features of Online and Offline Teaching Interactive Discourse

Jiaqi Liu(✉)

School of International Studies, Zhejiang University, Hangzhou 310058, China
Liujq557@zju.edu.cn

Abstract. From the perspective of computational linguistics, this study establishes a corpus of online and offline interactive discourse, and studies the lexical features of interactive discourse in English teaching through computational methods, in order to improve the online education model and teaching interaction. It fins that the length, frequency of word and word clusters of interactive discourse are affected by the text type, network system and teaching mode. There is a certain correlation between online textual discourse and offline spoken interactive discourse, but their interaction efficiency is different. Based on the computational analysis, the research points out that the online teaching mode and the teaching interaction mode still need to be discussed, and the teaching interaction system needs to be further improved.

Keywords: Lexical feature · online English teaching · offline English teaching · interactive discourse · computational linguistics

1 Introduction

As a branch of computational linguistics, corpus linguistics focuses on the application principles and corpus research in the study of language. Computational Linguistics uses accurate measurement, observation, simulation, modeling and interpretation of linguistic phenomena to find out the mathematic laws behind languages (Liu & Huang, 2012) [1], through which the internal mechanism of language can be better explained.

Lexical features change under the different influences of different languages and genres (Liu 2022) [2]. Sinclair (2004) [3] proposed that word is the starting point for building lexical models. Lexical features include word length, frequency and others which may have some effects on lexical collocation (Liu 2022) [2]. Ellis (2008) [4] classifies teacher-student interaction into the category of interaction purpose, believing that the purpose of interaction is to improve the efficiency of teaching and learning.

Nowadays, comparative studies based on computational linguistics are not enough, especially those using corpus method. Therefore, in order to find out the common and different features between various modes of interactive discourse, this paper builds corpus and uses computational methods to analyze the lexical features, and tries to answer:

D. Kumar et al. (Eds.): IEIT 2023, AHSSEH 10, pp. 701–707, 2023.
https://doi.org/10.2991/978-94-6463-230-9_84

Table 1. Some statistics of OFFIDC and OIDC

	OFFIDC	OIDC
tokens	60,189	19,830
types	3,299	1,416
Type-Token Ratio (TTR)	5.48%	7.14%
Average word length	5.27	3.89
Average sentence length	17.26	21.70

What are the distribution characteristics of lexical features of offline interactive discourse and online interactive discourse? Do they follow certain laws of models? What are the inspirations for online education related systems and models?

2 Data and Methods

The study searched college English courses through the text sorting function of a university's Zhiyun Class, and eliminated specialized and elective courses, as well as courses without live broadcast. Finally, 20 classes from September 2021 to September 2022 were collected, including 10 traditional classes and 10 online classes.

After automatic processing, manual proofreading and correction of the annotated text, data analysis software AntConc was used for further analysis. SPSS software was used to analyze Kendall's tau-b rank correlation coefficient. The basic information of the offline interactive discourse corpus (OFFIDC) and the online interactive discourse corpus (OIDC) is as follows:

In the Table 1, the TTR of OIDC is higher than OFFIDC, but the average word and sentence length are lower than OFFIDC. TTR is a commonly used method to compare the variation of lexical density in corpus linguistics, so the lexical density of OIDC is higher than OFFIDC. It can be preliminarily speculated that the offline interactive discourse tends to use more language units, with simplified vocabulary and low sentence complexity.

3 Results

3.1 Word Length

The difficulty of the discourse or text can be presented by the length of word, and it can also show the complexity of text's language unit. Due to the cognition and information processing mechanism of brain, human always try to use short phrases or simple words to make a conversation, which can make the conversation more concise. (Deng & Feng 2013: 37) [5].

Great progress has been made in the research of word length, as well as word frequency and their relationship, such as Chen & Liu (2014) [6] conducted a diachronic study on the distribution of the length of Chinese words. However, the studies on the

distribution features in offline spoken and online interactive discourses are not enough. From Table 1, it can be seen that the online interactive text's average word length is lower than that of offline interactive spoken text, and Table 2 presents the statistical information of the distribution of word length.

In Table 2, the economic principles of language can be presented by the relationship between the corpus's word length and frequency. In addition to genre differences, teaching models are also quite different, so online word length data can't be totally compared with the offline word length data in all directions. Therefore, this section focuses on analyzing its correlation. As shown in Table 3, the distribution of word length and frequency of OFFIDC and OIDC are significantly correlated ($p < 0.001$).

In Fig. 1, the word length distribution is in accordance with the labor-saving principle. The percentage of 5 to 7-letter word length of OFFIDC is higher than that of OIDC. Before the 5-letter word length, it always shows an increasing trend, but after the 5-letter word length, it shows a decreasing trend, without obvious fluctuation, and tends to 0 from the 14-letter word length. In OIDC, 3-letter words had the highest proportion, and showed an increasing trend before the highest point. After the 3-letter words, it showed a decreasing trend, and from the 7-letter words, it showed a stable decreasing trend.

Table 2. The statistical information of the distribution of word length

	OFFIDC	OIDC
1-letter words	747	403
2-letter words	3,991	3,196
3-letter words	4,033	3,978
4-letter words	8,110	2,402
5-letter words	13,010	2,044
6-letter words	11,209	1,865
7-letter words	8,102	1,906
8-letter words	4,706	1,508
9-letter words	2,894	1,007
10-letter words	1,843	735
11-letter words	697	359
12-letter words	332	204
13-letter words	431	150
14-letter words	72	37
15-letter words	0	32
16-letter words	12	3
17-letter words	0	1

Table 3. The results of correlation analysis

			OFFIDC	OIDC
Kendall's tau-b	OFFIDC	Correlation Coefficient	1.000	.775**
		Sig. (2-ailed)	.	.000
		N	17	17
	OIDC	Correlation Coefficient	.775**	1.000
		Sig. (2-ailed)	.000	.
		N	17	17

**. Correlation is significant at the 0.01 level (2-tailed).

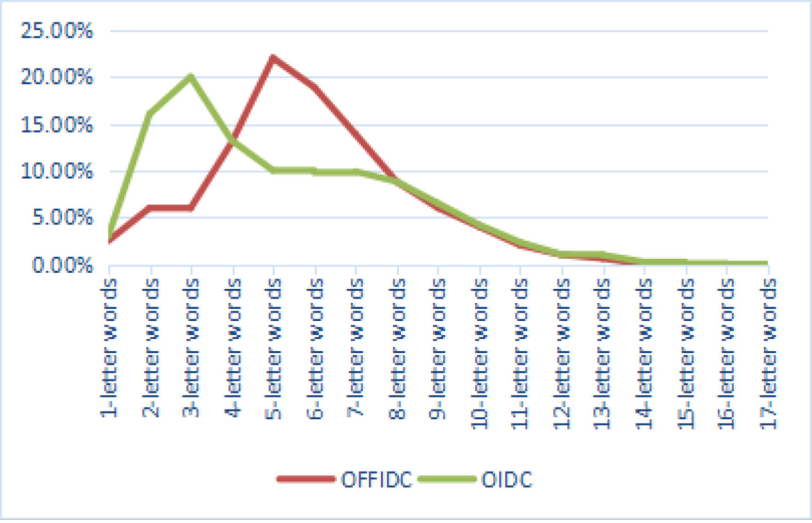

Fig. 1. Percentage distribution of word length

3.2 Word Clusters

Lewis (1993) [7] proposed that word cluster can be seen as the basic structure of language, rather than traditional grammar and vocabulary. Word clusters can be composed of several words that may not be complete in structure or meaning (Chen, 2014) [8], but it has specific discourse functions, which can play a significant role in language output and reflect the features of language reuse. This section takes 2–7 word clusters as the research object.

In Fig. 2, there are certain differences between the 2 to 7 word clusters with a frequency of more than 10 times in the two corpora. Although the number of words in OFFIDC is much larger than that of OIDC, there is no significant difference in the word clusters exceeding 10. Detailed information of word clusters with 2 to 7 words with frequencies greater than 10 times are further analyzed in Table 4.

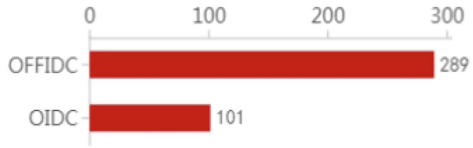

Fig. 2. Word clusters with frequencies greater than 10

Table 4. Numbers of word clusters with frequencies greater than 10

	OFFIDC	**OIDC**
word clusters of 2 words	79	7
word clusters of 3 words	89	14
word clusters of 4 words	65	47
word clusters of 5 words	37	19
word clusters of 6 words	17	13
word clusters of 7 words	2	1
Total	289	101

From Table 4 we can see that word clusters in OIDC is not as much as OFFIDC, but both of them focused on the 2–5 word cluster segments. If a word cluster contains fewer words, the number of word clusters will be more. Analyzing word cluster an effective method to discuss the characteristics of the text and judge the change of the text. There are many differences of the distribution of OIDC and OFFIDC. It can be seen that offline interactive spoken discourse tends to use shorter word clusters than online interactive text discourse. Compared with offline interaction, online interaction is more likely to use repetitive language, and this table also presents that the long language fragments will have more repetition if they contain content words in offline text interaction.

4 Discussion

As an important part of lexical features, word length can effectively measure the difficulty of discourse and is also one of the effective criteria for determining the complexity of language units. The results show that as for online and offline teaching texts, word length and frequency are affected by teachers' styles and teaching modes. Moreover, long words which contain many letters are hardly appear in the teaching discourses.

Word clusters can effectively reflect lexical and grammatical characteristics, which can also measure some features of the phenomenon of reuse (Wray, 2000) [9]. The research presents that the word clusters' frequency of online and offline teaching texts is almost the same. If a word cluster contains fewer words, the number of word clusters will be more.

Online text interaction tends to use longer word clusters, while offline spoken interaction is more prone to repetitive language fragments. Because text content is somewhat

more regulated than verbal interactions, written language is used more often to regulate specific words or phrases. While oral interactive language is relatively free and casual with more simple and short words, so that it is easier for the listener to quickly grasp and respond.

Online teaching text interaction takes time to browse and understand the meaning, which is easy to delay the teaching progress and affect the teaching efficiency to a certain extent. Therefore, compared with offline traditional teaching interaction, online interaction has less frequency, but it can also get a complete and clear interactive response, and improve the learning efficiency of individual interactive subjects. At the same time, offline interaction has more immediacy, which can extract effective interactive information faster and more accurately, and exchange teaching feedback in time with simple and clear words, so as to improve the efficiency of overall teaching.

5 Conclusion

By means of the relevant functions of corpus and corpus analysis software, this paper conducts a comparative study and finds that the interactive discourse text of online and offline English teaching have certain similarities and differences in terms of the length, frequency of word, word clusters and other characteristics, which will have a certain impact on the teaching efficiency.

This research tries to combine the lexical features with interactive discourse, enriches the study of language vocabulary features and language education level features, and provides some reference for related research. Although the differences and commonalities of some interactive discourse of different teaching modes can be discussed through the difference of lexical features, further exploration is needed to clarify more linguistic complexity and teaching relevance between the two. In addition, how to improve the online teaching interaction system, improve the interactive efficiency of the online teaching system, on the basis of traditional teaching interaction, to provide necessary technical support, still need to be further thought.

References

1. Liu, H. T. and Huang, W. (2012). Quantitative Linguistics: State of the Art, Theories and Methods. *Journal of Zhejiang University (Humanities and Social Sciences)*, (2): 178-192.
2. Liu, J. Q.(2022). Lexical Features of Economic Legal Policy and News in China Since the COVID-19 Outbreak. *Front Public Health.* 10:928965. doi: https://doi.org/10.3389/fpubh.2022.928965.
3. Sinclair, J. M. (2004). *Trust the Text: Language, Corpus and Discourse.* London/New York: Routledge.
4. Ellis, R. (2008). The study of second language acquisition (2nd ed.). Oxford: Oxford University Press.
5. Deng, Y. C. and Feng, Z. W. (2013). A Quantitative Linguistic Study on the Relationship between Word Length and Word Frequency. *Journal of Foreign Languages*, (3): 29-39.
6. Chen, H. and Liu, H. T. (2014). A Diachronic Study of Chinese Word Length Distribution. *Glottometrics*, (29): 81-94.

7. Lewis, M. (1993). *The Lexical Approach: The State of ELT and a Way Forward*. Language Teaching Publications (LTP).
8. Chen, C. (2014). Analysis of the Characteristics and Functions of Word Clusters in Alice Munro's Novels -- A Corpus based stylistic study. *Journal of PLA University of Foreign Languages*, 37(03): 151-159.
9. Wray, A. (2000). Formulaic Sequence in Second Language Teaching: Principle and Practice. *Applied Linguistics*, 21(4): 463-489.