CSCanada
Studies in Literature and Language
Vol. 11, No. 5, 2015, pp. 24-29
DOI:10.3968/7885

ISSN 1923-1555[Print]
ISSN 1923-1563[Online]
www.cscanada.net
www.cscanada.org

A Corpus-Based Analysis of N-Grams in English Texts Written by Chinese Learners

LIN Zhiyuan[a],*

[a]School of Foreign Studies, Henan Polytechnic University, Jiaozuo, Henan, China.
*Corresponding author.

Supported by Research Project of Education Department, Henan Province, China "Humanities and Social Science" (2014-qn-596).

Received 24 August 2015; accepted 6 October 2015
Published online 26 November 2015

Abstract

Recent linguistic studies have shown that proper use of multi-word expressions, or n-grams, is quite essential in foreign language acquisition. By contrast of the different types of n-grams from native corpus and Chinese learners' corpus, the present article is devoted to discussing distinctive features of n-grams use in English texts written by Chinese learners. Generally, Chinese learners have not developed a strong awareness of distinguishing written style from spoken style of English texts. Also they tend to produce various n-grams which reflect the Chinese social development. Besides, they usually overuse n-grams of "a personal pronoun + a modal or mental verb" to stress human subjective ideas, attitudes or feelings toward the objective world. In addition, they can not output n-grams about quantity in a native-like way. The article also explores the causes of those peculiar features above mentioned from linguistic, social, cultural and cognitive perspectives. The study may not only helps readers get a better understanding about the n-gram features involved in Chinese learners' English texts, but also brings some practical implications to EFL teaching.

Key words: Chinese learners; Written English; N-Grams; EFL teaching

Lin, Z. Y. (2015). A Corpus-Based Analysis of N-Grams in English Texts Written by Chinese Learners. *Studies in Literature and Language, 11*(5), 24-29. Available from: http://www.cscanada.net/index.php/sll/article/view/7885
DOI: http://dx.doi.org/10.3968/7885

INTRODUCTION

In the field of corpus linguistics, n-grams are usually defined as frequently occurring contiguous words that constitute a phrase or a pattern of use (e.g. *you know, on the, there is a, with the development of*). Two-word n-grams may also be referred to as bi-grams, three-word n-grams as tri-grams and so on. Sometimes, they are also termed lexical bundles, lexical phrases, clusters or chunks, which have attracted increasing attention from the academic circle. Studies in psychology, psycholinguistics, neurolinguistics, and Second Language Acquisition show that native speakers can produce complicated sentences in a quite smooth way, not because they have accumulated a large number of isolated words in their minds, but because they have stored many useful multiword units. Those units are usually extracted as a whole so that more time can be spared to be focused on the selection of sentence patterns and integration of meanings (Liang et al., 2010, p.14). Concerning traditional English learning in China, learners are usually affected by traditional linguistic theories and teaching modes, and therefore separate vocabulary from grammar. In this sense, grammar is considered to be a close system that determines the linguistic structure, which has great power of generation, generalization, and interpretation, while vocabulary is viewed as mixed and disorderly materials used to fill in the structure (Wei, 2011, p.12). Accordingly, grammar is placed in the first place and vocabulary second, which is cut apart from the former. However, vocabulary is actually no less important than grammar. What is more, one word does not exist by itself only, but is collocated with other words to produce a more specific meaning. The combination formed regularly or irregularly is nothing but the multiword unit, or n-gram in linguistic studies.

When it comes to n-grams, the recent studies were mainly focused on the grammatical structure, discourse

functions, features of specific style of texts (Biber et al., 1999; Cortes, 2002; Carter & McCarthy, 2006; Scott & Tribble, 2006; O'Keeffe et al., 2007; Hyland, 2008). However, those studies were mainly based on native languages rather than foreign languages. The present article is aimed to explore the general features of n-grams in English texts written by Chinese learners by means of a corpus-based approach, in the hope of revealing the regular rules and patterns in Chinese English and factors affecting English learning.

1. GENERAL DESIGN OF THE STUDY

1.1 Research Objective and Method

In the present study the corpus-based method is used to explore the features of n-grams in the written section of the Spoken and Written English Corpus of Chinese Learners 1.0 (Wen et al., 2009). By comparison between the learners' corpus and a native corpus and investigation of statistical evidence of various sorts of n-grams, the paper is aimed to disclose the typical features, grammatical structure and discourse functions of n-grams in English texts written by Chinese learners.

In the study, two questions will be investigated. First, compared with native people, what are the special features of n-grams that foreign language learners produce? Second, when foreign language learners produce n-grams, do they overuse, underuse or misuse them? And if yes, what are the causes?

1.2 Research Data

There are two corpora involved in the study. One is the written section of the Spoken and Written English Corpus of Chinese Learners 1.0, shortened as WECCL. The corpus covers 3,880 English texts, which were written by English majors from nine different types of colleges and universities, with the size of 1,254,123 million words. The texts are actually the writing assignments for the students with different titles, most of which are argumentative essays (Wen, 2009, p.5). The other is the native corpus of NS_WRITTEN, shortened as NS, used as a reference corpus in the study, which contains 9 texts with 1,557,835 words (Liang et al., 2010).

1.3 Research Tools

Two major tools are used to in the study. One is the AntConc 3.2.3, which is a portable cross-platform data processing software developed by Laurence Anthony from Faculty of Science and Engineering, Waseda University, Japan. The software is the combinations of multiple functions such as concordance, wordlist, key wordlist, collocation and n-grams. The other is the statistical tool of corpus data, Microsoft Office Excel 2003, which is mainly used in classification and calculation of data.

2. A STATISTICAL REPORT OF THE STUDY

With the help of the cluster function of the software AntConc, 2-grams, 3-grams and 4-grams in the two corpora WECCL and NS will be respectively retrieved and closely analyzed. However, we do not have enough space to discuss all of them but to choose the 30 most frequently used 2-grams, 3-grams and 4-grams as the representative data. About each type of the n-grams, the n-gram forms, standardized occurrence frequency of each n-gram per one million words (shortened as St. Fr.), and their ranks are given in different tables in the following sections.

2.1 2-Grams

2-grams are the combinations of any two consecutive words retrieved by computers. Since each 2-gram contains only two words, it is with the highest frequency among all the types of n-grams. The following table shows the 30 most frequently used 2-grams retrieved from the native corpus NS_WRITTEN on the left column and the learners' corpus WECCL on the right.

Table 1
The 30 Most Frequently Used 2-Grams Retrieved From NS and WECCL

Rank	St. Fr.	2-Grams (NS)	Rank	St. Fr.	2-Grams (WECCL)
1	10002	of the	1	3537	of the
2	5818	in the	2	3027	in the
3	3525	to the	3	2107	to the
4	2801	and the	4	2017	it is
5	2346	on the	5	1919	is a
6	2151	to be	6	1728	don 't
7	1876	for the	7	1534	we can
8	1839	that the	8	1522	and more
9	1661	by the	9	1483	I think
10	1639	from the	10	1444	on the
11	1633	of a	11	1426	is the
12	1585	with the	12	1330	more and
13	1455	at the	13	1329	want to
14	1302	it is	14	1311	can 't
15	1255	in a	15	1290	to be
16	1095	as a	16	1228	It is
17	903	with a	17	1227	with the
18	888	it was	18	1124	is not
19	857	have been	19	1099	you can
20	851	as the	20	1016	it 's
21	782	to a	21	1001	a lot
22	777	had been	22	990	we should
23	733	is a	23	975	and the
24	729	and a	24	966	each other
25	709	for a	25	960	will be
26	685	into the	26	942	high education
27	684	will be	27	940	computer games
28	657	has been	28	934	the world
29	656	can be	29	902	for the
30	623	the same	30	889	they are

From the Table 1, the following three points can be easily found.

Firstly, a lot of abbreviated 2-grams are frequently used in the learner's corpus WECCL, such as "don't", "can't", "it's". Their standardized frequencies per 1 million words are respectively 1,728, 1,311, and 1,016, and their ranks in the corpus are 6, 14, and 20. However, there is no single such type of abbreviated 2-grams in the native corpus NS. If more data is investigated, more types of such abbreviations will be easily found. Generally speaking, these abbreviations are used only in the informal spoken language rather than the formal written language. Therefore, it indicates that Chinese learners have not yet had a good command of the differences of styles between written and spoken English texts.

Secondly, content 2-grams with definite lexical meanings are far more frequently used in WECCL than in NS, for instance, "high education", "computer games", "the world", "mobile phone", "the internet", "their children". On the contrary, almost all of the 30 most frequently used 2-grams in the native corpus NS are functional. It can be easily seen that the most frequently used 2-grams in native corpus have little to do with the practical content of texts, while those in Chinese learners' corpus are highly related to the subjects of texts.

Thirdly, 2-grams of a special type are frequently used in the Chinese learners' corpus WECCL, such as "we can", "I think", "you can", "we should" and etc. This type of 2-grams can be referred to as "a personal pronoun (especially the first personal pronoun) + a modal or mental verb" pattern, which Chinese learners usually use to express ideas, abilities or responsibilities of themselves or others.

2.2 3-Grams

3-grams are formed by groups of three consecutive words, and their structures tend to be clearer and more fixed, which reflect to a larger extent their functions and cognitive features of users. Therefore, they are usually the major research objectives of scholars. The table below shows the data of the 30 most frequently used 3-grams from the corpora of NS and WECCL.

Table 2
The 30 Most Frequently Used 3-Grams Retrieved From NS and WECCL

Rank	St. Fr.	3-Grams (NS)	Rank	St. Fr.	3-Grams (WECCL)
1	335	one of the	1	1270	more and more
2	272	the end of	2	643	the development of
3	261	as well as	3	628	a lot of
4	211	part of the	4	517	In my opinion
5	202	some of the	5	388	I don t
6	198	end of the	6	388	to communicate with
7	197	the fact that	7	364	a degree certificate
8	180	in order to	8	364	with each other

To be continued

Continued

Rank	St. Fr.	3-Grams (NS)	Rank	St. Fr.	3-Grams (WECCL)
9	166	out of the	9	341	in order to
10	160	per cent of	10	337	is more important
11	158	there is a	11	336	parents and children
12	156	a number of	12	325	and so on
13	146	be able to	13	319	degree and certificates
14	130	because of the	14	307	the most important
15	126	that it is	15	291	to deal with
16	126	in terms of	16	281	a lifelong process
17	126	to be a	17	278	and more people
18	122	the same time	18	269	In a word
19	119	the number of	19	261	as well as
20	117	and it is	20	259	you want to
21	113	it is not	21	251	more attention to
22	113	there is no	22	242	the other hand
23	111	the use of	23	239	in the world
24	109	can not be	24	234	for us to
25	108	and in the	25	234	of the Internet
26	107	at the end	26	232	people think that
27	104	in which the	27	221	we can t
28	103	that it was	28	217	the mobile phone
29	102	have to be	29	214	the same time
30	101	side of the	30	211	I think the

Firstly, 3-grams that indicate tendency of social changes are frequently seen in the Chinese learners' corpus WECCL, such as "more and more", "is more important", "and more people", "more attention to", "the development of", "with the development", "development of the", whose Chinese equivalence are "越来越", "更加重要", "随着…的发展" and etc.. However, there is no single 3-gram of this kind in the native corpus. Obviously, the reason why Chinese learners prefer to use such English expressions is largely related to the increasing development of economy and living conditions in China during the past decades.

Secondly, the combination "a lot of" is the third most frequently used 3-grams in Chinese learners' corpus WECCL, while it does not exist in the 30 most frequently used 3-grams in the native corpus NS. Actually, the 3-gram "a lot of" is ranked 69 with 69 standardized frequencies. However, more specific expressions such as "a number of", "many of the", "most of the", "some of the" and etc. are widely used in the native corpus. One of the reasons is probably that Chinese learners can not clearly distinguish countable nouns from uncountable nouns so that they use the general expression to avoid committing mistakes. Besides, Chinese learners may be unfamiliar with these collocations.

Thirdly, Chinese learners tend to underuse the "noun + of" pattern. It can be easily found that there are quite a

few such type of 3-grams in the native corpus NS, such as "the end of", "part of the", "end of the", "the number of", "the side of", "members of the", "a result of", "the result of". However, there is none of those kinds of expressions only except the two "the development of" and "a lot of". It is sufficient to prove that Chinese learners have not yet had a full understanding of this type of collocation. In other words, they are not quite aware of the considerable use of nouns in English language.

Fourthly, Chinese learners tend to focus on expressing ideas, feelings and attitudes of their own, and therefore they have been quite accustomed to using multiword units such as "in my opinion", "I don t (think)", "I think the", "I think it" and etc.. On the contrary, native people seem to lay emphasis on the description of objective facts rather than subjective feelings. That is why they use more objective expressions such as "it is not", "there is no", "that it was", "it would be", "there was a", and so on and so forth.

2.3 4-Grams

Actually speaking, all four-word n-grams contain within them two three-word n-grams and three two-word n-grams. Although 4-word grams would offer a clearer range of structures and functions than 3-word bundles (Hyland, 2008, p.8), yet the majors features of the former have been more or less reflected in the latter. The 30 most frequently used 4-grams are listed here.

Table 3
The 30 Most Frequently Used 4-Grams Retrieved From NS and WECCL

Rank	St. Fr.	4-Grams (NS)	Rank	St. Fr.	4-Grams (WECCL)
1	153	the end of the	1	230	more and more people
2	95	at the end of	2	181	on the other hand
3	77	at the same time	3	181	with the development of
4	73	per cent of the	4	177	with the development of
5	68	as a result of	5	175	of the Internet on
6	68	for the first time	6	163	impact of the Internet
7	58	the rest of the	7	163	between parents and children
8	48	one of the most	8	162	become more and more
9	46	as well as the	9	162	I don t think
10	44	in the case of	10	162	more and more popular
11	44	by the end of	11	159	the Impact of the
12	42	the fact that the	12	154	the development of the
13	42	the top of the	13	144	more and more attention
14	41	in the form of	14	138	at the same time
15	39	a great deal of	15	138	the best way to
16	38	on the basis of	16	134	don t want to

To be continued

Continued

Rank	St. Fr.	4-Grams (NS)	Rank	St. Fr.	4-Grams (WECCL)
17	35	at a time when	17	128	communicate with each other
18	35	at the time of	18	126	slow down the modernization
19	34	at the same time	19	125	and more attention to
20	33	it is possible to	20	125	is a lifelong process
21	32	on the other hand	21	122	convenient fast and inexpensive
22	31	in the middle of	22	117	the patient the truth
23	31	the extent to which	23	112	becoming more and more
24	31	the way in which	24	110	as a lifelong process
25	31	a result of the	25	109	some people think that
26	31	on the other hand	26	108	speaking is more important
27	31	to be able to	27	107	education as a Lifelong
28	30	the time of the	28	104	is the most important
29	30	will be able to	29	104	as a Lifelong Process
30	30	on the part of	30	104	personal friendly and valuable

The Table 3 shows readers a list of the 30 most frequently used 4-grams retrieved from NS and WECCL. It can be easily seen that the 4-grams in native corpus are mostly functional clusters composed of prepositions, conjunctions and articles and simple frequently-used nouns, such as "the end of the", "at the same time", "for the first time", "one of the most" and so on. On the contrary, the 4-grams in the learners' corpus are mainly formed by content words including nouns, verbs, adjectives and adverbs that are closely related to the subjects of the texts. For instance, the grams, such as "more and more people", "with the development of", "of the Internet on", "between parents and children", "slow down the modernization", obviously indicate the specific subjects like "people", "social development" "the Internet", "parents and children" and "modernization". It cannot be denied that this has formed a sharp contrast between the native corpus and learners' one, which has more or less been proved to be true in the previous parts of the present article.

3. AN ANALYSIS OF THE CAUSES OF THE FEATURES OF N-GRAMS BY LEARNERS

From the statistical discussed above, it can be found that English learners in China have demonstrated some unique features about their n-grams using for various types of reasons. In the following section, four causes will be expounded in perspective of linguistics, society, culture and cognition.

First of all, English expressions used by Chinese learners are deeply affected by the ever-changing society. Since the reform and opening up, the living conditions of Chinese people have gone through a great and unbelievable change. Without any doubt, this social change will exert tremendous influence upon the linguistic development. For instance, there exist many of this type of language expressions which have already been rooted in minds of ordinary people, such as "with the increasing strength of China", "with the further development of reform and open-up policy", "with the development of economy and society", "with the development of science and technology", "people's living condition is becoming better and better", "national strength becomes stronger and stronger" and so on. Therefore, English learners in China can hardly avoid the impact that Chinese social development has exerted upon language, and that is why such language patterns such as "with the development of …" and "sth. is/becomes/will be more and more" have been overused in their writings. It may bring great significance to EFL (English as Foreign Language) teaching, during which teachers should at first be aware of such particular overused patter by Chinese students, and then give them a proper guidance about using proper words or phrases to convey the same ideas.

Secondly, English expressions used by Chinese learners are deeply interfered by the thinking patterns of their mother tongue. Generally speaking, people from English-speaking countries focus on objective thinking patterns, and lay emphasis on the influence that the objective world has exerted upon human lives. That may well explain the fact that language patterns such as the formal subject "it" pattern, "there be" pattern, and the inanimate subject pattern. On the contrary, Chinese people tend to focus on subjective thinking patterns. They stress the people-oriented idea, by which only man instead of inanimate things can produce actions consciously. That is why people's ideas, feelings and attitudes are usually much concerned. Consequently, in English writings by Chinese students, some peculiar language expressions are exceptionally overused, such as "I think", "I don't think", "in my opinion", "I want", "I don't want", "we should", "we can", "as far as I am concerned". To deal with this problem, English teachers should spend some time expounding the differences between English and Chinese thinking patterns and psychological features. If it has been done, firstly students' scopes of knowledge may be highly broadened, and secondly it may help students avoid producing much unidiomatic Chinese English.

Thirdly, Chinese students are hardly aware of the difference between formal written language and informal spoken language. As is well known to most English language teachers in China, the training of listening and speaking of English has been given unbelievable little attention in traditional EFL teaching, while too much stress has been laid on reading and writing. Affected by this traditional teaching theory, teachers would attach no importance to impart knowledge about style differences of written and spoken texts, let alone students. That is why many sentences involving informal colloquial language expressions can be easily found in students' writings. Shorten forms are a case in point, such as "don't", "can't", "it's" and so and so forth.

Fourthly, cognitive features and depth of vocabulary knowledge also affect the use of n-grams in English learners' language. In order to avoid grammatical mistakes, students tend to use the lexical clusters of which they are quite confident. As a result, certain language expressions will be usually overused. It is also likely that n-grams that convey general and abstract ideas will be frequently used. For instance, when it comes to the meaning of "many or much", Chinese learners will be likely to use the cluster "a lot of", perhaps because they can not distinguish whether the modified nouns are countable or uncountable. However, native people tend to use various types of clusters to express the meaning, such as "a number of", "many of the", "most of the", "some of the" and "a large amount of" and etc.. Besides, Chinese learners have not had a good command of in-depth lexical knowledge yet. They are only used to view the words, such as "many", "most" and "some", as adjectives used to define nouns, but not to use them as nouns to be followed by "of" structure. It will definitely bring some limitations to the choice of proper n-grams or even sentence structures. What is more, natives have a wide range of nouns to be used in the pattern "(articles +) noun + of", which is referred to as "Nominalization" or "Noun disease", one of the most conspicuous features of English language (Lian, 2010, p.54). The specific examples of such type of n-grams can be easily found from the native corpus, such as "the end of", "part of the", "end of the", "the number of", "the side of", "members of the", "a result of" and "the rest of" and so on. Comparatively speaking, Chinese learners tend to underuse the "noun + of" n-grams, only except some particular n-grams such as "the development of' and "a lot of".

CONCLUSION

By comparison between the native corpus and learners' corpus, it can be found that English written texts of Chinese learners have shown some unique features. Chinese learners are not good at distinguishing texts of the written style from those of the spoken style. They tend to overuse n-grams that are consisted of content words closely related to subjects under discussion, while they usually underuse n-grams that mainly show the grammatical relations or functional meanings. Chinese

learners are used to choose n-grams which are closely associated with the Chinese social development. Also they usually prefer subject-oriented n-grams that mainly indicate human attitudes or feelings toward the world, rather than object-oriented ones that mainly reflect descriptions of the objective world. In addition, they do not have a good command of proper expressions of quantity and "nouns + of" structure. The present study not only helps readers from different countries have a better understanding about the n-gram features of Chinese learners' English texts, but also brings some implications to EFL teaching. Teachers should consciously develop students' English thinking ways and give them more chance to taste English culture, so that students may have a strong sense of English language.

REFERENCES

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. Harlow, England: Pearson Education.

Carter, R. A., & McCarthy, M. J. (2006). *Cambridge grammar of English.* Cambridge: Cambridge University Press.

Cortes, V. (2002). Lexical Bundles in freshman composition. In D. Biber, S. Fitzmaurice, & R. Reppen (Eds.), *Using corpora to explore linguistic variation* (pp.131-45). Philadelphia, PA: John Benjamins.

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

Lian, S. N. (2010). *A coursebook on English-Chinese translation*. Beijing: Higher Education Press.

Liang, M. C., Li, W. Z., & Xu, J. J. (2010). *Using Corpora: A practical coursebook*. Beijing: Foreign Language Teaching and Research Press.

O'Keeffe, A., McCarthy, M. J., & Carter, R. A. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Wei, N. X. (2011). *Fundamentals of lexis*. Shanghai: Shanghai Foreign Language Education Press.

Wen, Q. F., Wang, L. F., & Liang, M. C. (2009). *Spoken and written English corpus of Chinese Learners*. Beijing: Foreign Language Teaching and Research Press.