A Study of Lexical Chunks Based on Linguistic and Medical Dissertation Corpora

WU Yanxia [1, a], Song Weicai [2,b]

[1] Institute of humanities, Jiangxi University of Traditional Chinese Medicine, Nan chang, 330004, China

[2] Institute of Computer Science, Jiangxi University of Traditional Chinese Medicine, Nan chang, 330004, China

[a]email: wuyanxia2000@sina.com

Corresponding Author, Song Weicai

Keywords: Linguistic dissertation corpus; Medical dissertation corpus; Lexical Chunks

Abstract. Lexical chunks abound in the academic domain. For language learners, the effective storage and proper use of lexical chunks are a prerequisite for improving language cognitive ability and expressive ability. The use of lexical chunks in authoritative academic journals should be regarded as an example of the use of lexical chunks. This study is based on the use of linguistic and medical dissertation corpora to compare linguistics and medical academic English lexical chunks, and analyze its similarities and differences in lexical structure fixation, structural form and discourse function in order to reveal the usage pattern of academic lexical chunks.

Introduction

Lexical chunks abound in the academic domain (Hyland 2008), such as "make the best of, be bound to, get the better of, on behalf of" etc which can promote the language expression native and language users as "insiders" in a certain language community. Therefore, for language learners, the effective storage and proper use of lexical chunks are a prerequisite for improving language cognitive ability and expressive ability. The use of lexical chunks in authoritative academic journals should be regarded as an example of the use of lexical chunks. The lexical chunks teaching based on this can make the learners quick to master the correct language expression method, so as to be handy in academic exchange. It can be seen that it is necessary to describe and summarize the academic English lexical chunks in different disciplines.

The lexical chunks, also known as the multiple word sequence, the repeating phrase, the word cluster, the prefabricated chunks, the N tuple, etc., refers to the recurring idiomatic string in the natural language (Biber & Conrad 1999: 90). For more than a decade, scholars at home and abroad based on corpora have mostly studied the lexical chunks by quantitative frequency statistics, but the research focuses are different. Foreign research focuses on the extraction of unique chunks of specific disciplines, to explore the discourse function of the lexical chunks. Cortes (2004) contrasts the structure and function of the lexical chunks in the historical and biology journal; Hyland (2008) points out that there are disciplinary differences in different academic lexical chunks; Durrant (2009) confirms the possibility of the compilation of common high-frequency phrases in different academic fields; Martine etal. (2009) extracted the lexical chunks in agricultural academic papers; Simpson-Vlach & Ellis (2010) has developed an interdisciplinary high-frequency academic lexicon covering four disciplines of humanities, arts, social sciences, natural sciences and medicine, science and technology and engineering.

China's research focuses on the actual situation of the use of English learners' lexical chunks and the influencing factors of lexical chunks learning. Ding Yan and Qi Yan(2005) examines how the three senior English learners can improve the fluency of the expression through the acquisition of the lexical chunks; Xu Fang (2011) explores the variety of lexical chunks in bachelor's, master's and doctor’s dissertations by comparing the domestic and international scholars' dissertation corpora;

Wang Min and Liu Ding (2013) compared the similarities and differences between Chinese English learners and foreign experts in the use of position markers.

In general, researches at home and abroad usually quantify and describe fixed lexical chunks, comparative studies of interdisciplinary chunks have just begun, and the discussion of fixed and semi-fixed levels of lexical chunks is rare (Staples et al. 2013) . It is important to study the formal structure of the lexical chunks, but it is also indispensable to study the fixedness of the lexical structure of different disciplines for a more in-depth and meticulous understanding of the characteristics of the interdisciplinary chunks. This study is based on the use of linguistic and medical dissertation corpora to compare linguistics and medical academic English lexical chunks, and analyze its similarities and differences in lexical structure fixation, structural form and discourse function in order to reveal the usage pattern of academic lexical chunks and provide some reference and guidance for the academic exchanges in the EAP (English for Academic Purposes) teaching and related areas. Specific research questions are as follows:

(1) How are the lexical chunks in the two corpora distributed?

(2) What are the most frequently used lexical chunks in the two corpora? Are there significant differences in usage?

(3) What are the characteristics of structural fixation and formal structure of high frequency four-word lexical chunks in two corpora? What are the discursive functions of these four-word phrases?

Research framework

This study is divided into three steps: creating corpus; extracting, sorting, perfecting lexical chunks list; analyzing and discussing the results. The research framework is as follows:

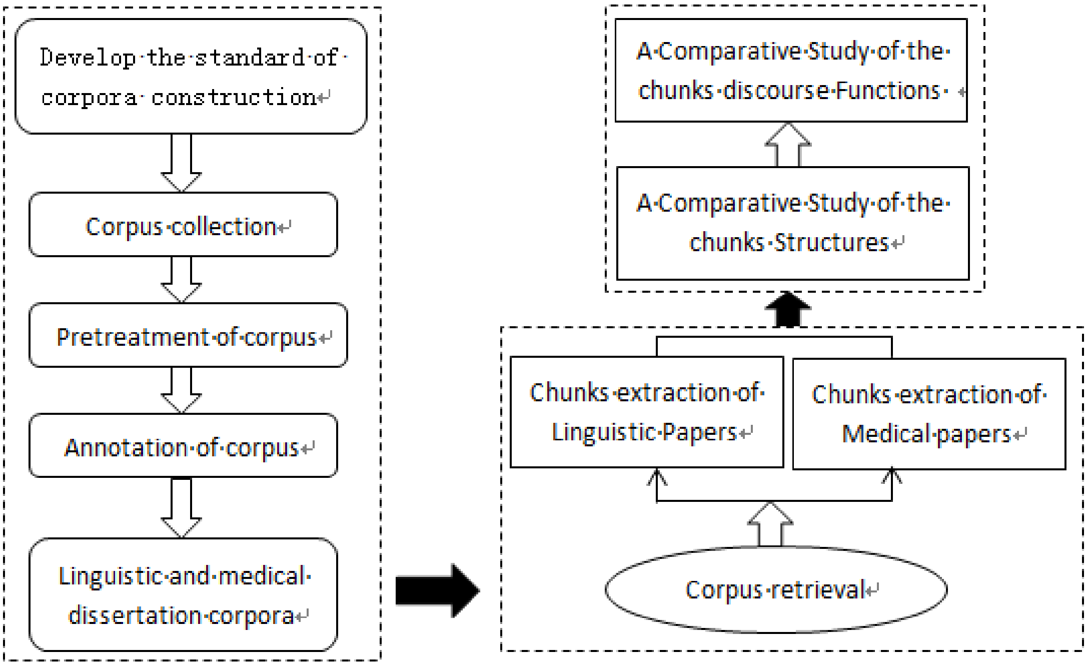

Develop the standard of corpora construction
Corpus collection
Pretreatment of corpus
Annotation of corpus
Linguistic and medical dissertation corpora
A Comparative Study of the chunks discourse Functions
A Comparative Study of the chunks Structures
Chunks extraction of Linguistic Papers
Chunks extraction of Medical papers
Corpus retrieval

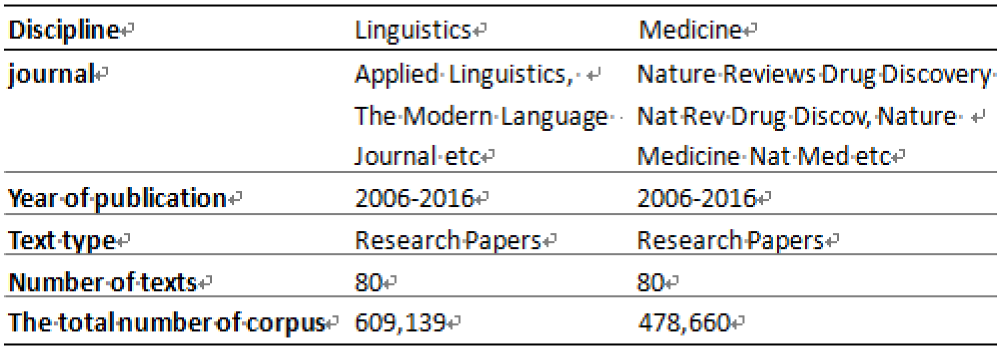
Fig.1 Research framework

Construction of linguistic and medical dissertation corpora

(1) The construction of corpus

This study uses two corpora of different disciplines, namely linguistic dissertation corpus and medical dissertation corpus. The former consists of linguistic SSCI journal articles, and the latter covers the international authoritative journals of Chinese medicine and Western medicine. As a representative of ordinary English and specialized English, the use of lexical chunks in the linguistics and medical disciplines can reflect the different cognitive, research methods and thesis writing methods. Linguistics Research Articles Corpus (LC) is composed of 80 academic papers randomly selected from eight kinds of linguistics SSCI journals. Medical Research Articles Corpus (MC) is composed of 80 papers from eight of the world's top academic journals recommended by TCM and Western medicine professionals. The first author of LC and MC papers must be an English-speaking scholars, or the academies and scientific research institutions of the papers are English-speaking countries. All papers are processed, with bibliographies, appendices, footnotes, acknowledgement, and charts deleted. Presents the specific situation of two corpus:

Table 1: Linguistics and medical dissertation corpora



Discipline	Linguistics	Medicine
journal	Applied Linguistics, The Modern Language Journal etc	Nature Reviews Drug Discovery Nat Rev Drug Discov, Nature Medicine Nat Med etc
Year of publication	2006-2016	2006-2016
Text type	Research Papers	Research Papers
Number of texts	80	80
The total number of corpus	609,139	478,660

(2) Corpus-based lexical chunks study

Before the lexical chunks are extracted, the length of the lexical chunks is determined first. This study mainly examines the four-word chunks in the academic discourse, since the vast majority of the four-word chunks themselves contain three-word chunks, such as "as a result of" containing "as a result". Furthermore, the four-word chunks are more frequent than the five-word chunks, and the structure and function are more extensive (Cortes 2004). Extraction and Frequency Statistics of the lexical chunks use the AntConc3.2 corpus retrieval and statistics tool, with the intercept frequency of 20 times per million words. In order to ensure that the extracted lexical chunks reflect the disciplinary characteristics rather than the author's personal style, the lexical range is defined as those lexical chunks that appear in at least 1/10 of the text in the respective corpus. The lexical chunks extraction is divided into five steps:

Table 2: The lexical chunks extraction steps

Step 1	Extract four-word lexical chunks. The minimum frequency is set to 12 times and 11 times, which is based on LC and MC storage capacity, and are converted from the standard conversion frequency of 20 times per million words.
Step 2	View the distribution rate. Use the Concordance feature to view and delete the lexical chunks with a distribution rate less than 1/10 of the LC or MC text.
Step 3	The artificial screening. To avoid the effect of overlapping chunks on statistical results, use the Concordance function to view and filter out four different four-word lexical chunks contained in the same five or six four-word lexical chunks.
Step 4	Use the Vlookup function of the Excel table's to compare the LC and MC lexical chunks, removing the coincident lexical chunks.
Step 5	Compares the resulted lexical chunks with the interdisciplinary high-frequency academic English lexical chunks of Simpson-Vlach & Ellis (2010), and removes the coincident lexical chunks in the fourth step to ensure that the final extracted lexical chunks are the specific lexical blocks with academic characteristics, and finally the LC and MC lexical scales are obtained.

(3) The comparative study of the lexical chunks structure and function

This study is to be compared from the following three aspects:

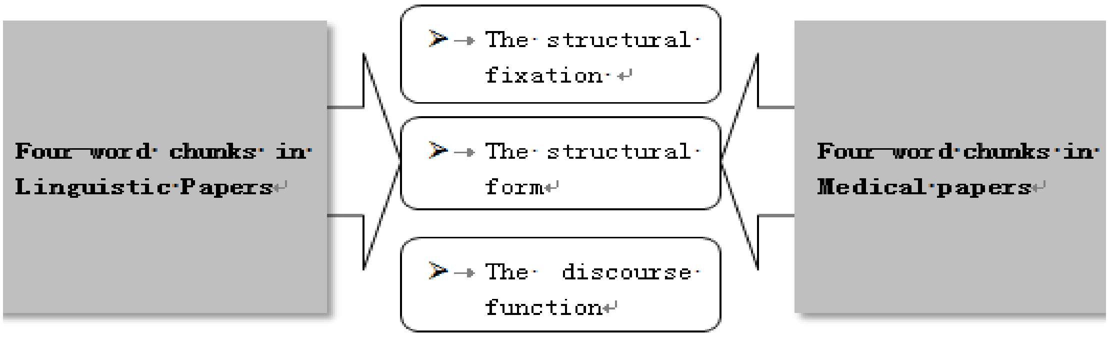

Four word chunks in Linguistic Papers
The structural fixation
The structural form
The discourse function
Four word chunks in Medical papers

Fig.2 Steps of lexical chunks comparison

1) The structural fixation of the lexical chunks

The internal structure of the lexical chunks is divided into a fixed continuous sequence and a semi-fixed combination of internal changes. Continuous sequence refers to a fixed lexical chunks of four words or consecutive three words or two consecutive words; internal changing lexical chunks refers to those lexical chunks with the first and fourth word fixed and the second or third word not fixed, or with the first and third word or the second and fourth word fixed and other words change, or only a word fixed lexical chunks (Biber 2009: 292). The purpose of detecting the structural stability of the lexical chunks is to investigate whether the lexical chunks is a fixed continuous sequence or a semi-fixed internal changing structure. Fixedness detection of lexical chunks follows Biber (2009) Criteria: If a word in a lexeme has a frequency greater than or equal to 50% in the same class, this term is a fixed lexeme.

2) The structural form of the lexical chunks

Biber (2009) found that although the lexical chunks is not a complete grammar unit, but its use and grammatical structure has a great relationship. Based on the study of the structural fixation of the lexical chunks, this study will compare the differences and commonalities of linguistics and medical English words in structural form and fixation.

3) The discourse function of the lexical chunks

Biber (2009) divides the lexical chunks into three categories according to discourse and pragmatic functions: indicator composition, position lexical chunks and discourse lexical chunks. The indicator lexical chunks labels an objective or abstract object or even the text itself, either to demonstrate the object itself or the object's characteristics. The position lexical chunks indicate the speaker's attitude or evaluation for the speech content. Discourse lexical chunks organize discourse or text, indicating the relationship between adjacent discourses and so on. Accordingly, we will discuss the close relationship between the two different types of lexical chunks structure and the text function.

Summary

A conclusion is gotten on the comparison above:

(1) Linguistics and medical scholars prefer to use a higher productivity semi-fixed lexical chunks.

(2) There are significant differences in the species and frequency of the two most commonly used lexical chunks in the two corpora. In the two corpora, the number of nouns and prepositions is mostly used. The use of phrasal lexical chunks is more than that of sentence fragments. However, the proportion of phrasal lexical chunks and sentence fragments in the two corpora are not the same.

(3) The high frequency four-word lexical chunks are similar in the structural form, and there are differences in the structure fixation, and the structure of the lexical chunks is closely related to the discourse function. Through the analysis of the structure of the lexical chunks, we can find that each discourse function of the lexical chunk is expressed as the relevant structural form.

Acknowledgements

This work is supported by Research on Humanities and Social Sciences in Jiangxi Province (GL1329).

References

[1] Biber D ＆ Conrad S. The Longman Grammar of Spoken and Written English［M］. London: Longman,1999.

[2] Cortes V. Lexical bundles in published and student disciplinary writing: Examples from history and biology［J］. English for Specific Purposes, 2004, 23(4): 397—423.

[3] Durrant P. Investigating the viability of a collocation list for students of English for academic purposes［J］. English for Specific Purposes, 2009, 28(3): 157—169.

[4] Hyland K. As can be seen: Lexical bundles and disciplinary variation ［J］. English for Specific Purposes, 2008, 27(1):4—21.

[5] Larsen-Freeman D. A complexity theory approach to second language development/acquisition［A］. In Atkin-son D (ed.). Alternative Approaches to Second Language Acquisition［C］. London: Routledge, 2011.

[6] Martinez I A,Beck S C & Panza C B. Academic vocabulary in agriculture research articles: A corpus- based study［J］. English for Specific Purposes, 2009, 28(3):183—198.

[7] Simpson- Vlach R & Ellis N C. An Academic Formulas List: New methods in phraseology research［J］. Applied Linguistics, 2010, 31(4):487—512.

[8] Staples S et al. Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section［J］. Journal of English for Academic Purposes, 2013,12(3): 214—225.