Glottometrics 47
2019

Quantitative Studies on English Textual Vocabulary

Dedicated to the Memory of Fengxiang Fan

Guest Editor

Yaqin Wang

Zhejiang University, China

RAM-Verlag

ISSN 1617-8351
e-ISSN 2625-8226

Glottometrics

Indexed in ESCI by Clarivate Analytics and SCOPUS by Elsevier

Glottometrics ist eine unregelmäßig er-scheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druck-version** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

Herausgeber – Editors

External Academic Peers for Glottometrics

Prof. Dr. Haruko Sanada
Rissho University,Tokyo, Japan (http://www.ris.ac.jp/en/);
Link to Prof. Dr. Sanada:: http://researchmap.jp/read0128740/?lang=english; mailto:hsanada@ris.ac.jp
Prof. Dr.Thorsten Roelcke
TU Berlin, Berlin, Germany (http://www.tu-berlin.de/)
Link to Prof. Dr.Roelcke: http://www.daf.tu-berlin.de/menue/deutsch_als_fremd_und_fachsprache/mitarbeiter/professoren_und_pds/prof_dr_thorsten_roelcke
mailto:Thosten Roellcke (roelcke@tu-berlin.de)

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de
Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 47 (2019), Lüdenscheid: RAM-Verlag, 2019. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 47 (2019)
online/ e-version ISSN 2625-8226 (print version ISSN 1617-8351)

Contents

Preface

Bright star, would I were steadfast as thou art?

–John Keats

Fengxiang Fan, an important quantitative linguist in China, died unexpectedly and lamentably on August 19th, 2018, aged at 68. Being one of the co-founders of the School of Foreign Languages at Dalian Maritime University, he had been a professor at the university for around thirty years until his retirement in 2015. As a pioneer in the field of quantitative linguistics (QL), he had published several monographs, numerous publications concerning a variety of important quantitative linguistic issues (Wang & Liu, 2018). He had also served as one of the editors of the *Glottometrics* journal and *Studies in Quantitative Linguistics* book series since 2008. Apart from his own notable achievements, his reputation can also be credited to his influence on new generations' academic progress. There are 60 graduates' M.A. theses which were supervised by him according to CNKI[1] from 2000 to 2015, whose research objectives range widely from maritime convention English to EFL learners' English. This shows the great potential of the applicability of quantitative methods to linguistic issues. Additionally, people may thus catch a glimpse of how QL was developed and also what the current situation is in China through their theses. Therefore, it's a necessary task, I believe, to collect a handful of short versions of theses and display their research questions, methods and certain results.

During the selection, the one closely related to QL was the prior choice, hence a number of theses using methods of corpus linguistics were not chosen (e.g., Zhang, 2007; Wang, 2009; Qi 2009), which were also undoubtedly excellent papers. Among the selected ones, articles that later were modified, to a large extent though, and published in academic journals were left out (e.g., Wang & Liu, 2014; Wang & Liu, 2017). Finally, eight articles in total were collected. By collecting these papers, this volume may shed new light on the development of QL and motivate new thoughts and works in this field. To honor and remember a scholar, nothing is more important than to inherit and develop his/her academic ideas. On the forthcoming first anniversary of Fan's death, this collection is dedicated to the memory of him.

The properties of English vocabulary have long been one of the Fan's research interests, including inter-textual vocabulary growth (Fan, 2006a), the dynamic inter-textual type-token relationship (2006b), textual vocabulary coverage (Fan, 2006c, 2008, 2013) and quantitative features of low frequency word classes and hapax legomena

[1] It is a key national information construction project established in 1996 and now has built China Integrated Knowledge Resources System, including journals, doctoral dissertations, masters' theses, proceedings, newspapers, yearbooks, statistical yearbooks, e-books, patents, standards, and so on. The website is http://www.cnki.net/.

(Fan, 2010a; Fan et al., 2014a, 2014b). Consequently, the majority of students' theses tentatively investigates lexical features using quantitative methods (e.g. He, 2007, Cao, 2008). Out of the collected papers, five deal with textual vocabulary of different text types/genres.

Yingying Xu explores the vocabulary size, inter-textual vocabulary growth patterns, and lexical density of the four sets of Chinese College English textbooks. **Zhao Gao** investigates the lexical characteristics of the commerce English both quantitatively and qualitatively, whose research objectives include the following: vocabulary distribution, vocabulary growth, entropy and perplexity, vocabulary and textual coverage. **Jingjie Li** discusses three issues concerning the inter-textual vocabulary growth patterns of maritime engineering English, including distributions of vocabulary sizes of individual texts, vocabulary growth models, and newly occurring vocabulary distributions of cumulative texts. Based on a maritime convention corpus, **Hong Su**'s study focuses on the characteristics of vocabulary growth of maritime convention English compared with general English from BNC and the mathematical models of vocabulary growth for the maritime convention corpus. Aiming at characterizing the lexis in Hong Kong learners' spoken English, **Yaobin Yan** explores both the quantitative features of the lexis in terms of vocabulary size, mean word length, lexical density and lexical coverage and the qualitative interpretation of these results. Many in-depth quantitative lexical studies on general English have been carried out while few are conducted on specific genres, thus the above studies can be regarded as innovative empirical studies to some extent in this field. These papers, furthermore, add to the growing body of research that indicates the pedagogical implications on English teaching in China, providing insights for different genres of English teaching and learning. In brief, these researches are of both theoretical and practical significance.

Fan also paid attention to other critical concepts in QL, that is, length (Fan et al., 2010) and complexity (Fan et al., 2013). Wang and Liu (2014) examines the interrelationship between length and complexity of sentential constituents and their positions in the sentence partly based on Wang's master thesis (Wang, 2013) supervised by Fan. Two of the collected papers accordingly discuss similar topics. **Pianpian Zhou** examines the relationship between the subjectival position and the sentential syntactic complexity of the spoken English. **Yujia Zhu** compares NPs in spoken and written English from three aspects, namely, length, complexity and pattern. Several mathematical models were also used to be fitted to different empirical distributions. Their studies shed light on the quantitative research regarding several linguistic concepts, i.e., length, position, pattern and NP. Specifically, Zhou's research also provides data support for information theory, the principle of end-focus and of end-weight. On top of these studies, **Fangfang Zhang** discusses issues related to computational stylistics. She made a comparison between American English and British English, aiming to reveal the stylistic characteristics of American English in terms of word length, TTR, high frequency vocabulary and sentence length. Her study helps English learners to obtain a better view of linguistic structure of American English and

enriches study in computational stylistics.

One may notice that all of these papers employed two programming languages, i.e., Visual Foxpro and Perl to address research questions. Needless to say, these studies followed Fan's footsteps and combined quantitative methods with linguistic issues for he had already published two monographs, teaching students without any programming background to use Visual Foxpro and Perl respectively (Fan, 2010b; 2010c). For researchers in this field, they are two excellent textbooks not to be missed. These state-of-the-art statistical methods employed lay a solid foundation for these articles.

To me, one of M.A. students whom he had mentored, he was like a torch in the darkness, always lighting up the road ahead of me. He was the one who led me to the field of QL and offered me valuable advice concerning both of the study and life. The way he acted as a linguist, a teacher, and a friend had lasting effects on me, hence I would like to consider that I have molded myself on his morals to some degree. Even though he is gone, his spirits will consistently inspire me to be as steadfast as he was.

I would like to express my sincere gratitude to all the authors for their prompt reply and time-consuming preparation for these articles since some of them do not work in academia now. This collection would not be made possible without their strenuous efforts. I am also grateful to **Lu Wang** for her kind help in getting in touch with some of Fan's students and editorial work, **Hua Wang** for her careful typesetting work, and **Michal Místecký** for his hard work in improving articles' English. My special thanks also go to **Haitao Liu** and **Gabriel Altmann**, who were Fan's colleagues in QL community, for their proposal for the collection, and their advice and warm support during the whole process.

Yaqin Wang

Department of Linguistics, Zhejiang University, China (mail: wyq322@126.com)

References

Cao, K. (2008). *An empirical study on mathematical models for vocabulary growth* (Master's thesis). Dalian Maritime University, Dalian, China.

Fan, F. (2006a). A Corpus-based empirical study on inter-textual vocabulary growth. *Journal of Quantitative Linguistics*, 13(1), 111–127.

Fan, F. (2006b). Models for dynamic inter-textual type-token relationship. *Glottometrics*, 12, 1–10.

Fan, F. (2006c). Quantitative lexical description of marine engineering English. *Journal of Dalian Maritime University (Social Sciences Edition)*, 5(3), 161–164. (In Chinese.)

Fan, F. (2008). A corpus-based study on random textual vocabulary coverage. *Corpus Linguistics and Linguistic Theory*, 4(1), 1–17.

Fan, F. (2010a). An asymptotic model for the English hapax/vocabulary ratio. *Computational Linguistics*, 36(4), 631–637.

Fan, F. (2010b). *Data Processing and Management for Quantitative Linguistics with Foxpro*. Lüdenscheid: RAM-Verlag.

Fan, F. (2010c). *Quantitative Linguistic Computing with Perl*. Lüdenscheid: RAM-Verlag.

Fan, F. (2013). Text length, vocabulary size and text coverage constancy. *Journal of Quantitative Linguistics*, 20(4), 288–300.

Fan, F., Grzybek, P. & Altmann, G. (2010). Dynamics of word length in sentence. *Glottometrics*, 20, 70–109.

Fan, F., Yu, Y., & Wang, H. (2013). Subjectival position and syntactic complexity in English sentences. In: Reinhard Köhler and Gabriel Altmann (eds.). *Issues in Quantitative Linguistics 3*. Lüdenscheid: RAM-Verlag, p. 137–149.

Fan, F., Wang, Y, & Gao, Z. (2014a). Some macro quantitative features of low-frequency word classes. *Glottometrics*, 28, 1–12.

Fan, F., Zhou, P., & Su H. (2014b). The use of the POR in macro-lexical analyses. In: Gabriel Altmann, Radek Čech, Ján Mačutek and Ludmila Uhlířová (eds.). *Empirical Approaches to Text and Language Analysis*. Lüdenscheid: RAM-Verlag, p. 60–68.

He, L. (2007). *A corpus-based study on the lexical change of the English language* (Master's thesis). Dalian Maritime University, Dalian, China.

Qi, X. (2009). *A corpus-based study on noun-phrase types and their syntactic functions* (Master's thesis). Dalian Maritime University, Dalian, China.

Wang, H. (2013). *The distribution of sentence-initial and sentence-final phrase length in written English* (Master's thesis). Dalian Maritime University, Dalian, China.

Wang, H., & Liu, H. (2014). The effects of length and complexity on constituent ordering in written English. *Poznan Studies in Contemporary Linguistics*, *50*(4), 477-494.

Wang, L. (2009). *A corpus-based study on the translation of Chinese political documents at the sentence and word levels* (Master's thesis). Dalian Maritime University, Dalian, China.

Wang, Y., & Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59, 135-147.

Wang, Y., & Liu, H. (2018). In Remembrance of Fengxiang Fan, 1950–2018, A Pioneer of Quantitative Linguistics in China. *Glottomerics*. 43: 91-96.

Zhang, X. (2007). *A descriptive study on two English versions of Hong Lou Meng* (Master's thesis). Dalian Maritime University, Dalian, China.

Glottometrics 47, 2019, 1-15

The Distribution of Word Families in Chinese College English Textbooks

Yingying Xu[1]

Abstract. Based on a corpus composed of four sets of College English textbooks used in mainland China, this paper examines the vocabulary size, inter-textual vocabulary growth patterns, and lexical density of the four sets of textbooks. Results show that: 1) the vocabulary size of the corpus decreases greatly after lemmatization, and is further reduced after turning lemmas into word families; 2) the inter-textual vocabulary growth patterns of the textbooks can be better described by the word family growth curves; the Brunet's model proves to be good for the description of the inter-textual word family growth for the four sets of textbooks; and 3) in terms of lexical density, the arrangement of teaching materials in some sets of textbooks is not sequenced according to band difficulty and text difficulty.

Keywords: *word family, vocabulary size, inter-textual vocabulary growth, lexical density*

1 Introduction

If language structures make up the skeleton of language, then it is vocabulary that provides the vital organs and flesh (Harmer, 1991). Developing lexical competence in the target language is now seen as the crucial factor in language acquisition (Alderson & Banerjee, 2002); nevertheless, as Cook (1991) notices, one of the major challenges of learning and using a language may not be being competent at syntax, but mastering all the aspects related to the lexicon. In English Foreign Learning (EFL), many learners, even those highly advanced, confirm that they are acutely aware of the lexical gap separating them from educated native speakers of the language, and they perceive vocabulary as an area being rather difficult to tackle. This is also true for Chinese university students who ultimately acquire impoverished lexicons despite years of formal study. On the basis of investigation and study of the research literature, it has been found out that vocabulary size has become a hindrance for Chinese university students in both inputting and outputting information in English, and that to enlarge Chinese students' vocabulary size is therefore critical to the teaching of College English (Du, 2004).

However, as the goal of developing a vocabulary size similar to that of a native

[1] Correspondence to Yingying Xu, School of Foreign Languages, Dalian Maritime University, Dalian, China. E-mail: xuyingying@dlmu.edu.cn

speaker seems beyond the reach of most EFL learners, it is more practical to investigate the minimal amount of words they need to know, and how to reduce the vocabulary learning burden facing them. According to Bauer and Nation (1993), a large amount of vocabulary learning can occur through control of the systematic word-building and affixation features of English. This notion is closely related to the idea of word family which is important for a systematic approach to vocabulary teaching and for assessing the vocabulary load of texts.

On the other hand, in China, College English is a compulsory subject for all the non-English major students at the tertiary level. College English textbooks, an integral part of teaching and learning in a Chinese College English classroom, play a crucial role for exposing learners to the English language. This indicates that to a large extent, learners' vocabulary development depends on the lexical input of textbooks. Therefore, four sets of college English intensive-reading textbooks, *Experiencing English* (EE), *New Horizon College English* (NHCE), *New College English* (NCE), and *New Era Interactive English* (NEIE), which are most widely used in mainland China, were chosen in this study. From the perspective of word family, the present study investigates the vocabulary size, vocabulary growth patterns, and lexical density of the four sets of textbooks.

2 Methods and Materials

2.1 Key Concepts

2.1.1 Token, Type, Lemma, and Word Family

Types refer to unique or different word forms in a text or corpus, while tokens are instances of a type. Therefore, for these 6 words, *compute*, *computes*, *computes*, *computed*, *computing*, *computation*, there are 6 tokens and 5 types. Then, the distinction between a lemma and word family lies in that a lemma consists of only a headword and its inflected forms, while a word family includes a headword and also closely related inflected and derived forms, even if the part of speech is not the same. For example, *admit*, *admits*, *admitted*, *admitting*, *admission*, *admittedly* are regarded as three lemmas – *admit, admission* and *admittedly*, but only one word family – *admit*.

2.1.2 Inter-textual Vocabulary Growth

From a quantitative perspective, many researchers have attempted to investigate the textual lexical growth pattern and proposed a number of vocabulary growth models to render a quantitative description (e.g., Herdan, 1964; Tuldava, 1980). Fan (2006) also tested the existing quantitative models, describing the relationship between vocabulary and text length. Among these models, Brunet's model

$$V = \alpha(\ln N)^{\beta} \quad (1)$$

proves to be robust in capturing the relationship between vocabulary and text length, and is suitable for use in lexical acquisition studies and EFL teaching.

In this paper, the inter-textual vocabulary growth refers to the growth of vocabulary with the cumulative input of texts. It was investigated from the perspective of type, lemma, and word family, so as to compare how the types, lemmas, and word families grow in the textbooks.

2.1.3 Lexical Density and Word Family/Token Ratio

Lexical density is frequently calculated by dividing the number of word types by the number of word tokens. In terms of measuring vocabulary size, Type/Token ratio is much less useful than Lemma/Token ratio, which is, in turn, much less appropriate than Word family/Token ratio (WTR). However, all these measures are sensitive to the corpus size. A standardized measure is a solution to this problem. This study therefore adopts standardized Word family/Token ratio to calculate lexical density of the textbooks. It refers to the ratio between the average number of word families, which exists in 10 sets of 300 randomly drawn word tokens, and 300 (i.e. the number of word tokens).

2.2 Corpus Compilation

Four sets of College English textbooks, from band 1 to band 4[1], were compiled into a corpus for present research. The general information about the textbooks is illustrated in Table 1.

Table 1
Information on the size of the four set of textbooks

Textbook	Number of Texts	Range of Text Size	Tokens
NCE	64	730—2,008	71,874
NHCE	120	564—993	97,313
EE	64	332—930	43,439
NEIE	80	320—917	58,715

Table 1 shows that altogether 328 texts were collected in the corpus totalling 271,341 word tokens. NCE has 64 texts with each band consisting of 8 units with 2 texts in each unit, totaling 71,874 word tokens; NHCE includes 120 texts and 97,313 word tokens, with each band consisting of 10 units and 3 texts in each unit; EE consists of 43,439 tokens and 64 texts, with 16 texts in each band; and each band of NEIE is

[1] Band 1 is designed to be used by college freshmen in their first semester, band 2 by freshmen in their second semester, band 3 by sophomores in their third semester, and band 4 by sophomores in their fourth semester.

made up of 10 units with 2 texts in each unit, totalling 80 texts with 58,715 word tokens.

2.3 Procedures for Data Processing

To get word families, a comparison algorithm written with the Perl programming language was used in this study. It is similar to the Porter Stemmer algorithm (Porter, 1980), except for the fact it also uses several supporting wordlists. This algorithm was applied sequentially to strip inflectional suffixes, remove prefixes, decompose compounds, and strip derivational suffixes from words, corresponding to the process of lemmatization, deprefixing, decompounding, and stemming.

3 Results and Discussion

3.1 Lexical Description of the Textbooks

After tokenization, lemmatization, and particularly turning words into word families, the vocabulary size of each set of textbook decreases greatly, which can be observed from Table 2.

Table 2
General information on the vocabulary size of the four sets of textbooks

Textbook	Tokens	Types	Lemmas	Word Families
NCE	71,874	8,301	5,461	3,654
NHCE	97,313	8,830	5,490	3,713
EE	43,439	6,193	4,163	2,911
NEIE	58,715	7,564	4,963	3,417

It can be noticed that when we use lemma as the counting unit, the number of vocabulary in a corpus is greatly reduced, by about 2,800, 3,300, 2,000, and 2,600 words in NCE, NHCE, EE, and NEIE respectively. Moreover, when the counting unit is word family, the vocabulary size is further reduced, by about 1,200, 1,500, 1,700, and 1,800 in EE, NEIE, NHCE, and NCE.

3.2 Inter-textual Vocabulary Growth Patterns of the Textbooks

3.2.1 Inter-textual Vocabulary Growth Curves of the Textbooks

The following figure displays the type, lemma, and word family growth curves and illustrates the dependence of cumulative word types, lemmas, and word families on the consecutive input of tokens. In this figure, the dotted lines represent the word type

growth curves, the dashed lines the lemma growth curves, and the solid lines the word family growth curves.

The type growth curves in the four panels are higher than the lemma growth curves, which are, in turn, higher than word family growth curves. Therefore, in terms of growth rate, word family growth rate in each panel is the slowest. This proves the idea that use of word family can be in a better position to describe the vocabulary size of the textbooks and the lexical learning burden confronting learners than either word type or lemma.

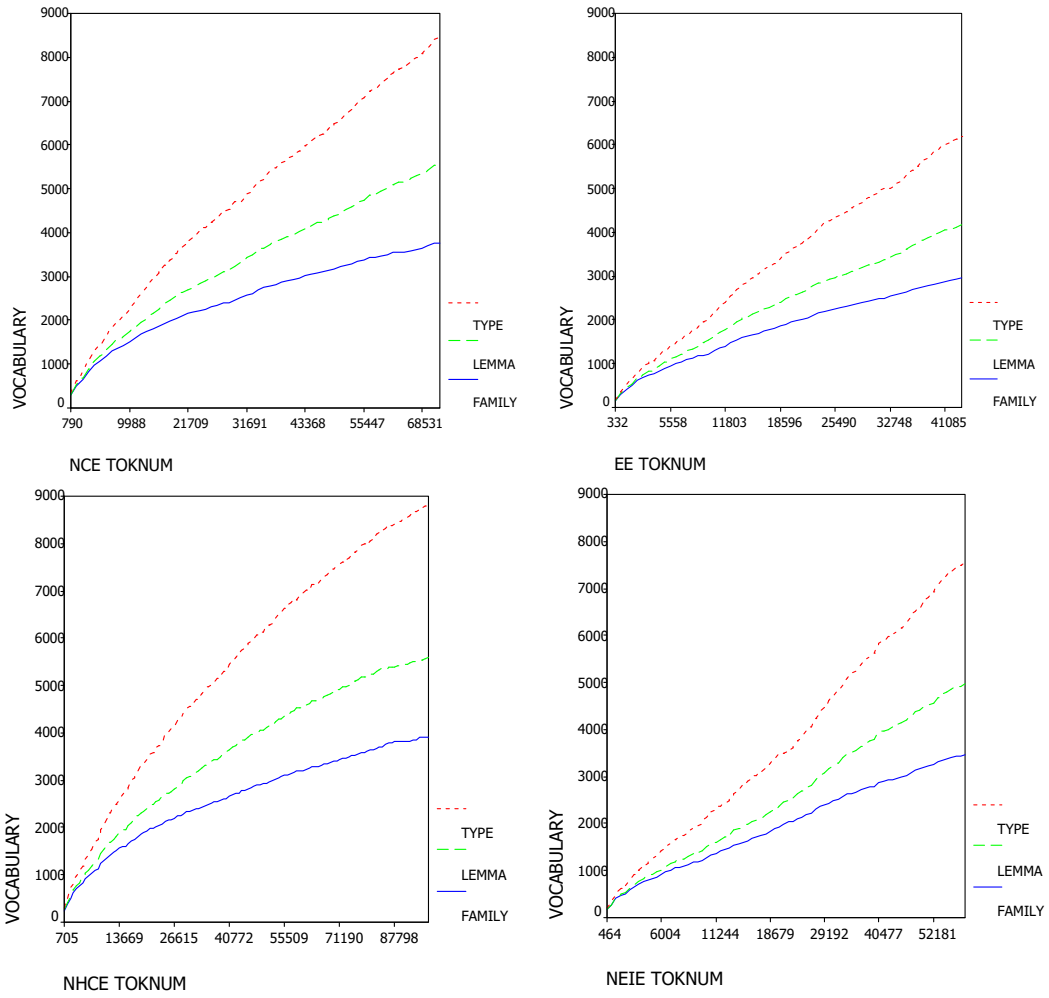

Figure 1. Inter-textual vocabulary growth patterns of the textbooks

Despite the similarities, some differences can be observed. On one hand, in terms of the curve slope, NCE and NHCE exhibit similar patterns, varying sharply from the curves of EE and NEIE. It can be seen that initially, the vocabulary increase rate of NCE is larger than NHCE, which, in turn, is larger than that of NEIE, with EE being the smallest. On the other hand, NCE, EE, and NHCE vocabulary growth curves are comparatively smooth and distinctively different from that of NEIE, revealing there is a steady increase of vocabulary across the four bands in NCE, EE as well as NHCE. In NEIE, nonetheless, except the initial rapid growth, there is another rapid growth of

vocabulary from 21,025 word tokens (band 3) on. That explains why an obvious dividing line can be detected between the first two and the last two bands of NEIE, with the vocabulary size in band 3 increasing suddenly.

3.2.2 Test of Brunet's Model for the Textbooks

As concluded in the previous part, the vocabulary growth pattern of each set of the textbooks can be better described by the word family growth curve, so the Brunet's model fit was tested on the word family growth of the textbooks in this section. The parameters of the Brunet's model parameters for each set of the textbooks are listed in the table below.

Table 3

The parameters of the Brunet's model for the textbooks

Name of Textbooks	α	β
EE	0.008	5.417
NEIE	0.004	5.698
NCE	0.038	4.764
NHCE	0.021	4.979

Based on the parameters in Table 3, the observed word family growth curves (the dotted lines) of the four sets of textbooks and their expected word family growth curves (the solid lines) using Brunet's model are drawn in Figure 2.

In close examination, it can be found out that the fit of Brunet's model has very mild deviations throughout the observed word family growth curves of the four sets of textbooks. For example, in EE, the Brunet's model underestimates the observed vocabulary size of EE at the beginning and at the end of its vocabulary growth curve, whilst it makes an overestimation at the middle of this curve.

To analyse the goodness-of-fit of Brunet's model to each set of the textbooks, R^2, which is a measure of the accuracy of a regression, was calculated. The value of R^2 ranges from 0 to 1, which displays a direct relation with the model's goodness of fit. Due to the fact that a high value of R^2 indicates a good model fit, the fit of the Brunet's model to each of the word family growth curves is good since the values of R^2 for the four sets of textbooks (EE: 0.9987; NEIE: 0.9993; NCE: 0.9986; NHCE: 0.9993) are close to 1.

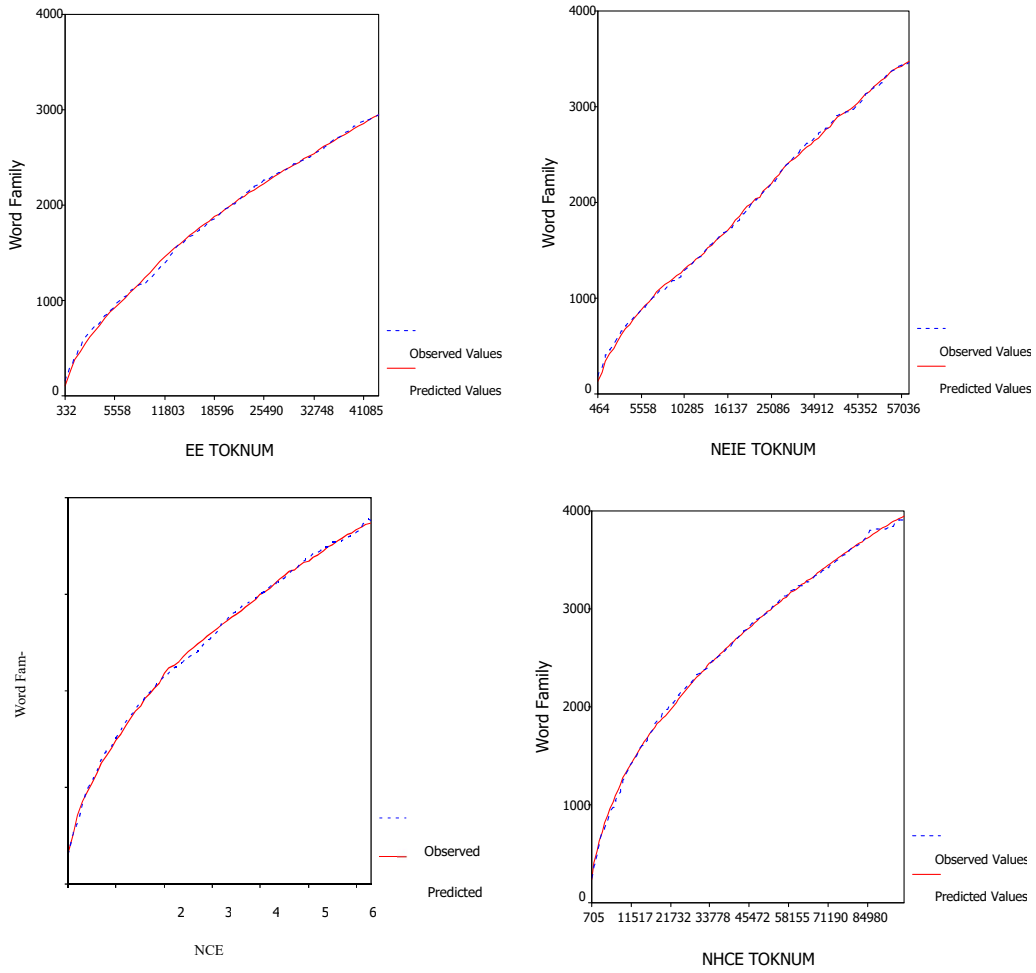

Figure 2. The Brunet's model fit to the observed word family growth of the textbooks

Therefore, the Brunet's model is appropriate for the study of lexical acquisition of the textbooks, and can be used as a mathematical estimator for textual vocabulary coverage to estimate the textbooks' vocabulary coverage over naturally occurring texts of any given length. Making use of the Brunet's model, Fan (2008) proposes the following formula to estimate the textual vocabulary coverage:

$$1-\frac{(\ln N')^{\beta}-(\ln N_i)^{\beta}}{(\ln N_j)^{\beta}} \qquad (2)$$

In the formula, N_i and N_j stand for the number of word tokens in the covering text and covered text; N' equals to N_i plus N_j; and β is a parameter. For example, the total number of word tokens in EE is 43,439, and if we assume the length of an English text has 2,000 word tokens, then, based on the formula and using β listed in Table 3, the estimated vocabulary coverage of EE over the 2,000-word text is:

$$1-\frac{(\ln 45439)^{5.417}-(\ln 43439)^{5.417}}{(\ln 2000)^{5.417}}=0.753.$$

Similarly, the vocabulary coverage of NEIE, NCE, and NHCE over the 2,000-word general English text is, respectively, 0.858, 0.926, and 0.931. It can be seen that none of them reaches 95% coverage.

3.3 Lexical Density in the Textbooks

3.3.1 Frequency Distribution of WTR in the Whole Set of Textbooks

The following figure displays the frequency distribution histograms of WTR in the four sets of textbooks, in which the values on the horizontal axis stand for lexical density in terms of WTR, and the values on the vertical axis represent the frequencies of WTR.

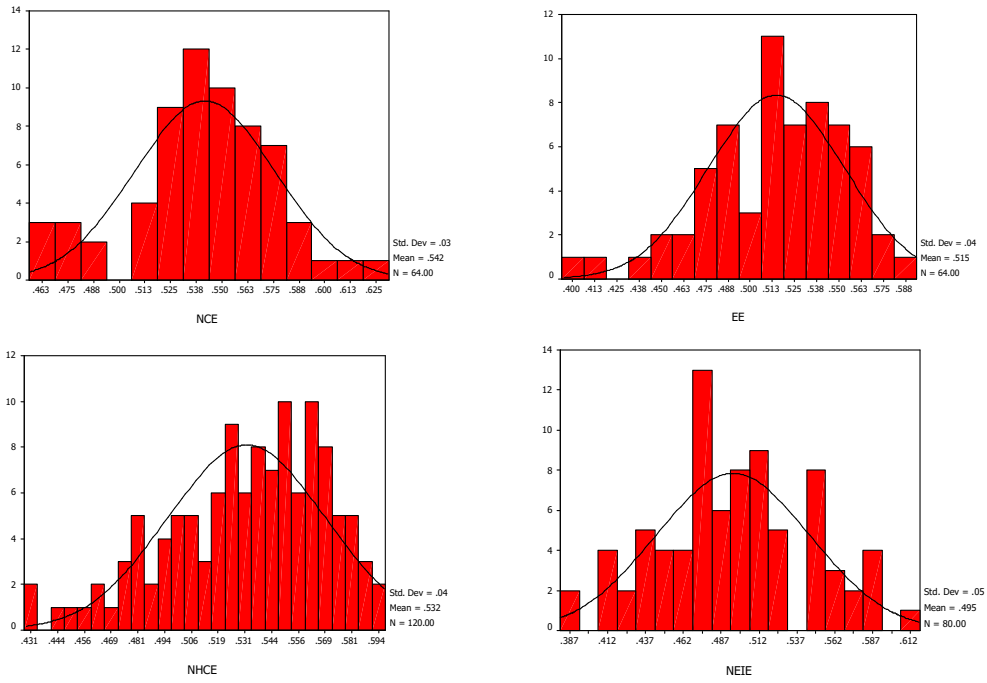

Figure 3. Frequency distribution of WTR in the four sets of textbooks

As shown in the figure, the distribution of WTR is normal in NCE, ranging from the minimum value 0.4610 to the maximum value 0.6223. Most of the WTR values fall between 0.519 and 0.582. The mean value is 0.542 and standard deviation are 0.03. Similarly, the WTR are normally distributed in EE, with the mean value being 0.515 and standard deviation 0.04. The minimal WTR value is 0.4057, the maximum is 0.5817, and most of the WTR values are falling between 0.469 and 0.569. In NHCE, the distribution curve of WTR is roughly normal, with the mean value and standard deviation being 0.532 and 0.04. The minimal and maximal WTR values are 0.4307 and 0.5967

respectively. Most of the WTR values are clustering from around 0.488 to 0.594. The normal distribution of WTR of NEIE can also be observed. Ranging from 0.3817 to 0.6063, most WTR values cluster from around 0.431 to 0.531. The mean value and standard deviation are 0.495 and 0.05 respectively.

Comparatively, NCE has the highest mean value of WTR, which indicates that it is the most difficult for learning, due to the increasing the likelihood of more unfamiliar lexis in its texts, at the same time, it has the smallest standard deviation as well as the smallest range value, indicating the difficulty among texts does not vary greatly from one to another. Noticeably, NEIE has the smallest mean value, but the highest standard deviation and range value, suggesting that it is the easiest for learning and that the difficulty of each text varies greatly from the others. NHCE has the second higher mean value, the third highest standard deviation and range value, which reveals that NHCE contains a richer vocabulary and that there is no great variation in terms of the degree of difficulty among texts. Similarly, EE requires fewer efforts to learn when compared with NCE and NHCE, but is more difficult than NEIE.

3.3.2 WTR between the Bands in Each Set of Textbooks

Based on the calculation of WTR in each text of the four series of textbooks, this section explores WTR of each band in the textbooks in order to show whether the vocabulary makes smooth transitions from band to band.

Table 4
WTR description of each band in NCE

Band	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
			Lower Bound	Upper Bound		
1	.5414	.0324	.5241	.5587	.4733	.5917
2	.5352	.0238	.5225	.5479	.4830	.5703
3	.5392	.0412	.5172	.5611	.4707	.6087
4	.5457	.0418	.5234	.5679	.4607	.6137

As can be noted in Table 4, in NCE, the mean value of WTR of band 1 is larger than those of band 2 and band 3, but only slightly smaller than that of band 4. The data indicate that band 1 has a relative greater lexical diversity and may pose a higher vocabulary learning burden than band 2 and band 3. Still, besides the fact that the difference value between neighbouring bands is too small, the level of difficulty among the four bands is nearly on the same scale, from 0.5352, the smallest WTR value, to 0.5457, the largest value. Thus, the difficulty degree from band 1 to band 4 in NCE is not steadily escalating, but rather remaining on the same level.

Table 5
WTR description of each band in EE

Band	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
1	16	.4873	.0350	.4686	.5059	.4090	.5387
2	16	.5191	.0270	.5048	.5335	.4737	.5613
3	16	.5198	.0458	.4954	.5442	.4057	.5717
4	16	.5328	.0297	.5170	.5487	.4633	.5817

Table 5 demonstrates that the mean value of WTR is increasing from band to band, conveying that the arrangement of bands in EE is sequenced by difficulty. In addition, compared with the big difference value between band 1 and band 2, the difficulty gap between band 2 and band 3 is very small.

Table 6
WTR description of each band in NHCE

Band	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
1	30	.5205	.0386	.5061	.5349	.4520	.5790
2	30	.5279	.0457	.5108	.5449	.4307	.5903
3	30	.5348	.0227	.5263	.5433	.4840	.5847
4	30	.5451	.0340	.5324	.5578	.4320	.5967

Data in Table 6 reveals that the mean value of WTR of each band in NHCE is increasing both gradually and steadily, conveying a smooth transition between bands, and also a steady increase in the level of difficulty from band 1 to band 4.

As described in Table 7, the mean value of the four bands in NEIE is increasing gradually, witnessing a change from a lower lexical burden of band 1 to a greater lexical density within band 4. In terms of the difficulty level between adjacent bands, there is a steady and smooth transfer from one to another.

Table 7
WTR description of each band in NEIE

Band	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum
				Lower Bound	Upper Bound		
1	20	.4531	.0300	.4391	.4671	.3877	.5123
2	20	.4806	.0426	.4606	.5005	.3817	.5840
3	20	.5103	.0516	.4862	.5345	.4077	.6063
4	20	.5368	.0346	.5206	.5529	.4720	.5873

3.3.3 WTR of Each Text in the Textbooks

This section investigates the WTR of each text in the four sets of textbooks. The information is illustrated in the following four figures.

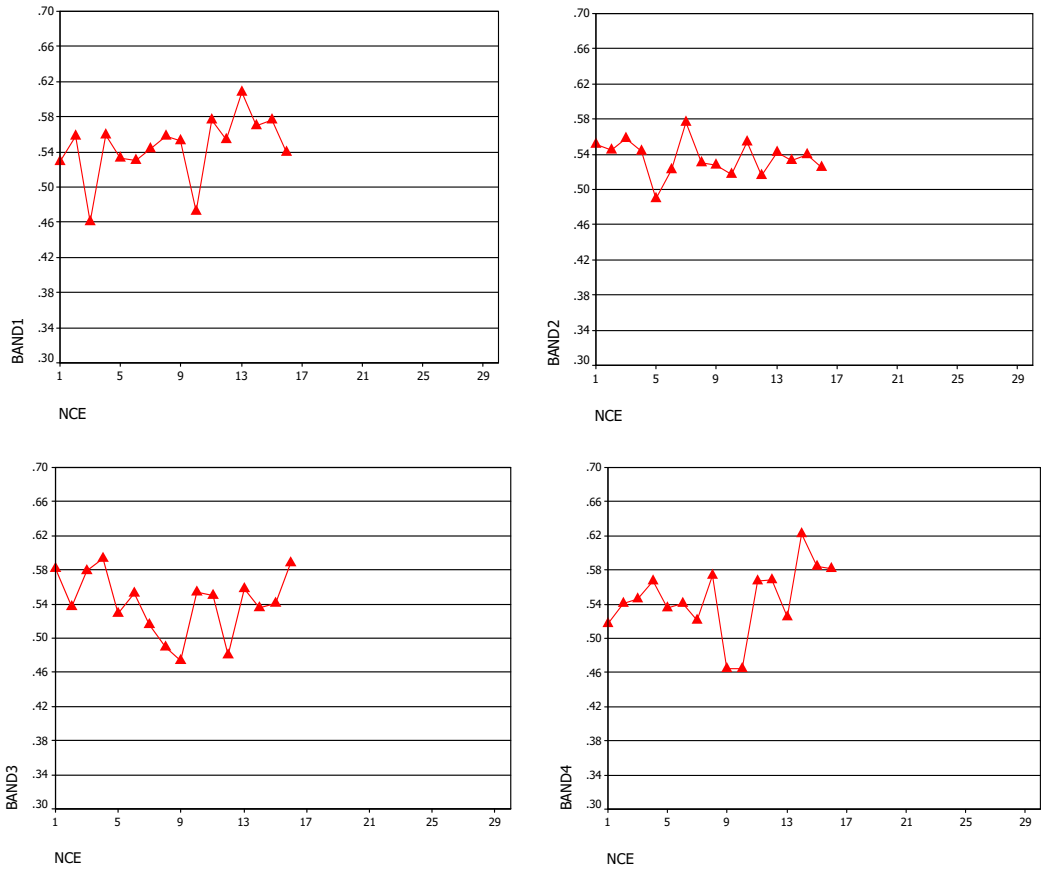

Figure 4. WTR of each text in NCE

In NCE, the curve rises gradually except two extreme values, indicating that the

texts in band 1 are arranged in accordance with the degree of difficulty, only with texts 3 and 10 being too easy. In band 2, the curve fluctuates around the mean value 0.54, which suggests that there is no great variation among texts within the band. Furthermore, text 5 is comparatively simple, and text 7 is relatively difficult. In band 3, the curve initially rises and then falls in the first half of the book, this is followed by rising in the latter half of the book, showing that texts in band 3 are sequenced from the difficult to the easy ones, and then they become difficult again. Moreover, text 9 and 12 are comparatively too simple. In the last graph, the curve shows a general ascending tendency, except two extreme values. Therefore, band 4 is also arranged in order of text difficulty, except for the texts 9 and 10 that are too simple. In a word, band 1 and band 4 in NCE are arranged in accordance with text difficulty.

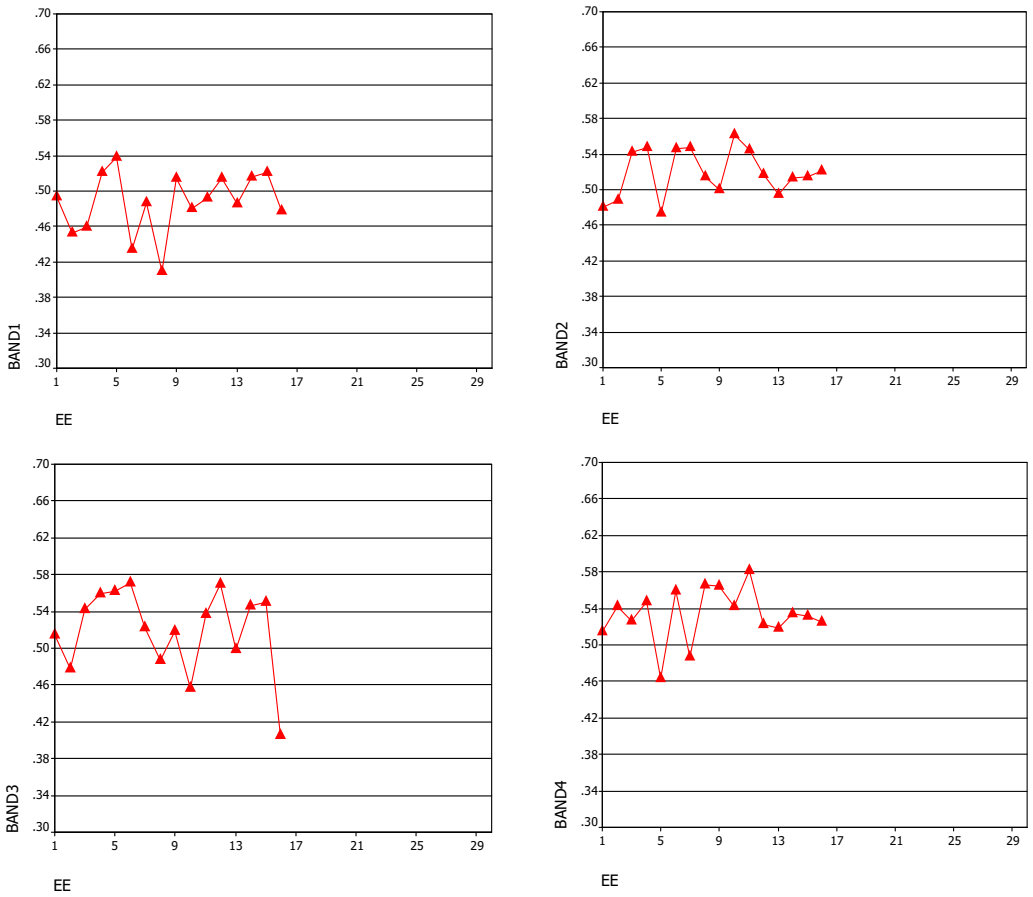

Figure 5. WTR of each text in EE

Figure 5 exhibits the curve of WTR value of each text in the four bands of EE. In band 1, the general rising curve indicates that the arrangement of texts is generally from the easy to the difficult ones, with text 5 being too difficult and text 8 too simple. In band 2, texts are initially arranged from the easy to the difficult ones, except for the fact that text 5 is relatively simple; however, from text 10 till the end, the curve gradually decreases, suggesting texts are becoming easier and easier from text 10 to text 16. The arrangement of band 3 is not well-balanced, as shown in the third graph, in which the

last text is too simple. Band 4 exhibits the similar trend as band 2, initially increasing and then declining. Therefore, only band 1 in EE can be said to be arranged according to the order of text difficulty.

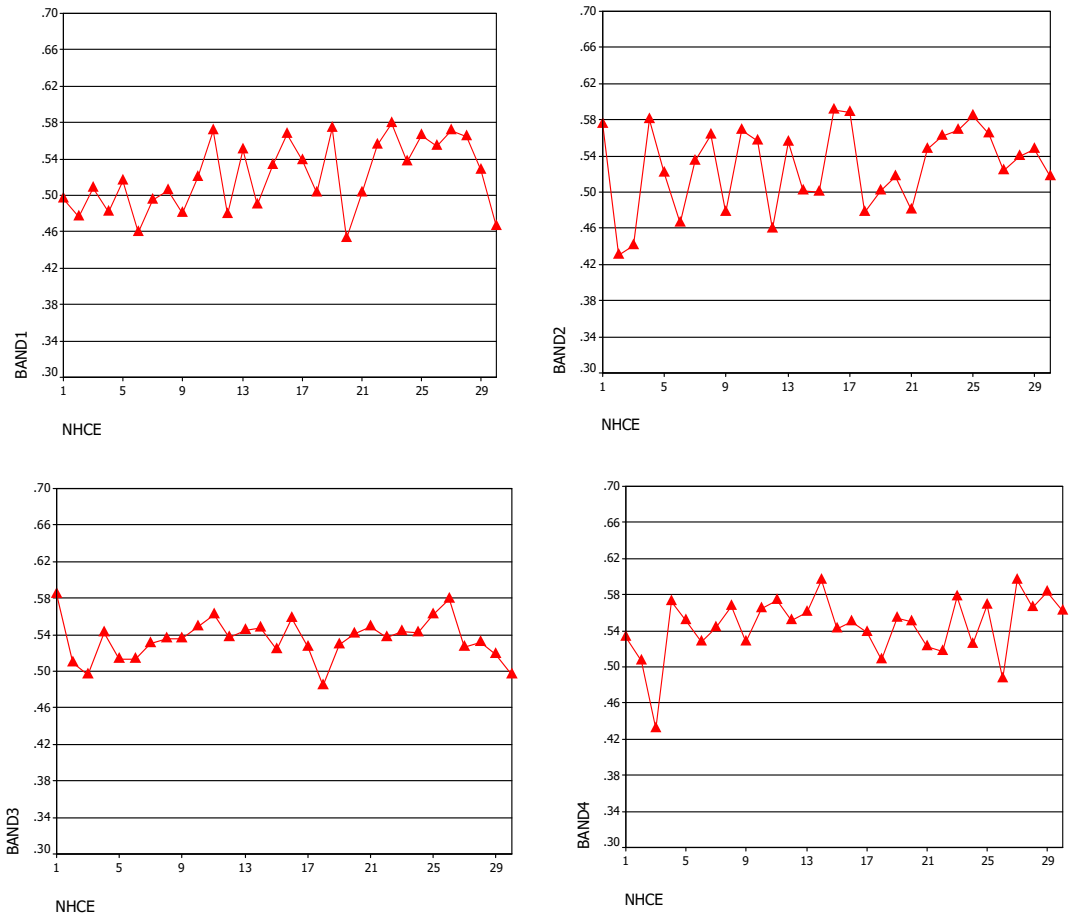

Figure 6. WTR of each text in NHCE

As shown in Figure 6, the steadily escalating curve of band 1 conveys the smooth transition of difficulty among texts. However, texts 20 and 30 are a little bit simple when compared with other texts. In band 2, the curve fluctuates around the mean value of WTR, and the high WTR of text 1 indicates it is rather difficult to comprehend as the first text. Texts in band 3 are arranged on the basis of the ascending difficulty, as the third graph shows. However, text 1 is too difficult as the first input text, and text 18 is comparatively simple. As depicted in the last graph, the general ascending tendency is noticeable; however, the WTR of text 3 is too low, thus text 3 is too simple when compared with the previous and following texts. Therefore, band 1, band 3, and band 4 in NHCE are sequenced in accordance with the order of difficulty.

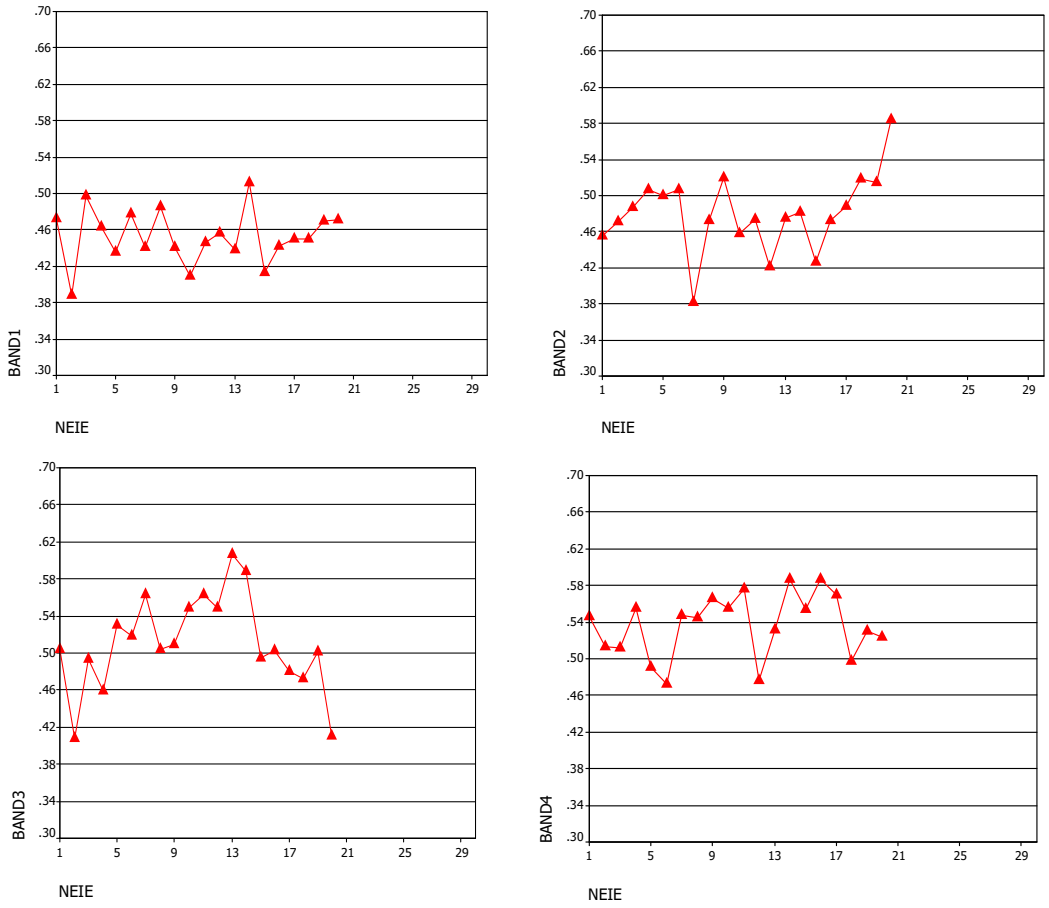

Figure 7. WTR of each text in NEIE

According to Figure 7, band 1 is sequenced roughly from the easy to the difficult ones. Texts in band 2 are generally arranged from easiness to difficulty with text 7 being too simple. In band 3, the curve is initially increasing and then declining, which reveals that band 2 is arranged from the easy texts to the difficult ones, and then turns to easy texts again. Text 2 and text 20 in band 3 are rather simple. In band 4, except for text 6 and 12 being too simple, texts are arranged in accordance with the order of difficulty. In a word, in NEIE, band 1, band 2, and band 4 are arranged in accordance with text difficulty.

4 Conclusion

This study explores the distribution of word families in the four sets of College English textbooks which are most widely used in mainland China. Through the investigation of vocabulary size, vocabulary growth, and lexical density, the following conclusions have been reached: 1) the vocabulary size of each set of textbooks decreases greatly after lemmatization (about 2,000—3,300 words), and further declines after turning words into word families (about 1,200—1,800 words); 2) the inter-textual vocabulary growth patterns of the four sets of textbooks can be better described by the word family growth

curves; the Brunet's model proves to be good for the description of the inter-textual vocabulary growth for the four sets of textbooks, and therefore, by virtue of this model, it has been found out that none of the four sets of textbooks' vocabulary coverage over 2,000-word texts reaches 95% coverage rate (EE: 75.3%; NEIE: 85.8%; NCE: 92.6%; NHCE: 93.1%); and 3) in terms of lexical density, or rather WTR, the difficulty level of the four sets of textbooks does not vary greatly, but the arrangements of teaching materials in some sets of textbooks are not sequenced according to band difficulty and text difficulty. However, some issues are not addressed in this paper and remain open for further studies, such as the distribution of newly occurring word families, which are also important in vocabulary acquisition and which can affect the process of reading comprehension.

References

Alderson, J., & Banerjee, J. (2002). Language testing and assessment (Part 2): State of the art review. *Language Teaching*, 35(2), 79–113.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, 6, 253 – 279.

Cook, V. (1991). *Second Language Learning and Language Teaching.* London: Edward Arnold.

Du, H. (2004). Reflections on Vocabulary Size of Chinese University Students. *International Education Journal*, 5, 571–581.

Fan, F. (2006). A Corpus-Based Empirical Study on Inter-textual Vocabulary Growth. *Journal of Quantitative Linguistics*, 13, 111–127.

Fan, F. (2008). A corpus-based study on random textual vocabulary coverage. *Corpus Linguistics and Linguistic Theory*, 4(1), 1–17.

Harmer, J. (1991). *The Practice of English Language Teaching*. London: Longman.

Herdan, G. (1964). *Quantitative Linguistics*. London: Butterworths.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.

Tuldava, J. (1980). A mathematical model of the vocabulary—text relation: COLIN 1980, *Proceedings of the 8th International Conference on Computational Linguistics*, Tokyo, September 1980.

Glottometrics 47, 2019, 16-35

A Quantitative Lexical Study on Commercial English

Zhao Gao[1]

Abstract. This study is corpus-based and employs the theory and methodology of quantitative linguistics investigating the lexical characteristics of the Commerce Domain of British National Corpus (hereafter referred to as CDBNC), both quantitatively and qualitatively**.** The samples of CDBNC were drawn randomly from BNC. As a reference, other eight groups of samples in other eight domains from BNC were drawn randomly. The contents of the research include the following: the lexical statistics, vocabulary distribution, vocabulary richness, vocabulary growth, entropy and perplexity, vocabulary and textual coverage by CET–4 and CET–6 over CDBNC, Brunet's model, and Tuldava's model fit. The major findings of the present research can enrich empirical studies in the field of quantitative linguistics.

Keywords: *CDBNC, corpus, quantitative linguistics, models, vocabulary growth, vocabulary coverage*

1 Introduction

Commercial English has a significant role in modern social and economic life and the importance is remarkable. In international commercial activities, commercial English has a wide use and becomes a unique style of written language which is called *commercial English* in Britain and *business English* in the USA. Commercial English is a variety which has a strong social function and aims at serving for commercial activities. In today's economic globalization, commercial English has become an important domain of the English language, and gradually developed into the lingua franca.

The present research undertakes the task of making a quantitative lexical study on commercial English. The significance of this research is explained in the following paragraphs.

Firstly, linguists have made many in-depth quantitative lexical studies on general English and scientific and technological English, but there are very few related studies on the quantitative characteristics of commercial English. So, it is an innovative research and can enrich the empirical study in this field.

[1] Correspondence to Zhao Gao, Dalian Smart City Administration Supervision and Command Center, Dalian, China. Email: melody_vvi@163.com

Secondly, the present research has an important academic significance for quantitative lexical research, machine translation, natural language processing, and other fields.

Thirdly, the present research can shed light on commercial English learning and teaching. Now, the main teaching content of college English in China is General English or Basic English; this kind of examination-oriented English teaching and learning may not adapt to the economic globalization and internationalization. This research can provide some strong empirical evidence and theoretical support for the reform of college English education, the reform of the curriculum design, teaching material compilation, and English teaching for college English tests CET–4 and CET–6.

All the previous research was carried out mainly using relatively small amount of data and conventional methods. Few researchers have adopted the quantitative linguistic theories and methodologies so far. The present research investigates the quantitative lexical characteristics of commercial English. The research questions are as follows:

1. What are the characteristics of the lexical statistics of commercial English?
2. What is the vocabulary richness of commercial English?
3. What are the vocabulary growth, hapax growth, and hapax vocabulary ratio of commercial English?
4. What are the vocabulary coverage and textual coverage provided by CET-4 and CET-6 of commercial English?

The present research is corpus-based. The quantitative linguistic theory, statistical methods, and computer language processing were used in this research.

2 Methods and Materials

2.1 Data Source

2.1.1 Description of BNC

This research is corpus-based and statistically driven. The corpus used is the British National Corpus, also known as the BNC. It contains about 100 million words in texts of different registers and genres in British English, written in the late 1980s and early 1990s. It contains 4,124 files, 90% of which are written and 10% spoken.

According to Gries (2007), corpora are inherently variable in different linguistic aspects, both internally and externally. BNC's written section consists of nine domains. They are: applied sciences, arts, belief and thought, commerce, imaginative, leisure, natural sciences, social sciences, and world affairs. These domains have their own lexical and syntactic characteristics. Table 1 shows the numbers of tokens of the nine domains of BNC.

Table 1
The numbers of words of the nine domains of BNC

Domain	No. of words in domain
Commerce	7,516,450
Applied science	7,737,395
Arts	6,915,501
Belief and thought	3,099,344
Imaginative	19,149,864
Leisure	12,528,823
Natural sciences	3,924,190
Social sciences	14,469,798
Word affairs	17,820,390

2.1.2 The Commerce Domain of BNC

The text samples of commercial English were drawn randomly from the *commerce* domain of the written part in BNC; the number of samples and the size of individual samples of CDBNC was determined by the following factors:

First, the number of samples should be large enough to capture lexical characteristics of the domain.

Second, the cumulative volume of the sample texts should be able to reach the vocabulary size of native speakers. Nation (1990) and Nagy (1997) place the vocabulary size of educated native speakers at 20,000; Radford et al. (1999) gives it a much higher figure, around 30,000. Usually, an EFL learner's vocabulary size does not exceed those estimates. A test revealed that cumulative texts of 2,000,000 words can roughly cover or exceed the vocabulary size of the native speaker.

Third, the size of texts used in intermediate and advanced EFL teaching should be taken into consideration; this is usually around 2,000 words. A text which is either too small or too big in size will influence the validity or the representativeness of the research.

Based on the above considerations, 1,000 2,000-word text samples were randomly drawn from the *commerce* domain in BNC, totalling 2,000,000 words.

2.2 Data Processing

BNC contains a variety of POS tags, text formatting codes, and other non-textual material in the standard generalized mark-up language coding sets. These non-textual tags were removed initially. Then, the lemmatization was carried out. Lemmatization is the process of gathering word-forms and arranging them into lemmas or lemmata (Sinclair, 1991). A lemma, in turn, is "a set of lexical forms having the same stem, the same major part-of-speech, and the same word-sense" (Jurafsky and Martin, 2000: 195). For

example, words such as “go”, “goes”, “going”, “went”, “gone”, and “conference”, “conferences”, and so on are different word forms of the lemmas “go” and “conference”. The current study devised a lemmatisation algorithm called the comparison algorithm (Fan, 2010b). The lemmatisation accuracy of this algorithm is 98%.

2.3 Tools for Data Processing

To obtain the necessary data for this research, the programming language Foxpro, Perl, and the regression analysis software NLREG were used.

Foxpro is a powerful and widely used computer database management system operated by using a programming language. It can perform complicated math operations and string manipulation by a set of user-friendly, natural-language-like commands and functions for table handling (Fan, 2010a). In this study, a Foxpro programme was used for text sampling and lemmatization.

In order to obtain the data which cannot be easily processed by Foxpro, another powerful programming language, Perl, was applied in this research. Perl is a powerful and efficient computer language. In this research, it was used for computing vocabulary growth, vocabulary coverage, textual coverage, and entropy.

NLREG, short for nonlinear regression, can handle linear, polynomial, exponential, logistic, periodic, and general nonlinear functions. In this study, NLREG was utilized for estimating the statistical regression analysis and estimating the value of parameters for linear and general nonlinear functions.

3 Results and Discussion

This chapter discusses the results in both quantitative and qualitative ways. It is divided into four sections. Section 3.1 displays the lexical statistics of CDBNC and the other eight domains of BNC. Section 3.2 shows the vocabulary richness of CDBNC using TTR, textual TTR distribution, and the relationship between text length and TTR using Tuldava’s model. Section 3.3 analyses the vocabulary growth. The Brunet’s model is used to describe the vocabulary growth of CDBNC. Section 3.4 analyses the vocabulary coverage and textual coverage. The vocabulary and textural coverage of CET–4 and CET–6 over the 1,000 text samples are discussed. The vocabulary not covered by CET–4 and CET–6 is also discussed.

3.1 Lexical Statistics

3.1.1 Overall Lexical Statistics

A very important lexical distribution characteristic of CDBNC and other eight domains is shown in Table 2. The overall sample size, overall vocabulary size, overall number of hapax and overall TTR of the nine domains were computed. The relationship

between the vocabulary size and TTR can be found from the following analysis.

The overall vocabulary sizes of the 2,000,000-word text samples from the nine domains are different; there are significant differences in other three lexical statistics of the nine domains. We find that CDBNC has the smallest vocabulary size, hapax, and the TTR, which are respectively 30,044, 10,622, and 0.015; while arts has the largest number of the three lexical statistics, which are respectively 49,685, 19,259, and 0.0241.

Table 2

Descriptive statistics of the sample size, vocabulary size, number of hapax, and the overall TTRs of the nine domains

Domain	Size of samples	Voc.	No. of hapax	Overall
Commerce	2,000,000	30,044	10,622	0.0150
Applied sciences	2,000,000	37,259	14,015	0.0188
Arts	2,000,000	49,685	19,259	0.0241
Belief and thought	2,000,000	32,753	11,046	0.0158
Imaginative	2,000,000	33,770	11,064	0.0152
Leisure	2,000,000	44,972	16,922	0.0213
Natural sciences	2,000,000	38,195	14,294	0.0194
Social sciences	2,000,000	31,525	11,211	0.0158
Word affairs	2,000,000	44,630	17,731	0.0222

In short, the dispersion of the vocabulary size of each domain is large, from 30,044 to 49,685. The hapax is from 10,622 to 10,259, and TTR is from 0.015 to 0.0241. The relationship between the vocabulary size and TTR is very close. TTR serves as a measure for vocabulary diversity, giving exact information of how lexically varied the text is.

3.1.2 Entropy and Perplexity

Entropy is a measure of lexical richness and represents the amount of information in language. According to the statistics, we find that the lowest perplexity is corresponding to the lowest entropy; the greater the entropy value is, and the greater the amount of information will be. Table 3 shows the entropy and perplexity of the nine domains of BNC.

Table 3 displays the entropy and perplexity of the nine domains appearing in the BNC. They are ranked in the descending order, according to the counts of entropy and perplexity. From the top to the bottom, we notice that the entropy and perplexity of each domain are ranked in the same order respectively. The highest ranking is leisure, with a number of 10.132 for entropy and 1122.078 for perplexity. Arts is the second largest,

with a number of 10.093 for entropy and 1092.172 for perplexity. The following are world affairs, applied sciences, natural sciences, commerce, social science, belief and imaginative. Commerce is ranked number 6; its entropy and perplexity are much smaller than the above 5 domains, 9.676 and 818.024 respectively. Imaginative has the smallest values; its entropy and perplexity are 9.500 and 723.994, respectively.

Table 3

The entropy and perplexity of the nine domains of BNC

DOMAIN	ENTROPY	PERPLEXITY
Leisure	10.132	1122.078
Arts	10.093	1092.172
World Affairs	10.021	1038.721
Applied Sciences	10.010	1031.417
Natural Sciences	10.001	1024.414
Commerce	9.676	818.024
Social Sciences	9.675	817.422
Belief and Thought	9.573	761.538
Imaginative	9.500	723.994

3.1.3 Textual Vocabulary Distribution

The vocabulary size distribution of the individual 1,000 text samples from CDBNC is listed in Table 4. The distribution of the vocabulary size is shown in Figure 1.

Table 4

Descriptive statistics of the vocabulary distribution of the individual 1,000 samples

	Valid texts no.	Minimum voc. size	Maximum voc. size	Mean voc. size	Std. Deviation	Range
Vocabulary distribution	1,000	271	854	600.080	104.180	583

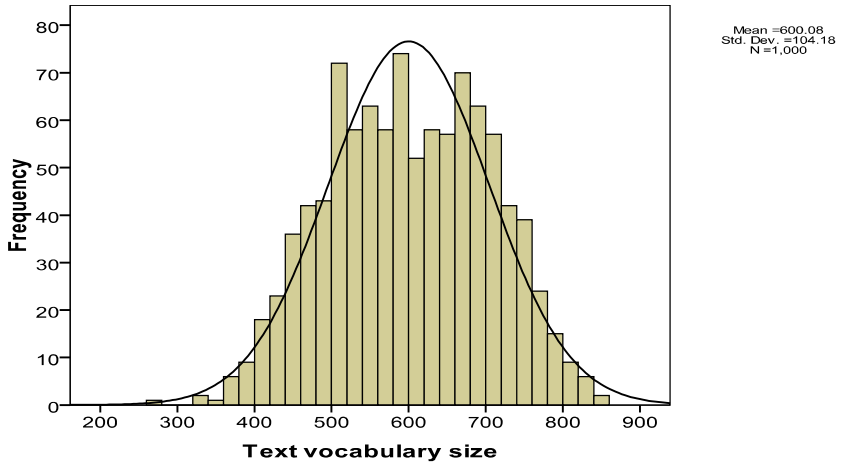

Figure 1. Textual vocabulary distribution with the normal distribution curve

The K–S test was applied to test whether the text vocabulary distribution of the individual samples is normal. The K–S result shows that *Z* is 0.606 and the *p*-value is 0.856. The *p*-value is much greater than 0.05, which shows that the text vocabulary distribution is basically normal.

The tolerance interval formula (Devore, 2000) below (1) is used to predict the possible values of the vocabulary sizes of any text of 2,000 words of CDBNC

$$\textit{Tolerance Interval} = \textit{mean} \pm (\textit{tolerance critical value}) \times S \quad (1)$$

S is the standard deviation of a set of normally distributed values. *Tolerance critical value* is provided by many books on statistics; it is determined by the number of observed values and the percentage of all such possible values the tolerance interval intends to include. To capture 95% of all the possible values of a normally distributed population at the 95% confidence level, the tolerance critical value is 1.96 for the number of observed values > 300. For any text of Commercial English, the size of which is similar to the one of the samples from CDBNC, its estimated vocabulary size with its lower and upper bounds can be obtained by (1), using the related textual lexical statistics for the sample sets listed in Table 3.3. For example, there is a 95% probability that the vocabulary of a text of CDBNC of about 2,000 words will lie between 600.08 ± 1.96 × 104.18, which mean 804.2728 and 395.8872, respectively.

Figure 2 is the box plot for the textual vocabulary size of the 1,000 texts.

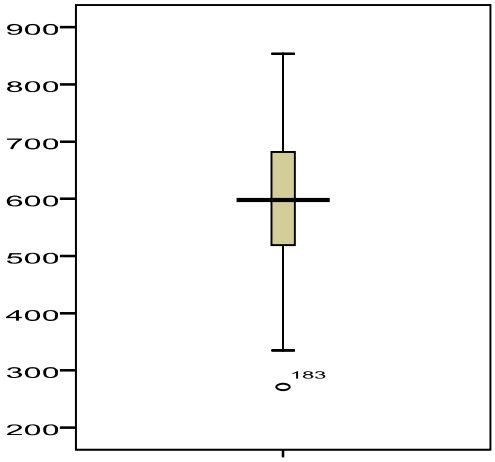

Figure 2. Textural vocabulary distribution boxplot

The above box plot reflects the vocabulary distribution of CDBNC. The central box of the figure represents the majority of the vocabulary size of the 1,000 samples; the horizontal line is the mean, which is 600.08. The minimum value is 271 and the maximum value is 854. The outlier is the number 183 sample, which is 271.

3.2 Vocabulary Richness of CDBNC

This section displays the overall TTR statistics, textual TTR distribution, and the relationship between text length and TTR. For the samples of the same size, TTR is a vocabulary richness indicator. As shown in Table 3.1, the overall of TTR of CDBNC is 0.015, while from the top to bottom, the TTR of the nine domains are respectively arts, world affairs, leisure, natural sciences, applied sciences, belief and thought, social sciences, imaginative, and commerce. This shows that CDBNC has the lowest TTR and vocabulary density from all the domains.

3.2.1 Textual TTR Distribution

Table 5 shows the statistical information for the individual textual TTR distribution of 1,000 2,000-word samples. In addition, out of the 1,000 TTR values, 250 are below 0.263, 250 over 0.339, and 500 in between. Figure 3 displays the frequency histogram of the textual TTR distribution of CDBNC.

Table 5

Descriptive statistics of textual TTR distribution of the individual samples

	Valid texts no.	Minimum TTR	Maximum	Mean TTR	Std. deviation	Range
Vocabulary distribution	1,000	0.1422	0.424	0.300	0.052	0.282

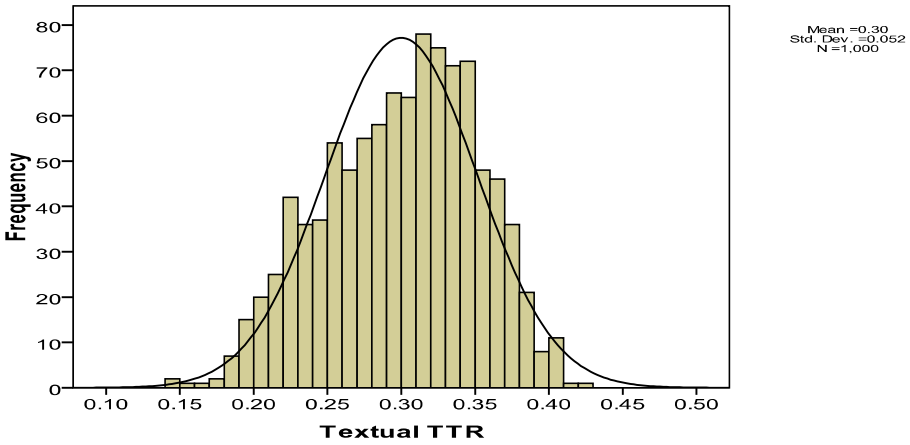

Figure 3. Textual TTR distribution of the individual 2,000-word samples of CDBNC with the normal distribution curve

A one sample K–S test was used to test whether the TTR is normally distributed. The K–S Z result is 0.957, and the p-value is 0.319; the *p*-value is much greater than 0.05, which shows that the textual TTR distribution is normal.

The tolerance interval formula of (1) is used to predict the TTR of any 2,000-word text of commercial English. There is a 95% probability that the TTR of a text of CDBNC of about 2,000 words will lie between 0.299964 ± 1.96 × 0.0516849, which are 0.1987 and 0.4013, respectively.

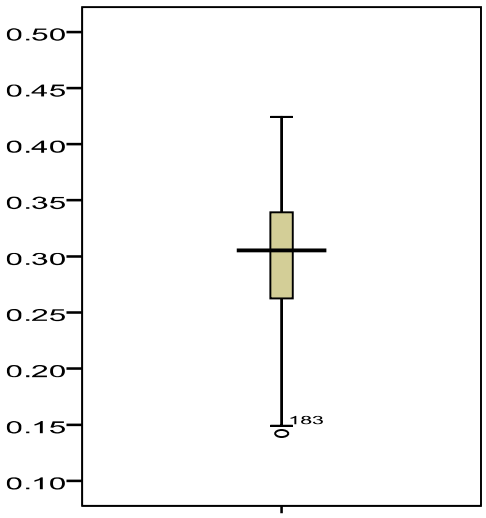

Figure 4. TTR distribution boxplot of the individual samples

The box plot in Figure 4 reflects the TTR distribution of CDBNC. The central box represents the majority of TTR of the 1,000 samples; the horizontal line is the mean, 0.299964. The minimum value TTR is 0.1422, and the maximum value is 0.4244. The outlier is the number 183, which scores 0.1422.

3.2.2 Relationship between Text Length and TTR

There is a certain relationship between text length and TTR. It can be captured by Tuldava's model shown below:

$$TTR = Ne^{-\alpha(\ln N)^{\beta}} \tag{2}$$

In the above formula, N stands for text length, e is 2.71828, and α, β are parameters.

It can be found that the model fits the observed data very well, with $\alpha = 12.921$; $\beta = -0.456$; $R^2 = 0.998$. Figure 5 shows the fit. The smooth solid line is the model fit, and the dotted line represents the empirical values.

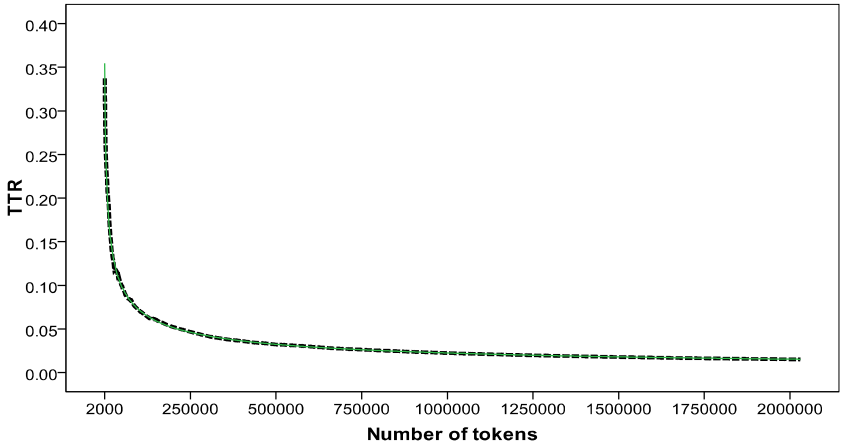

Figure 5. The TTR distribution curves of the Tuldava's model fit to CDBNC

As shown in Figure 5, the solid line almost overlaps the dotted line, which means the Tuldava's model is very good for the description of the TTR decrease as the text length increases.

3.3 Vocabulary Growth

This section covers on the vocabulary growth of CDBNC, Brunet's Model, residual analysis, and the HVR.

3.3.1 Vocabulary Growth of CDBNC

According to Gries (2007), corpora are inherently variable in different linguistic aspects both internally and externally. BNC's written section consists of nine domains. These domains have their own vocabulary growth characteristics. To study the vocabulary growth of each of the domains at N = 2,000,000, 1,000 2,000-word samples were randomly drawn from each of the domains.

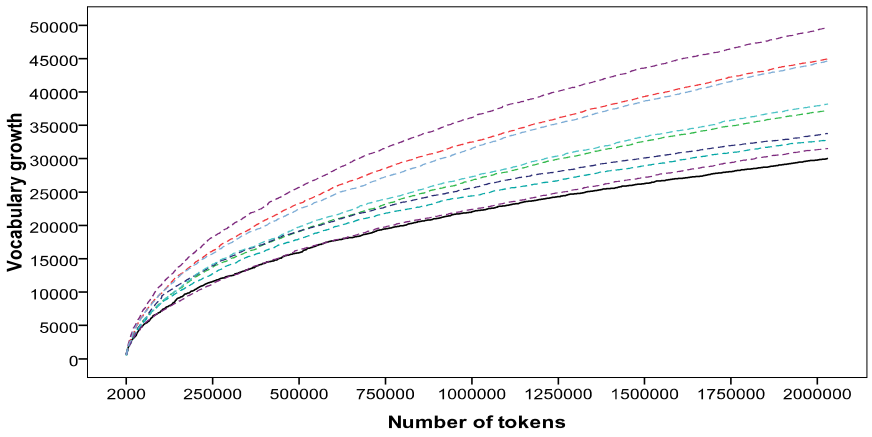

Figure 6. Vocabulary growth of the nine domains of BNC. The solid line is the vocabulary growth curve of CDBNC. The dotted lines are the vocabulary growth curves of the other domains of BNC

The vocabulary growth curves of the nine domains are displayed in Figure 6. The vocabulary growth curves of the individual samples from each of the nine domains are distinctive. These curves from top to bottom respectively are: arts, leisure, world affairs, natural sciences, applied science, imaginative, belief and thought, social science, and commerce. Out of the nine domains, CDBNC has the smallest vocabulary, while *arts* have the largest.

3.3.2 Brunet's Model

The Brunet's model was tested on the vocabulary growth of the samples from CDBNC at a 2,000-word interval. The parameters are listed in Table 6.

Figure 7 displays the Brunet's model fits to the vocabulary growth curve of the samples from CDBNC. The fit of the Brunet's model to the CDBNC vocabulary growth data is very good.

Table 6

Parameters of the Brunet's model for the individual 2,000-word samples of CDBNC

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
α	0.002	0.000	0.002	0.002
β	6.147	0.004	6.140	6.154

$R^2 = 1$

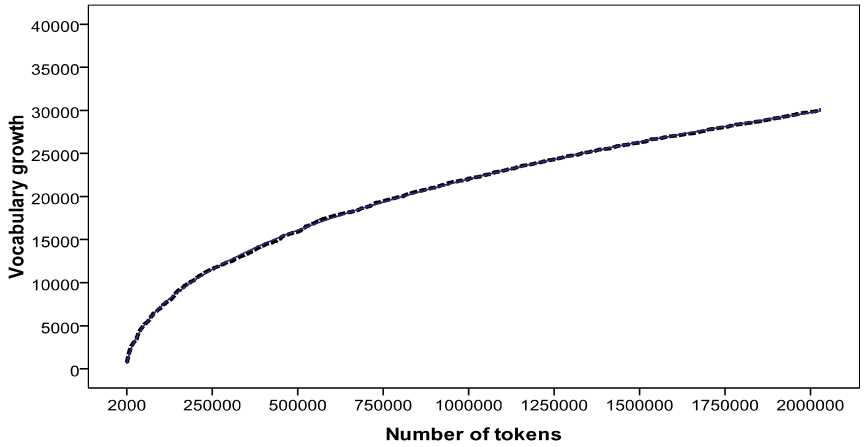

Figure 7. The vocabulary growth curve of the Brunet's model fits to CDBNC. The solid line is the Brunet's model fit; the dotted line represents the observed values.

3.3.3 Residual Analysis

Mendenhall & Scheaffer (1973) suggests that if the model makes correct predictions of the observed values, the residual should randomly alternate between the positive and negative values, and there should be no coherent patterns in the residuals. The greater the coherent structure of the residual, the greater the chance that part of the estimated trend is actually due to other sources of variability unaccounted for in the model, which would result in a greater statistical uncertainty. The residuals of the Brunet's model are shown in Figure 8.

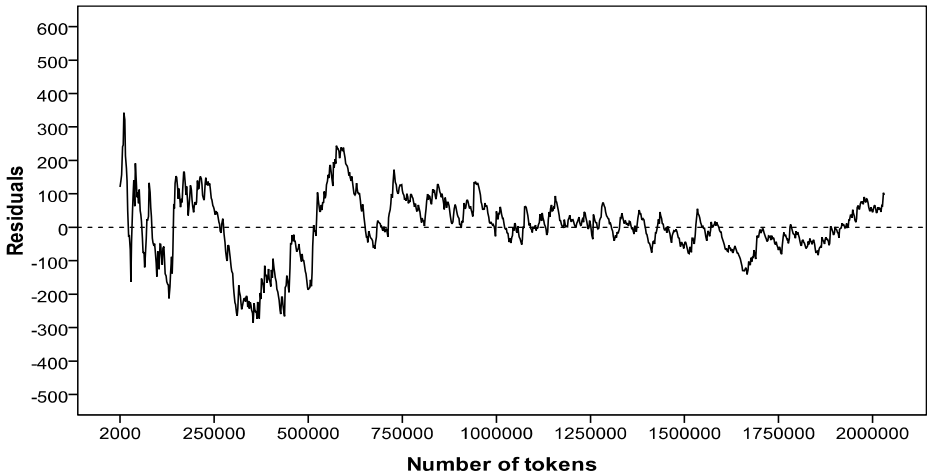

Figure 8. Residuals of the Brunet's model and the observed values

As shown in Figure 8, the residuals are distributed evenly on both sides of the dotted zero line; the biggest residual is below 350. The residuals exhibit no coherent patterns, which means that the Brunet's model provides a very good fit for CDBNC's vocabulary growth.

3.3.4 Hapax Growth and the Hapax-Vocabulary Ratio (HVR)

In this section, the dynamic relationship among the vocabulary growth, hapax, and HVR is examined in each domain of BNC. Figure 9 shows the hapax growth curves of the nine domains of BNC.

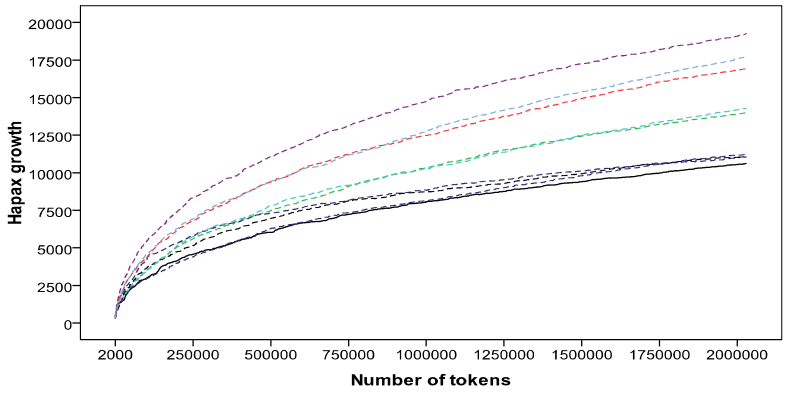

Figure 9. The hapax growth curves of the nine domains of BNC

As shown in Figure 9, the curves represent the hapax growths of the nine domains of the BNC. The solid line is CDBNC and the dotted lines are other eight domains. As the text length keeps increasing, hapax keeps increasing, too. All the hapax growth curves of the nine domains present a trend of increasing as the text length becomes larger and larger. The hapax growth curves from the top to the bottom are arts, world affairs, leisure, applied sciences, imaginative, belief and thought, natural sciences, social sciences, and commerce. The arts in the upper curve increases the fastest, and the curve of the *commerce* is the lowest. It indicates that the hapax growth curve of CDBNC is the smallest among the nine.

In order to make the analysis of hapax and HVR clear, first, we explain the relationship of HVR and the text length through an assumption.

Suppose there is a text T with a length N, a vocabulary size V, consisting of H hapaxes. If the text has only one word, i.e., $N = 1$, then $V = H = N = HVR = 1$. As N gradually increases, this relationship may still maintain for a short while, until N reaches a certain value, from which this relationship starts to change; the HVR would be smaller than 1. As N keeps increasing, the HVR would keep decreasing. As N approaches ∞, all the words in the language would have occurred more than once, and the number of hapaxes would be zero, and so would the HVR. Figure 10 shows the relationship of HVR and text length in the nine domains of BNC.

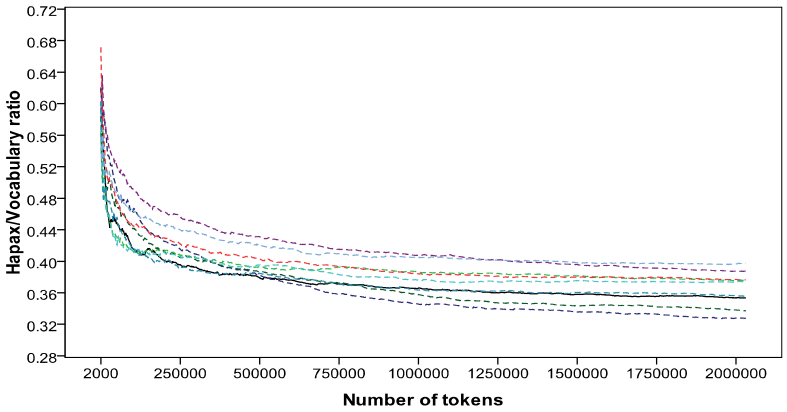

Figure 10. The HVR curves of the nine domains

In Figure 10, the curves of HVR of the nine domains are displayed from the top: world affairs, arts, leisure, applied science, natural science, commerce, social science, imaginative, and belief and thought. It is inferred that HVR declines continuously along with the number of tokens getting larger. What would happen to hapaxes and HVRs when *N* is much larger than 2,000, and approaches 2,000,000? – Table 7 will explain it in detail.

Table 7 displays the initial HVRs, the final HVRs, the initial hapaxes, and the final hapaxes of the nine domains.

As is shown in Table 7, as the tokens of cumulative texts increase, the HVR decreases. The study on CDBNC shows that, when the text length is 2,000, the initial number of hapax is 434, the initial HVR is 0.6191; as the text length approaches 2,000,000, the final number of hapax is 10,622, and the final HVR is 0.3535.

Table 7

Descriptive statistics of HVR. The initial HVRs (HVR_i), the final HVRs (HVR_f), the initial hapax no. (Hap_i), the final hapax no. (Hap_f)

Domain	HVR_i	HVR_f	Hap_i	Hap_f
Commerce	0.6191	0.3535	434	10,622
Applied	0.5648	0.3762	305	14,015
Arts	0.5954	0.3876	362	19,259
Belief and Thought	0.5838	0.3373	296	11,046
Imaginative	0.6208	0.3276	424	11,064
Leisure	0.6714	0.3763	519	16,922
Natural	0.6024	0.3742	444	14,294
Social	0.6035	0.3556	347	11,211
Word	0.5579	0.3973	323	17,731

3.4 Vocabulary Coverage and Textual Coverage

This section examines the vocabulary coverage and textural coverage by CET–4 and CET–6 over CDBNC. Sections 3.4.1 and 3.4.2 describe the vocabulary coverage distribution and the textual coverage distribution of CET–4 and CET–6 on CDBNC respectively. To see whether the vocabulary coverage and textual coverage of CET–4 and CET–6 are normally distributed, a one sample K–S test was performed. In section 3.4.3, the vocabulary not covered by CET–4 and CET–6 is analysed.

3.4.1 Vocabulary Coverage

The mean vocabulary coverage of CET–4 over the 1,000 text samples is 0.7747. It ranges from 0.6112 to 0.9002, with a median of 0.7744 and a mode of 0.7778. The standard deviation is 0.0442, indicating a moderate dispersion. Out of the 1,000 CET–4 vocabulary coverage values, 250 are below 0.7446, 250 over 0.8062, and 500 in between. A one sample K–S test shows that the CET–4 vocabulary coverage over the 1,000 text samples is basically normal, with the K–S Z being 0.664 and the p-value being 0.77. The left panel of Figure 11 displays the vocabulary coverage distribution of CET–4. The mean vocabulary coverage of CET–6 over the 1,000 text samples is 0.8170. It ranges from 0.6563 to 0.9173, with a median of 0.8193 and a mode of 0.7997. The standard deviation is 0.0435, smaller than the one of the coverage of CET–4. Out of the 1000 CET–4 vocabulary coverage values, 250 are below 0.7881, 250 over 0.8491, and 500 in between. As with the coverage of CET–4, a one sample K–S test was used to test the distribution of the CET–6 coverage, and the result shows that the CET–6 vocabulary coverage distribution is basically normal, with the K–S Z being 0.794 and the p-value of 0.554. The right panel of Figure 11 displays the vocabulary coverage distribution of CET–6. Both of the p values are much greater than the significance level 0.05; that is to say the vocabulary coverage distributions of CET–4 and CET–6 are roughly normal. Although CET–6 contains 1,098 more lemmas than CET–4 does, its mean vocabulary coverage is increased by only 0.0423, indicating that the relationship between vocabulary size and its vocabulary coverage is non-linear.

Figure 11 shows that the vocabulary coverage distributions of CET–4 and CET–6 are basically normal. The 95% confidence intervals of the vocabulary coverage distributions of CET–4 and CET–6 are obtained by formula (1):

CET–4: $0.7747101 \pm 1.96 \times 0.0442360957 \approx 0.688, 0.8614$

CET–6: $0.8170302 \pm 1.96 \times 0.0434653568 \approx 0.7318, 0.9022$

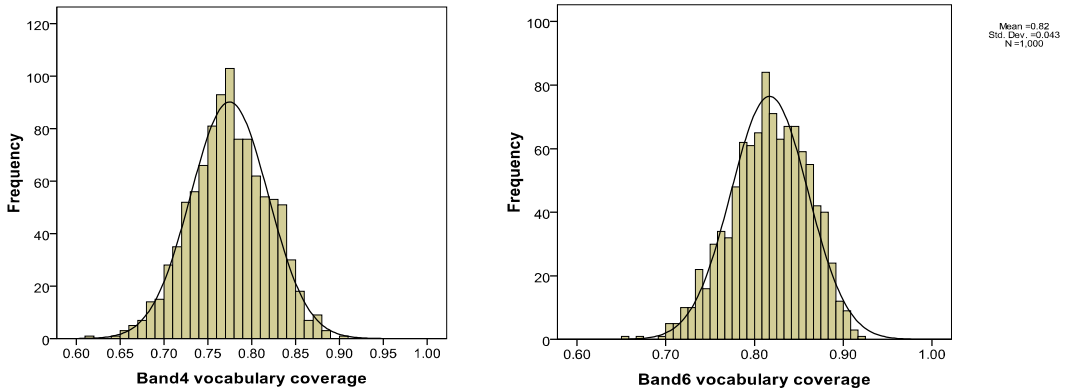

Figure 11. The vocabulary coverages of CET–4 (left panel) and CET–6 (right panel), with normal distribution curves

These intervals provide good estimates of the number of types of any randomly selected texts with roughly 2,000 word tokens; that is, 95% of such texts have types within the above corresponding intervals.

Figure 12 shows the box plots for the vocabulary coverages of CET–4 (left panel) and CET–6 (right panel). The area in the box is where vocabulary coverage converges, while the horizontal line in the box is the mean of vocabulary coverage of the 1,000 samples. Theoretically, the values for vocabulary coverages should be confined within the two horizontal lines outside the box. The number of the outliers – represented by the small circles exceeding this range – is small.

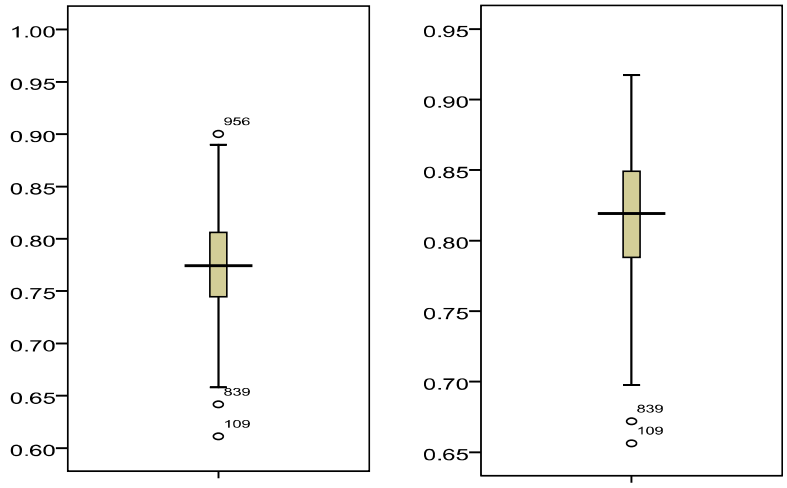

Figure 12. The vocabulary coverages of CET–4 (left panel) and CET–6 (right panel) in the individual samples

3.4.2 Textual Coverage

The textual coverages of CET–4 and CET–6 are considerably higher than the vocabulary coverages of CET–4 and CET–6. For CET–4, its textual coverage over the 1,000

text samples ranges from 0.7563 to 0.964, with a mean of 0.872, a median of 0.8713, and a mode of 0.8377. The standard deviation is 0.0326. Out of the 1,000 CET–4 textual coverage values, 250 are below 0.8493, 250 over 0.8951, and 500 in between. To see whether the individual textual coverage of CET–4 is normally distributed, the one sample K–S test was performed, and the result shows that the textual coverage distribution of CET–4 is basically normal, with the K–S Z being 0.749 and a p-value of 0.63. The left panel of Figure 13 shows the textual coverage distribution of CET–4. The CET–6 textual coverage over the 1,000 text samples ranges from 0.8015 to 0.9686, with a mean of 0.8955, a median of 0.8958, and a mode of 0.9186. The standard deviation is 0.032. 250 of the individual textual coverage are below 0.8727, 250 over 0.9203, and 500 in between. The textual coverage distribution of CET–6 is basically normal, with the K–S Z being 0.786 and the p of 0.568. The right panel of Figure 13 shows the textual coverage distribution of CET–6.

The 95% confidence intervals of the lexical overlap distributions of the individual samples are as follows:

CET–4: $0.8720455 \pm 1.96 \times 0.0326173376 \approx 0.8081, 0.934$

CET–6: $0.895485 \pm 1.96 \times 0.0320488 \approx 0.8327, 0.9583$

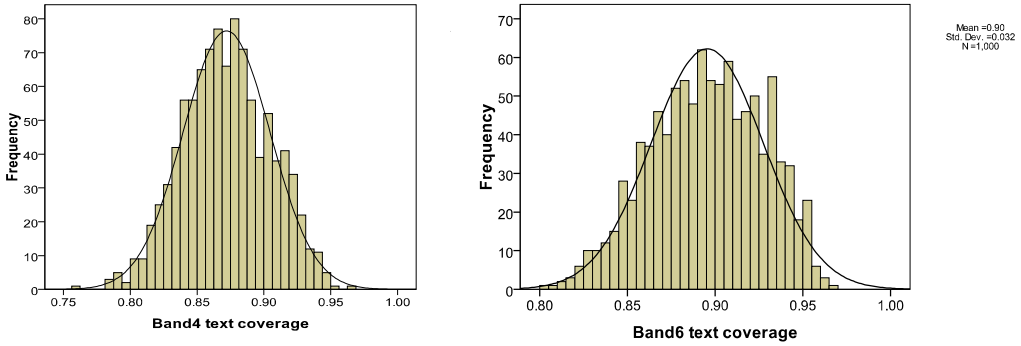

Figure 13. The textual coverages of CET–4 (left panel) and CET–6 (right panel), with the normal distribution curves

In Figure 14, the area in the box is where the textual coverage converges, while the horizontal line in the box is the mean of vocabulary coverage of the 1,000 samples. Theoretically, the values of the textual coverage should be confined within the two horizontal lines outside the box. The outliers, represented by the small circles, are fewer than the ones of the vocabulary coverage.

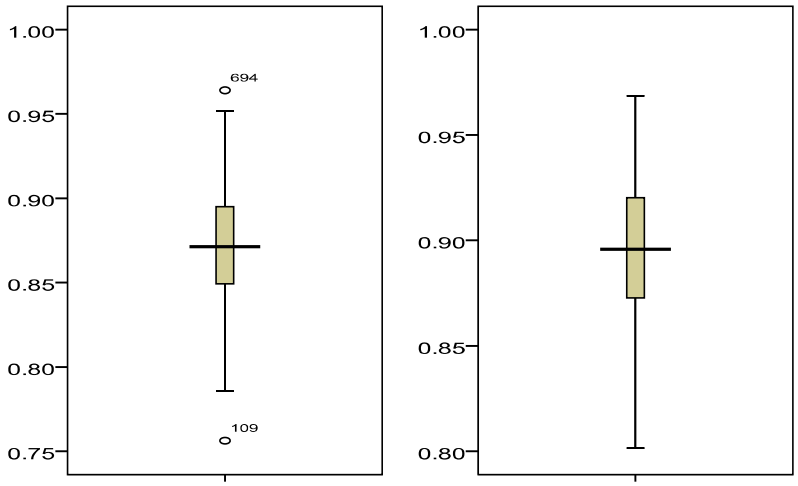

Figure 14. The textual coverages of the CET–4 (left panel) and CET–6 (right panel) individual samples

3.4.3 Vocabulary not Covered by CET–4 and CET–6

Another important finding is the uncovered words by CET–4 and CET–6 calculated in CDBNC, which account for a very large proportion of the vocabulary. The total number of the words uncovered by CET–4 and CET–6 are 25,697 and 24,650, respectively.

The frequency of these uncovered words is from 170 to 941 in CDBNC. Those words are very popular and important in the commercial activities. Most of them are professionalisms and common words which often occur in commercial written and spoken English; for example, the words *inflation*, *asset*, *consumer*, *shareholder*, *organization*, *investor*, and so on.

4 Conclusion

In this corpus-based research, through analysing and comparing the data of CDBNC and other eight domains of BNC, we discussed the characteristics of CDBNC and discovered the differences between CDBNC and other eight domains. The major findings of the study are summarized here.

Firstly, through the analysis and calculation on the 2,000,000-word samples, a general picture of CDBNC is revealed. Being a member of the nine domains of BNC, CDBNC has the smallest vocabulary size (30,044), the hapax of 10,622, and the TTR of 0.015. The dispersion of the vocabulary size of the nine domains is large, from 30,044 to 49,685. The number of hapax spans from 10,622 to 19,259, and TTR from 0.015 to 0.0241.

Secondly, compared with other eight domains, the entropy and perplexity of CDBNC are smaller. We find that the lowest perplexity is corresponding to the lowest entropy. Entropy represents the amount of information in language; the greater the

entropy value is, and the greater the amount of information will be.

Thirdly, TTR is normally distributed in the texts of the same length; the TTR decreases as the length of text increases. Tuldava's model fits the observed data very well, and it is very good for the description of the relationship between text length and TTR.

Fourthly, the Brunet's model was tested on the vocabulary growth of the samples from CDBNC, at a 2,000-word interval. It has been found that the Brunet's model provides a very good fit for CDBNC's vocabulary growth.

Lastly, the vocabulary coverages and individual textual coverages of CET–4 and CET–6 are normally distributed. The mean vocabulary coverage of CET–4 over 1,000 text samples is 0.7747. The mean vocabulary coverage of CET–6 over 1,000 text samples is 0.8170. The textual coverages of CET–4 and CET–6 are considerably higher than the vocabulary coverages of CET–4 and CET–6. The mean textual coverage of CET–4 is 0.872. The mean textual coverage of CET–6 is 0.8955.

To sum up, the major findings of the present research can enrich empirical study in the field of quantitative linguistics. As one of the nine domains of BNC, CDBNC has a large developing space in theoretical study, teaching and learning reform, textbook complication, etc.

References

Chujo, K. (2004). Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list. *English Corpora under Japanese Eyes*: 215–237.

Devore, J. (2000). *Probability and Statistics*. Pacific Grove: Brooks/Cole.

Fan F. (2008). A corpus-based study on random textual vocabulary coverage. *Corpus Linguistics and Linguistic Theory* 4(1), 1–17.

Fan, F. (2010). An asymptotic model for the English hapax/vocabulary ratio. *Computational Linguistics* 36(4), 620–631.

Fan, F. (2010a). *Data Processing and Management for Quantitative Linguistics with Foxpro*. Lüdenscheid: RAM-Verlag.

Fan, F. (2010b). *Quantitative Linguistic Computing with Perl*. Lüdenscheid: RAM-Verlag.

Gries, S. Th. (2007). Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2), 109–151.

Jurafsky, D. & Martin, J. (2000). *Speech and Language Processing, an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River: Prentice Hall.

Mendenhall, W., & Scheaffer, R.L. (1973). *Mathematical Statistics with Applications*. Massachusetts: Duxbury Press.

Nagy. W. (1997). On the role of context in first– and second–language vocabulary learning. In Schmitt, N., & McCarthy, M. (Eds)., *Vocabulary: Description, Acquisition and Pedagogy* (pp. 64–83). Cambridge: Cambridge university press.

Nation, I. (1990). *Teaching and Learning Vocabulary.* New York: Newbury House Publishers.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program* 14, 116–132.

Radford, A., M. Atkinson, D. Britain, H. Clahsen and A. Spencer. (1999). *Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Glottometrics 47, 2019, 36-51

Inter-textual Vocabulary Growth Patterns for Marine Engineering English

Jingjie Li[1]

Abstract. This paper explores the three fundamental issues concerning the inter-textual vocabulary growth patterns for marine engineering English. These are distributions of vocabulary sizes of individual texts, vocabulary growth models, and newly occurring vocabulary distributions of cumulative texts. The research is carried out on the basis of the MEE corpus. The vocabulary sizes of individual texts with the same text size conform closely to the normal distribution. Four existing models (Brunet's model, Tuldava's model, Guiraud's model, and Herdan's model) are tested against the empirical growth curve for marine engineering English. A new growth model is derived from the logarithmic function and the power law. The theoretical mean vocabulary size and the 95% upper and lower bound values are calculated as functions of the sample size. The new growth model can make accurate estimates not only on the vocabulary size and its intervals for a given textbook, but also on the volume of texts that are needed to produce a particular vocabulary size.

Keywords: *MEE corpus, inter-textual vocabulary growth, growth models*

1 Introduction

Due to the exclusivity of vocabulary for the English language, effective vocabulary acquisition is of particular importance for EFL learners (Ellis, 2004) who frequently acquire impoverished lexicons despite years of formal study. As Smith (1999:50) once commented on the significance of vocabulary, "the lexicon is of central importance and is even described as being potentially the locus of all variation between languages, so that apart from differences in the lexicon, there is only one human language." The mushrooming amount of experimental studies and pedagogical publications also demonstrates beyond all doubt that the field of vocabulary studies is now anything but a neglected area (Schmitt, 2002). Since lexicons are items that should be acquired in certain quantities at certain rates, it is essential for both teachers and learners to be aware of the characteristics of inter-textual vocabulary growth patterns, such as word frequency distributions and vocabulary growth models. These characteristics are of practical significance not only for the resolution of a series of problems in vocabulary

[1] Correspondence to Jingjie Li, College of Foreign Languages, Donghua University, Shanghai, China. Email: lijingjie@dhu.edu.cn

acquisition, but also for the theoretical explanation of some important issues in lexical statistics.

A great number of attempts have been made to construct an appropriate mathematical model that would express the dependence of the vocabulary size *V(N)* on the sample size *N*. Some models describe the vocabulary growth pattern for general English with sufficient precision, but none of them have ever been tested against the vocabulary development for marine engineering English. Therefore, in this paper, four existing mathematical models (Brunet's model, Tuldava's model, Guiraud's model, and Herdan's model) will be evaluated with regard to fitting the empirical growth curve for marine engineering English. A new vocabulary growth model will be constructed by multiplying the logarithmic function and the power law. Three fundamental issues concerning the inter-textual vocabulary growth pattern will also be explored in this paper. These are distributions of vocabulary sizes of individual texts, vocabulary growth models, and newly occurring vocabulary distributions of cumulative texts.

2 Methods and Materials

2.1 Key Concepts

One of the key concepts in morphology, as well as in lexical statistics, is that of *word*. Although many people take this concept for granted, the definition of the term *word* is really ambiguous. Philosopher C. S. Peirce made the distinction between word types being distinct letter strings and word tokens being instances of a type (Lyons, 1995). Lexicographer R. Beard replaced *word* with two separate notions: lexeme and word form (Matthews, 2000). He defines lexeme as the abstract unit underlying the smallest unit in the lexical system of a language that appears in different grammatical contexts. In this sense, *cat* and *cats* are two separate word forms (word tokens) of one lexeme (word type) *CAT*.

In the present study, the vocabulary size refers to the number of lexemes in a given text, and the sample size is the total number of word forms, some of which are used more than once. Tokenization is the process of counting the frequencies of each word token of a given text; lemmatization is the process of calculating the total number of word types of a text, the frequencies of their occurrences.

2.2 The MEE Corpus

The MEE (Marine Engineering English) Corpus was used as the data source to establish the inter-textual vocabulary growth patterns. It contains 1,030,522 word tokens (word instances) and 18,766 word types (different words). 959 text samples were collected in the corpus; the length of each text varies from 349 to 2,070 word tokens, and the average text size reaches 1,075 tokens. The corpus texts were all selected from today's most influential marine engineering English materials that are produced or published either

in Britain, or in the USA. The publishing time ranges from 1987 to 2004. About 85% of the text samples were taken from papers and books, 10% from lectures and discussions, and the remainder from such sources as brochures and manual instructions. The corpus texts were also proportionally selected from various constituent subject fields of marine engineering English, including marine diesel engine, marine power plant, electrical installation, steering gear, marine refrigeration, gas exchange, lubricating system, propelling system, maintenance and repair, etc.

2.3 Procedures for Data Processing

With the aid of the statistical software SPSS and data-managing system FoxPro, the task of data processing was carried out through the following steps.

The first step was to evaluate the vocabulary distributions of individual texts. The text samples of the MEE corpus were grouped into four sub-corpora in terms of the sample size. Sub-corpus One consists of the individual texts that have roughly 500 word tokens, Sub-corpus Two of the texts with roughly 1,000 tokens, Sub-corpus Three of the texts with about 1,500 tokens, and Sub-corpus Four of the texts with 2,000 tokens. Detailed information for the four sub-corpora is presented in Table 1.

Table 1
Sampling information for the four sub-corpora of the MEE corpus

CORPUS NAME	TOKENS	TYPES	NUMBER OF TEXTS	AVERAGE TEXT SIZE
Sub-corpus One	252,444	11,567	436	579
Sub-corpus Two	254,694	10,871	227	1,122
Sub-corpus Three	261,908	10,551	164	1,597
Sub-corpus Four	261,476	10,313	131	1,996

The distribution of vocabularies of the texts was explored for each sub-corpus. Take Sub-corpus One as an example. The first FoxPro programme was designed to perform tokenization and lemmatization to each text sample, and 466 values of vocabulary sizes were obtained. SPSS was then used to plot the histogram of vocabularies of the texts for Sub-corpus One, and to calculate the 95% confidence interval for the population vocabulary of individual texts. By the same means, the other three sub-corpora were statistically analyzed. The four output histograms will be presented in Section 3.1 for further discussion.

The second step was to explore inter-textual vocabulary growth models for marine engineering English. All the text samples of the MEE corpus were regrouped into two sets, the Sample Set and the Test Set, each of which has equally 500 thousand word tokens. The Sample Set served as the database for evaluating the four existing mathematical models and constructing a new vocabulary growth model. The Test Set was

used to verify the descriptive power of the new model with regard to fitting the empirical growth curve of marine engineering English. The second FoxPro programme was designed to select texts randomly from the MEE corpus and rearrange them to constitute the Sample Set and the Test Set. Detailed information for the two sets is listed in Table 2.

Table 2
Descriptive data for the Sample Set and the Test Set of the MEE corpus

NAME	TOKENS	TYPES	NUMBER OF TEXTS	AVERAGE TEXT SIZE
Sample Set	513,658	15,583	479	1,072
Test Set	516,864	15,436	480	1,077

The final step was to explore the distributions of newly occurring vocabularies of cumulative texts. The newly occurring vocabulary size $V(N)_{new}$ was calculated and plotted as a function of the sample size N for both the Sample Set and the Test Set. The third FoxPro programme was written to calculate the number of new word types and the corresponding number of tokens; SPSS was used to plot the scatter-grams of $V(N)_{new}$ against N for the two sets separately.

3 Results and Discussion

3.1 Vocabulary Distributions of Individual Texts

Tables 3 and 4 list the statistical information for the distributions of vocabularies of individual texts with the respective text sizes of 500, 1000, 1500, and 2000 word tokens.

Tables 3
General information for the vocabulary distributions of individual texts in the four sub-corpora

SUB-CORPUS NAME	NUMBER OF TEXTS	SUB-CORPUS SIZE	VOCABULARY SIZE
Sub-corpus One	436	252,444	11,567
Sub-corpus Two	227	254,694	10,871
Sub-corpus Three	164	261,908	10,551
Sub-corpus Four	131	261,476	10,313

Tables 4
Statistical data of the vocabulary distributions of individual texts
for the four sub-corpora

SUB-CORPUS NUMBER	TEXT MEAN SIZE	TEXT MEAN VOCABULARY	STANDARD DEVIATION	MAXIMUM VOCABULARY	MINIMUM VOCABULARY
Sub-corpus One	500	212	29.34	290	113
Sub-corpus Two	1,000	344	51.20	487	174
Sub-corpus Three	1,500	461	73.19	645	220
Sub-corpus Four	2,000	562	92.98	776	270

The four sub-corpora have roughly the same corpus sizes, but their vocabulary sizes vary regularly, with the values decreasing steadily from Sub-corpus One to Sub-corpus Four ($V(N)_{Sub\text{-}corpus\ One}$ = 11,567, $V(N)_{Sub\text{-}corpus\ Two}$ = 10,871, $V(N)_{Sub\text{-}corpus\ Three}$ = 10,551, $V(N)_{Sub\text{-}corpus\ Four}$ = 10,313). Sub-corpus One has the smallest mean text size, 500 word tokens, but it has the largest vocabulary size of the four sub-corpora. Sub-corpus Four has the biggest average text size, 2,000 word tokens, but it is characterized by the least vocabulary richness.

Figure 1 plots the four histograms of the vocabulary distributions of individual texts with the respective text sizes of 500, 1,000, 1,500 and 2,000 word tokens. The bell-shaped line represents the corresponding theoretical normal curves.

The One-sample Kolmogorov–Smirnov Test is applied to check whether the vocabulary sizes of individual texts in each sub-corpus come from the normal population distribution. The Kolmogorov–Smirnov values under "Asymp. Sig." are 0.718 for Sub-corpus One, 0.611 for Sub-corpus Two, 0.543 for Sub-corpus Three, and 0.517 for Sub-corpus Four. All these values are much greater than the significance level of 0.05. Therefore, the distributions of vocabulary sizes of individual texts conform very closely to the normal distribution. As we can see, the "Asymp. Sig." values decrease steadily from Sub-corpus One to Sub-corpus Four (i.e., 0.718 > 0.611 > 0.543 > 0.517). Thus, the normal fit to the four histograms (Figure 1) is in a decreasing sequence from the upper left panel to the lower right panel. The normal curve fits Sub-corpus One best (upper left panel of Figure 1) because Sub-corpus One contains the largest number of text samples (436 texts in Sub-corpus One, 227 texts in Sub-corpus Two, 164 in Sub-corpus Three, and 131 in Sub-corpus Four). In the view of Fleiss (1981), the more samples are randomly selected, the closer the sample mean is to the population mean; the fewer samples are measured, the less accurate the estimate of population is, as a result of the sampling error and the extremes.

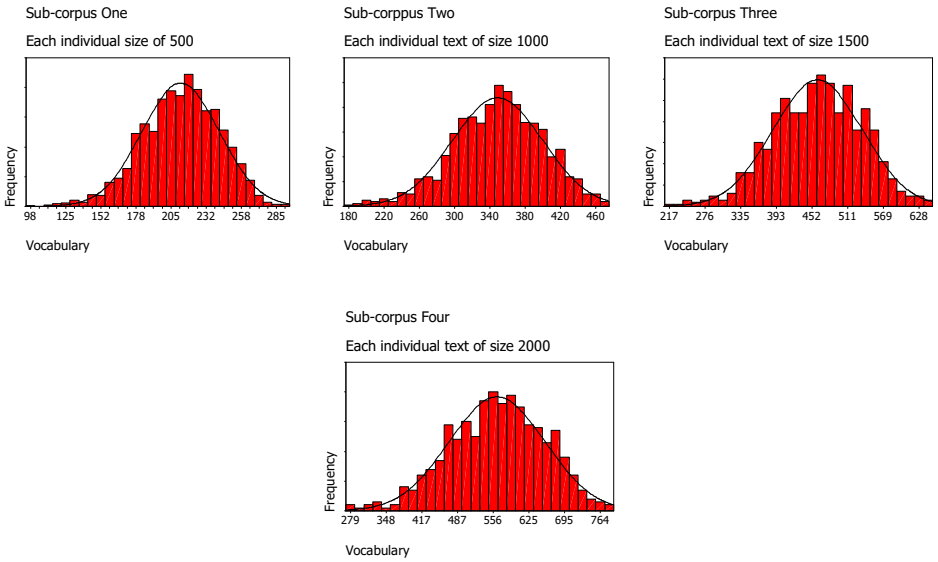

Figure 1. Four frequency histograms of vocabulary distributions of individual texts and their theoretical normal curves (The panels from left to right plot the vocabulary distributions of texts with 500, 1,000, 1,500 and 2,000 word tokens, respectively.)

Instead of being symmetric about their highest points, the histograms of Figure 1 are distributed lopsidedly in the negative direction. The values of skewness are: –0.310 for the upper left panel, –0.283 for the upper right panel, –0.234 for the lower left panel, and –0.222 for the lower right panel. Thus, the vocabulary distributions of individual texts are approximately normal, with a slight degree of negative skewness.

Expression 1 is known as the tolerance interval formula. It is used to capture at least 95% of the vocabulary sizes of individual texts with the same word tokens:

$$tolerance\ interval = \overline{x} \pm (tolerance\ critical\ value) \times s \tag{1}$$

For reasonably large samples, $\overline{x}$ is the mean value of a random sample taken from a normally distributed population; s is the standard deviation of the random sample. Replacing $\pm$ by + gives an upper tolerance bound, and using $-$ in place of $\pm$ results in a lower tolerance bound. For capturing the possible values in a normal population distribution with a confidence level 95%, the tolerance critical value is 1.96. Consider Figure 1 for Sub-corpus One, where $n_{sample} = 436$ (number of text samples), $\overline{x} = 212$, $s = 29.34$, and the vocabulary distributions of individual texts are approximately normal. For a confidence level of 95%, the two-sided tolerance interval is –

$$\begin{aligned} 212 \pm 1.96 \times 29.34 &= 212 \pm 57.5064 \\ &= [154.4936,\ 269.5064] \\ &\approx [154,\ 270] \end{aligned}$$

The particular vocabulary size varies from one sample to another, but we are highly confident that at least 95% of the individual texts with 500 word tokens have the vocabulary sizes between 154 and 270 word types.

By the same means, the respective 95% tolerance intervals are calculated for the possible text vocabularies for the other three sub-corpora: [248, 449] for Sub-corpus Two, [317, 605] for Sub-corpus Three, and [381, 746] for Sub-corpus Four.

3.2 Vocabulary Growth Models for Marine Engineering English

3.2.1 Test of the Existing Vocabulary Growth Models

Table 5 lists part of the statistics to plot the observed and expected growth curves for the Sample Set using Brunet's model, Tuldava's model, Guiraud's model, and Herdan's model. The respective formula of the four models is displayed in Table 7. Refer to Baayen (2001) for detailed explanation of the existing models.

Figure 2 plots the empirical vocabulary growth curve (tiny circles) for the Sample Set, as well as the corresponding expectations (solid line) using the aforementioned models.

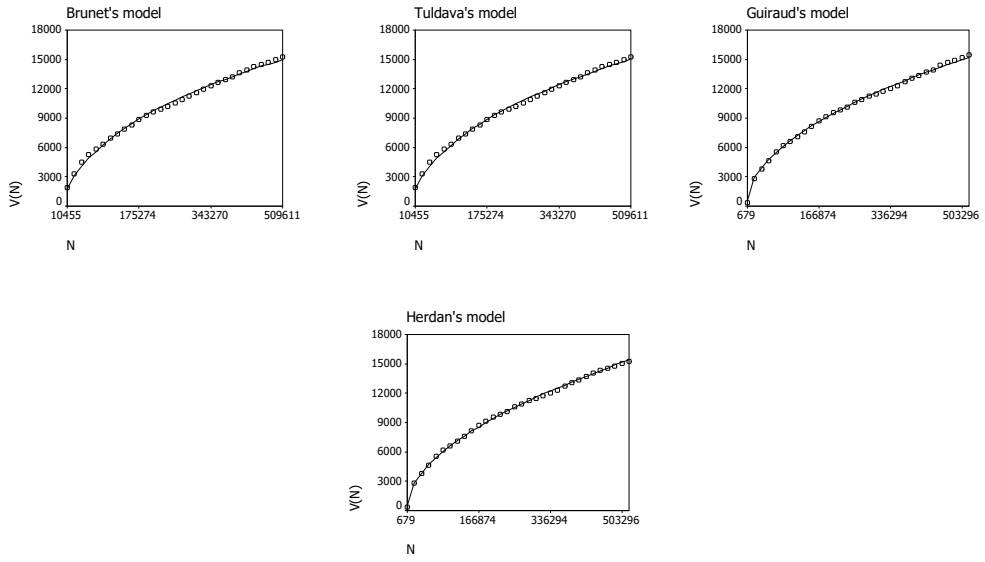

Figure 2. Observed (tiny circles) and expected (solid line) vocabulary growth curves for the Sample Set

Table 5

Observed vocabulary sizes for the Sample Set of the MEE corpus, and the predicted values (*Pre_*) and the corresponding residuals (*Resi_*) using Brunet's model, Tuldava's model, Guiraud's model, and Herdan's model

Tokens	*Types*	*Pred_B*	*Resi_B*	*Pred_T*	*Resi_T*	*Pred_G*	*Resi_G*	*Pred_H*	*Resi_H*
25,202	3,169	2,947	222	3,001	168	3,436	-267	3,225	-56
50,471	4,658	4,500	158	4,513	145	4,863	-205	4,644	14
75,420	5,719	5,679	40	5,670	49	5,944	-225	5,734	-15
100,929	6,611	6,688	-77	6,667	-56	6,876	-265	6,682	-71
125,251	7,408	7,531	-123	7,502	-94	7,660	-252	7,484	-76
150,861	8,151	8,326	-175	8,295	-144	8,407	-256	8,252	-101
175,622	9,065	9,028	37	8,996	69	9,071	-6	8,937	128
200,998	9,594	9,692	-98	9,662	-68	9,704	-110	9,593	1
225,305	10,258	10,286	-28	10,258	0	10,274	-16	10,186	72
250,376	10,871	10,862	9	10,838	33	10,830	41	10,766	105
275,847	11,303	11,414	-111	11,395	-92	11,368	-65	11,328	-25
300,330	11,695	11,918	-223	11,904	-209	11,862	-167	11,845	-150
325,548	12,301	12,413	-112	12,405	-104	12,349	-48	12,357	-56
350,479	12,817	12,882	-65	12,879	-62	12,814	3	12,846	-29
375,305	13,315	13,329	-14	13,333	-18	13,260	55	13,316	-1
400,817	14,002	13,772	230	13,782	220	13,703	299	13,783	219
425,830	14,306	14,190	116	14,206	100	14,124	182	14,228	78
450,526	14,712	14,589	123	14,612	100	14,528	184	14,768	-56
475,815	15,015	14,984	31	15,014	1	14,930	85	15,082	-67
500,185	15,363	15,354	9	15,390	-27	15,308	55	15,483	-120

Brunet's model fits the empirical growth curve well. The value of R^2 (the coefficient of determination) reaches 99.876%. The upper left panel shows that Brunet's model underestimates the observed vocabulary sizes both at the beginning and towards the end of the growth curve. The predicted values are presented in the third column of Table 5.

The fit of Tuldava's model (the upper right panel) is close to the empirical growth curve. The R^2 is 99.910%. Tuldava's model tends to underestimate the observed vocabulary sizes at the beginning of the curve. The predicted values are listed in the fifth column of Table 5.

Guiraud's model fit the empirical growth curve with less precision. The R^2 is 99.812%, the lowest value among the four models. The lower left panel shows that Guiraud's model overestimates the observed vocabulary sizes from the beginning of the growth curve to nearly the middle. The expected values are listed in the seventh column of Table 5.

Though similar in expression to Guiraud's model, Herdan's model describes the empirical growth curve best of the four models. The R^2 reaches the highest value – 99.942%. However, the lower right panel shows that Herdan's model tends to overestimate the observed vocabulary sizes towards the end of the growth curve. The expected values are presented in the ninth column of Table 5.

3.2.2 A New Vocabulary Growth Model for Marine Engineering English

Based on the lexico-statistical study of the Sample Set, a new mathematical model (Expression 2) is suggested to fit the vocabulary growth curve of marine engineering English.

$$V(N) = \alpha \times \log N \times N^{\beta} \tag{2}$$

The new model is constructed by multiplying the logarithmic function and the power law. Parameter α is the coefficient of the whole expression; parameter β is the exponential of the power function part of the model. Parameters α and β do not have fixed values; they have to change slightly with specific sample sizes to realize sufficient goodness of fit.

The theoretical considerations for proposing the new model are as follows:

- Brunet's model is in essence a complex logarithmic function, with $\log_W N$ being the independent variable, and $\log_W V(N)$ the dependent variable. It tends to *underestimate* the observed vocabulary sizes both at the beginning and towards the end of the empirical growth curve.
- Herdan's model is a generalized power function. It tends to *overestimate* the observed values towards the end of the vocabulary growth curve.
- The mathematical combination of the logarithmic function and the power law may provide a good fit to the empirical growth curve for marine engineering English.

Figure 3 plots the dependence of *V(N)* on *N* for the Sample Set (tiny circles) and the corresponding expectations (solid line), using the new mathematical model.

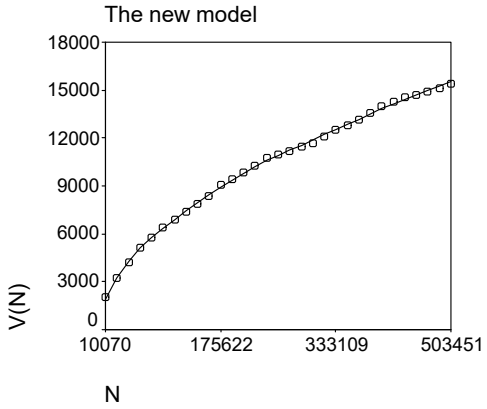

Figure 3. Vocabulary size *V(N)* as a function of the sample size *N*, measured in units of texts

The new mathematical model provides a very good fit to the empirical growth curve. The value of R^2 reaches 99.945%, higher than any other vocabulary growth models, particularly than Brunet's model and Herdan's model. The respective values of parameters α and β are 3.529696 and 0.442790 for the Sample Set.

3.2.3 Goodness-of-Fit Test for the New Vocabulary Growth Model

The Test Set of the MEE corpus is used to verify the descriptive power of the new vocabulary growth model. Figure 4 plots the empirical growth curve for the Test Set (tiny circles) and the corresponding expectations, using the new mathematical model (solid line) with α = 3.529696 and β = 0.442790.

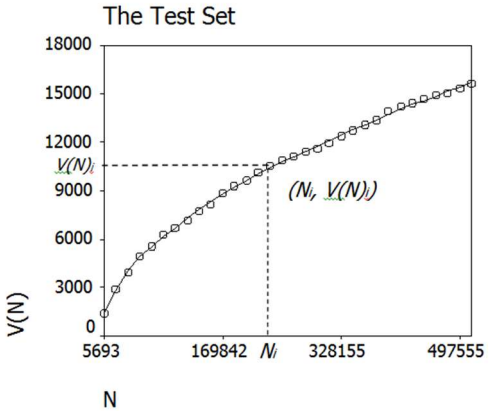

Figure 4. Dependence of the vocabulary size *V(N)* on the sample size *N*, measured in units of texts

The Test Set has the same corpus size as the Sample Set, and they both represent the contemporary marine engineering English. Thus, the values of parameters α and β, derived from the Sample Set, also apply to the growth curve for the Test Set. Figure 4 shows that the fit of the new model is very close to the empirical growth curve. The value of R^2 reaches 99.917%.

The vocabulary sizes of individual texts basically conform to the normal distribution. Thus, for each point (N_i, $V(N)_i$) on the theoretical growth curve fitting the new model (Figure 4), $V(N)_i$ is the mean value of the normally distributed vocabulary sizes of individual texts with the same text size N_i. The upper and lower bound values of each point (N_i, $V(N)_i$) are calculated for a confidence level of 95%, using Expression 3 that is derived from the tolerance interval formula.

$$\begin{aligned} V(N)_{upper} &= V(N)_i + (tolerance\ critical\ value) \times s \\ V(N)_{lower} &= V(N)_i - (tolerance\ critical\ value) \times s \end{aligned} \tag{3}$$

The standard deviation s determines the extent to which each normal curve spreads out

around its mean value $V(N)_i$. The texts of the MEE corpus are randomly selected and rearranged ten times to constitute ten sets of text samples, based on which ten sets of vocabulary sizes and the corresponding sample sizes are obtained (see the Appendix), and the standard deviation mean $\overline{s}$ is calculated for each given point $(N_i, V(N)_i)$ on the expected growth curve. Devore proposed that when $n_{sample} = 10$ (number of samples), the tolerance critical value is 3.379 for capturing at least 95% of the vocabulary sizes in the normal population distribution.

SPSS calculates the two-sided tolerance interval for each given point $(N_i, V(N)_i)$ on the expected growth curve with a confidence level of 95%. Figure 5 plots the empirical growth curve for the Test Set and the upper and lower bounds (dashed line) using the new mathematical model. The solid lines show that all the observed vocabulary sizes fall within the 95% two-sided tolerance bounds, with no exceptions.

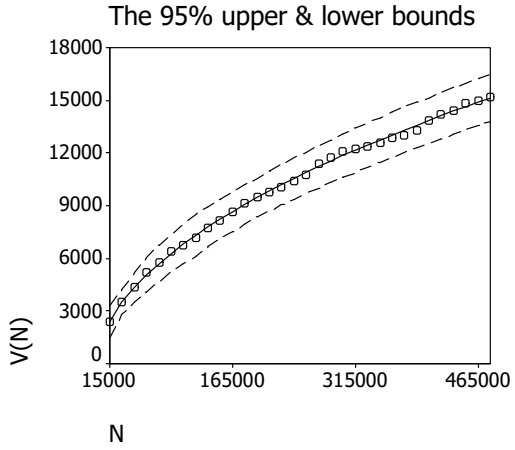

Figure 5. Vocabulary sizes *V(N)* as functions of the sample size *N*, measured in units of texts

3.2.4 Analysis of Residuals for the New Vocabulary Growth Model

An effective approach to assessment of model adequacy is to compute the residual errors. A residual error is the difference between an empirical vocabulary size and the corresponding theoretical value predicted by a model. The residuals from fitting the growth models will be analysed from two aspects: variance of the residuals and the degree to which there are coherent patterns in the residuals.

The variance of residuals shows how accurately the growth models give predictions of the observed vocabulary values. A smaller variance implies a better model fit to the empirical growth curve and lower statistical uncertainty of the model components. SPSS calculates the mean square (a measure of variance) of residuals on the basis of each vocabulary growth model (Table 6).

The new mathematical model fits the empirical growth curve best, with the mean square value of residuals being the lowest. Herdan's model, Tuldava's model, Brunet's model and Guiraud's model follow in a sequence of decreasing goodness-of-fit.

Table 6

Mean square values of residuals on the basis of fitting the five models to the empirical vocabulary growth curve for the Test Set

Brunet's model	Tuldava's model	Guiraud's model	Herdan's model	New model
19,484	13,371	22,185	11,303	10,512

In the view of Scheaffer (1973), if the model makes correct predictions of the observed values, there should be no coherent patterns in the residuals. The greater the coherent structure of the residuals, the greater the chance that the observations deviate from their expected values. Figure 6 is the residuals plot against the sample size on the basis of the five vocabulary growth models.

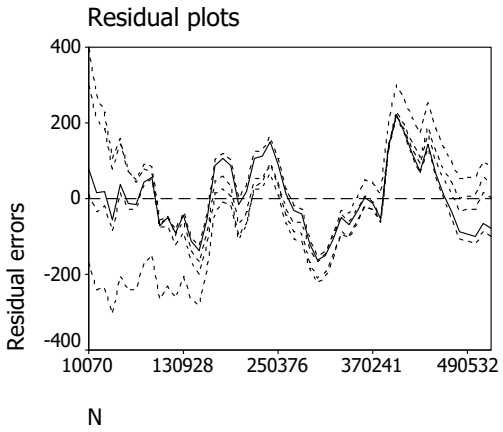

Figure 6. Residuals as functions of the sample size *N*, measured in units of texts (The solid line represents the residuals from the new model fit, the dotted lines the residuals of fitting the other four growth models, and the dashed line the expected value for each residual error.)

The residuals from the new model fit exhibit no coherent patterns (solid line); they do not change systematically as the sample size is increased. The residuals are randomly distributed around their expected value 0 (dashed line) according to a normal distribution. All but a very few residuals lie between –160 and 160.

The residuals of the other four growth models (dotted lines) exhibit coherent patterns, such as a curved pattern for Brunet's model and an increasing function for Guiraud's model. The respective residuals fall within the interval of [–223, 397] for Brunet's model, [–209, 300] for Tuldava's model, [–305, 299] for Guiraud's model, and [–150, 220] for Herdan's model.

The new model accounts for the empirical growth curve best of the five models for both the Sample Set and the Test Set. Therefore, it is reasonable to conclude that the new model gives accurate predictions of the vocabulary growth patterns for marine

engineering English. Tables 7 and 8 present the detailed information and the basic statistics for the goodness of each model fit.

Table 7
Expressions and parameters of the vocabulary growth models for the Sample Set and the Test Set

MODELS	EXPRESSIONS	PARAMETERS
Brunet's model	$\log_w V(N) = \frac{1}{\alpha}\log_w(\log_w N)$	α =18.21789; β = 0.15653
Tuldava's model	$V(N) = Ne^{-\alpha(\ln N)^{\beta}}$	α = 0.02584; β = 1.90470
Guiraud's model	$V(N) = \alpha\sqrt{N}$	α = 21.64422
Herdan's model	$V(N) = \alpha N^{\beta}$	α = 15.75641; β = 0.52506
New model	$V(N) = \alpha \times \log N \times N^{\beta}$	α = 3.529696; β = 0.442790

Table 8
Goodness-of-fit statistics (R^2 and mean square values) for the five models with regard to approximating the empirical growth curves for the Sample Set and the Test Set

MODELS	THE SAMPLE SET		THE TEST SET	
	R^2	MEAN SQUARE	R^2	MEAN SQUARE
Brunet's model	99.876%	17,970	99.847%	19,484
Tuldava's model	99.910%	13,076	99.895%	13,371
Guiraud's model	99.812%	27,273	99.826%	22,185
Herdan's model	99.940%	8,395	99.911%	11,303
New model	99.945%	7,845	99.917%	10,512

3.3 Newly Occurring Vocabulary Distributions of Cumulative Texts

As the sample size $N_{i,}$ increases, the new vocabulary $V(N)_{new}$ that a new input text produces decreases. At any given N_i, the new vocabulary $V(N)_{new}$ of a sample with N word tokens can be estimated with Expression 4.

$$V(N)_{new} = V(N_i + N) - V(N_i) \quad (4)$$

Figure 7 plots the newly occurring vocabulary sizes $V(N)_{new}$ as functions of the sample size N for the Sample Set and the Test Set.

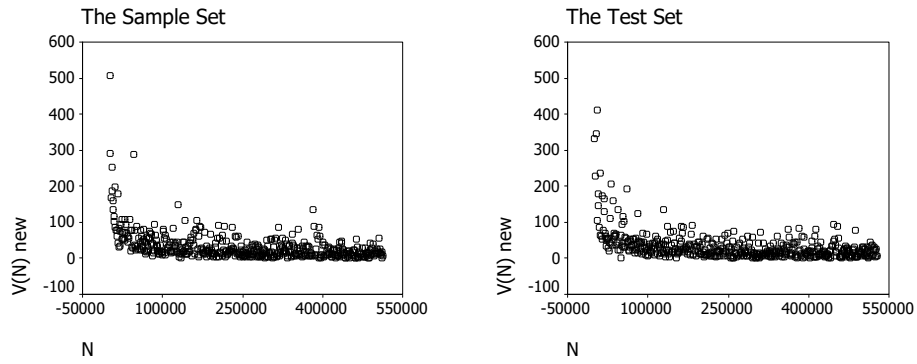

Figure 7. Scatter-grams of $V(N)_{new}$ against N for the Sample Set (left panel) and the Test Set (right panel)

The left panel for the Sample Set and the right panel for the Test Set reveal surprisingly similar distributions of newly occurring vocabulary sizes. The new vocabulary $V(N)_{new}$ is a rapidly decreasing function of the sample size N with a long tail of the low values of $V(N)_{new}$. Initially, $V(N)_{new}$ decreases sharply to some steady-state value that decreases slowly and smoothly. Then, as N increases, the empirical values of $V(N)_{new}$ are distributed around the stable value in a wide dispersion till the end of the scatter plot.

The first input text of the left panel produces more than 500 new word types. Then $V(N)_{new}$ decreases sharply as N increases. When the sample size reaches 62,305 word tokens, the value of $V(N)_{new}$ has diminished to less than 150. However, from N = 62,305 onwards till the end of the scatter plot, the newly occurring vocabulary size decreases very slowly and the observed values of $V(N)_{new}$ are distributed in a fluctuant way. The last input text produces 6 new word types, which proves that new vocabulary still occurs even when the sample size has reached 500 thousand word tokens.

4 Conclusion

This paper explores the three fundamental issues concerning the vocabulary growth patterns for marine engineering English. They are inter-textual vocabulary distributions of individual texts, vocabulary growth models, and newly occurring vocabulary distributions of cumulative texts.

Four sets of text samples with 500, 1,000, 1,500, and 2,000 word tokens, respectively, were selected from the MEE corpus to explore the vocabulary distributions of individual texts. Through a series of statistical procedures, the vocabularies of individual texts are proven to be normally distributed, though the goodness of fit varies with the number of text samples. Four mathematical models are verified and analysed with regard to fitting the empirical growth curve for marine engineering English. A new growth model $V(N) = \alpha \times \log N \times N^{\beta}$ is constructed by multiplying the logarithmic function and the power law. Parameter α is the coefficient of the whole expression; parameter β is the exponential of the power function part of the model. The new model

describes the empirical growth curve best of the five models for marine engineering English, with the R^2 reaching the highest value and the mean square being the lowest. The upper and lower tolerance bounds are calculated to capture at least 95% of the possible vocabulary sizes in the normal population distribution. The residuals from fitting the five growth models are also compared and analysed.

The present study has been exploratory in nature, and some difficult issues have not yet been tackled adequately. These issues would leave the field open to further fruitful research. First and foremost, the new vocabulary growth model was constructed on the basis of studies of the Sample Set, which consists of 480 individual texts, totalling about 500,000 word tokens. The model has been verified to provide a good fit to the empirical vocabulary sizes within the boundary of the sample size of not more than 500,000 word tokens. However, the possibilities of extrapolation of this new growth model in the direction of larger than 500,000 tokens need further consideration and verification. Second, it has been demonstrated in this paper that the vocabulary sizes of individual texts basically conform to the normal distribution. However, to be more exact, the vocabulary size distributions are lopsided in the negative direction to a slight degree. Further research may be needed to illustrate this property from both the theoretical and the empirical perspectives.

References

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Berlin: Springer Science & Business Media.

Ellis, R. (2004). *Second Language Acquisition*. Shanghai: Foreign Language Teaching Press.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. New York: Wiley.

Lyons, J. (1995). *Linguistic Semantics: An Introduction.* London: Cambridge University Press.

Matthews, P. H. (2000). *Morphology.* Beijing: Foreign Language Teaching and Research Press.

Schmitt, N. & McCarthy, M. (2002). *Vocabulary: Description, Acquisition and Pedagogy.* Shanghai: Shanghai Foreign Language Education Press.

Smith, N. (1999). *Chomsky: Ideas and Ideals*. Cambridge: Cambridge University Press.

Scheaffer, R. L. (1973). Tests for uniform clustering and randomness. *Communications in Statistics-Theory and Methods*. Abingdon-on-Thames: Taylor & Francis.

Appendix:

Statistical Data for the Ten Sets of Text Samples

N	*V(N)*									
	Set_1	*Set_2*	*Set_3*	*Set_4*	*Set_5*	*Set_6*	*Set_7*	*Set_8*	*Set_9*	*Set_10*
15,000	2,073	2,699	2,262	2,189	2,338	2,288	2,842	2,182	1,590	1,837
30,000	3,200	3,422	3,170	3,375	3,056	3,606	3,597	3,030	2,583	3,427
45,000	3,764	4,323	4,148	4,040	4,274	4,374	4,393	3,896	4,007	4,308
60,000	5,003	4,885	4,676	4,720	4,861	5,058	4,961	4,948	5,266	5,362
75,000	5,410	5,541	5,097	5,628	5,343	6,115	5,472	5,474	5,855	6,033
90,000	5,792	6,076	5,800	6,208	5,640	7,025	6,004	5,939	6,322	6,526
105,000	6,137	6,421	6,317	6,580	5,997	7,552	6,393	6,306	6,749	6,841
120,000	6,715	6,788	6,880	7,021	6,686	7,812	6,790	6,636	7,046	7,226
135,000	7,091	7,326	7,316	7,798	7,467	8,168	7,457	7,025	7,596	7,602
150,000	7,481	7,757	7,654	8,140	8,038	8,387	7,688	7,422	8,271	7,874
165,000	8,095	8,227	7,930	8,430	8,829	8,700	7,960	8,103	8,829	8,293
180,000	8,400	8,840	8,410	8,775	9,189	9,188	8,390	8,628	9,247	8,927
195,000	8,864	9,394	8,951	9,265	9,564	9,732	8,893	8,970	9,553	9,376
210,000	9,442	9,718	9,521	9,632	9,914	10,256	9,134	9,234	9,849	9,840
225,000	9,832	10,122	10,163	9,933	10,296	10,670	9,608	9,470	10,060	10,108
240,000	10,270	10,427	10,638	10,168	10,561	10,913	10,051	9,885	10,373	10,325
255,000	10,538	11,031	10,885	10,371	11,012	11,308	10,487	10,187	10,815	10,611
270,000	10,703	11,537	11,242	10,661	11,643	11,592	11,084	10,437	11,075	10,888
285,000	10,936	11,748	11,528	10,942	11,881	11,817	11,489	10,631	11,327	11,233
300,000	11,129	12,174	11,775	11,146	12,049	12,006	11,652	10,970	11,581	11,486
315,000	11,546	12,500	11,933	11,573	12,302	12,362	11,995	11,226	11,928	11,955
330,000	11,860	12,680	12,262	12,029	12,566	12,530	12,206	11,521	12,148	12,187
345,000	12,075	12,931	12,769	12,242	12,827	12,910	12,837	11,832	12,565	12,661
360,000	12,521	13,147	13,201	12,491	13,323	13,200	13,196	12,069	12,959	13,078
375,000	12,744	13,374	13,348	12,830	13,524	13,458	13,706	12,303	13,125	13,300
390,000	13,067	13,608	13,618	13,273	13,596	13,646	13,935	12,613	13,320	13,511
405,000	13,471	14,081	13,880	13,795	13,969	13,851	14,135	12,958	13,632	13,845
420,000	13,716	14,480	14,004	14,168	14,262	14,155	14,312	13,363	13,802	14,010
435,000	13,893	14,690	14,153	14,321	14,438	14,360	14,544	13,593	14,038	14,355
450,000	14,196	15,003	14,377	14,450	14,653	14,662	14,727	13,994	14,263	14,733
465,000	14,400	15,252	14,605	14,812	14,874	14,856	15,032	14,327	14,679	15,075
480,000	14,638	15,373	14,841	15,239	15,226	15,041	15,376	14,553	15,021	15,282
495,000	15,074	15,472	15,159	15,631	15,481	15,248	15,621	15,044	15,258	15,581
510,000	15,368	15,678	15,406	15,762	15,699	15,545	15,810	15,243	15,635	15,741

Glottometrics 47, 2019, 52-65

A Study on Inter-textual Vocabulary Growth Patterns for Maritime Convention English

Hong Su[1]

Abstract. Based on the maritime convention corpus, this paper attempts to address two issues – that is, the characteristics of vocabulary growth of maritime convention English compared with general English from BNC, and the mathematical models of vocabulary growth for the maritime convention corpus. This research tests three existing mathematical models (Brunet's model, Tuldava's model, and Herdan's model). It is revealed that all these three models can be well fitted to the vocabulary growth of the maritime convention corpus, yet Herdan's model is the most suitable one, not only concerning the goodness of fit, but also the brevity of expression.

Keywords: *maritime conventions, vocabulary growth, mathematical models*

1 Introduction

Nowadays, there are three main branches concerning lexical research on the basis of their contribution to the applied linguistic theory of vocabulary. Specifically speaking, they are vocabulary description, vocabulary acquisition, and vocabulary pedagogy. Due to different research methods, vocabulary description can be divided into quantitative lexical study and qualitative lexical study. Although in many occasions, the qualitative method is able to accomplish the research objectives owing to the introspective approach, in many situations, a large number of statistical data is needed to prove the scientific argumentation, and the method of quantitative lexical study is of great significance. Issues like word distribution, text length, wordlist coverage, vocabulary growth patterns are covered by the research on quantitative lexical study. Quantitative lexical study not only provide scientific grounds for vocabulary description, but also a reasonable basis of the applied linguistic theory on vocabulary.

As to quantitative lexical study, vocabulary growth pattern is significant, involving complicated research procedures. Many scholars have made efforts to study the textual vocabulary growth pattern – for instance, what the lexical growth pattern is in a certain text. Many characteristic constants and models have been put forward to describe the growth pattern of a certain text. As for the inter-textual vocabulary growth pattern, there is some research concerning the vocabulary growth curve for specific texts. For example, Fan (2006) tested inter-textual vocabulary growth for general written English with

[1] Correspondence to Hong Su, Weiming Highschool, Chongqing, China. Email: suzihesuhong@126.com

1,000,000 words, and Li (2006) created a mathematical model based on a corpus of maritime engineering English. However, no reliable models have been examined to describe the inter-textual growth pattern for maritime convention English. Thus, this paper will emphasize on how the vocabulary growth curve behaves in maritime convention English.

This paper emphasizes the research of the quantitative lexical characteristics of vocabulary growth pattern and dynamic relationship between text size and vocabulary size for maritime conventions.

Specifically speaking, this paper is intended to answer the following questions: What is the inter-textual vocabulary growth pattern for maritime conventions? – How well can the existing models put forward by quantitative linguists fit inter-textual vocabulary growth for maritime conventions, and which model is the most suitable one describing its vocabulary growth?

2 Methods and Materials

2.1 Maritime Convention Corpus and British National Corpus

There are two corpora in this study: the Maritime Convention Corpus and the British National Corpus. The Maritime Convention Corpus (MCC) includes conventions concerning all aspects of maritime affairs. The British National Corpus (BNC) is one of the largest corpora of general English.

In this study, the first step was the breakdown of the maritime convention corpus into individual texts of an approximately equal length, as the text lengths of conventions in the MCC and the texts from the BNC vary greatly. Since the research of intermediate and advanced EFL teaching tends to use average-length texts, this study divided the two corpora into individual texts of approximately 1,000 tokens. The detailed description of the individual sample texts from the MCC and the BNC is shown in Table 1.

Table 1

Descriptive statistics of sample texts from the Maritime Convention Corpus and the British National Corpus

	Number of texts	Number of tokens	Text length
The MCC	360	372,266	1,000
The BNC	369	369,000	1,000

Table 1 displays the descriptive statistics of the Maritime Convention Corpus and the British National Corpus. The total number of tokens of the MCC and the sample corpus from BNC is 372,266 and 369,000, respectively. The MCC was divided into

360 individual texts of roughly 1,000-word tokens. The BNC was divided into 369 texts. In the first process, some crucial issues, namely, the lemmatization and the

tokenization were addressed by the Perl programs.

In the second step, SPSS was utilized to calculate the word frequencies and type-token ratio and to analyse the difference in the vocabulary growth patterns between the MCC and general English from the BNC.

The final step was accomplished with the aid of SPSS; the programme drew the predicted vocabulary growth model curve and figured out the most suitable mathematical model. SPSS was used to draw the vocabulary growth curve for the MCC and the predicted vocabulary growth curves of three mathematical models (Brunet's model, Tuldava's model, and Herdan's model). The determination coefficient R^2 and residual errors of three mathematical models were also calculated and drawn by SPSS.

2.2 Three Mathematical Models

2.2.1 Brunet's Model

In 1988, Brunet (Brunet, 1988) made a great contribution to quantitative lexical studies by proposing a complicated logarithmic relation between *N* and *V(N)*. The expression is as follows:

$$V(N) = (\log_N W)^{\frac{1}{\alpha}} \tag{1}$$

In this function, parameter *W* (Brunet's constant) has no realistic interpretation, and parameter α is constantly fixed as 0.17, a heuristic value that can lead to the wanted result of the relation between *N* and *V(N)*. To make it easier, let

$$\alpha = \ln \mathrm{W}^{\frac{1}{\alpha}}$$

and

$$\beta = (\ln N)^{\frac{1}{\alpha}};$$

as a consequence, the expression below is transformed into:

$$V(N) = \alpha (\ln N)^{\beta}, \tag{2}$$

which is another format of Brunet's model.

2.2.2 Tuldava's Model

Based on the principle of "the restriction of variety", Tuldava (1996) founded a mathematical model to research the relationship between the vocabulary size *V(N)* and the sample size *N*. The variety of lexicon is defined as the relation between the vocabulary size and the sample size in the form of *V(N)/N* (type-token ratio) or *N/V(N)* (mean type

frequency). Hence, in Tuldava's model the relation between *V(N)* and *N* is estimated by the power function:

$$\frac{V(N)}{N} = \alpha N^{\beta} \tag{3}$$

To make the function clearer and simpler, Tuldava applied logarithms to both variables of *V(N)* and *N* and established a vocabulary growth model as follows:

$$V(N) = N\mathrm{e}^{-\alpha(\ln N)^{\beta}} \tag{4}$$

e: natural base, 2.71828

Parameters α and β have no probabilistic interpretation; they are coefficients of variety that are considered to be related to the probabilistic process of selecting "new" (unused) and "old" (already utilized in the text) words at each stage of text processing.

2.2.3 Herdan's Model

After observing that the growth curve of the vocabulary appears as roughly as a straight line in the double logarithmic plane, Herdan (1964) established a mathematical model. In Herdan's research, a linear relation between *log V(N)* and *log N* would be reasonable as follows:

$$\log V(N) = \log \alpha + \beta \log N = \log(\alpha N^{\beta}) \tag{5}$$

Therefore,

$$V(N) = \alpha N^{\beta} \tag{6}$$

This power model for the growth of vocabulary is known as Herdan's model, where α and β are observed parameters without a sensible interpretation.

3 Results and Discussion

3.1 Vocabulary Growth Models for the Maritime Convention Corpus

In order to study the inter-textual vocabulary growth pattern for the MCC, the vocabulary growth curve is studied and drawn. In view of the large volume of the MCC, the vocabulary growth curve was measured at roughly equal 70,000 running word intervals.

Figure 1 shows the relation of vocabulary size *V(N)* and the sample size *N*. The *y*-

axis represents the number of word tokens (*N*) of the MCC. The *x*-axis stands for the number of word types (*V(N)*) of the MCC. Generally speaking, the vocabulary growth curve of *V(N)* exhibits a nonlinear function of the sample size *N*, and *V(N)* increases with the increase of *N*. In terms of the slope, the vocabulary growth curve ascends sharply in the beginning and the rates of increase gradually slow down.

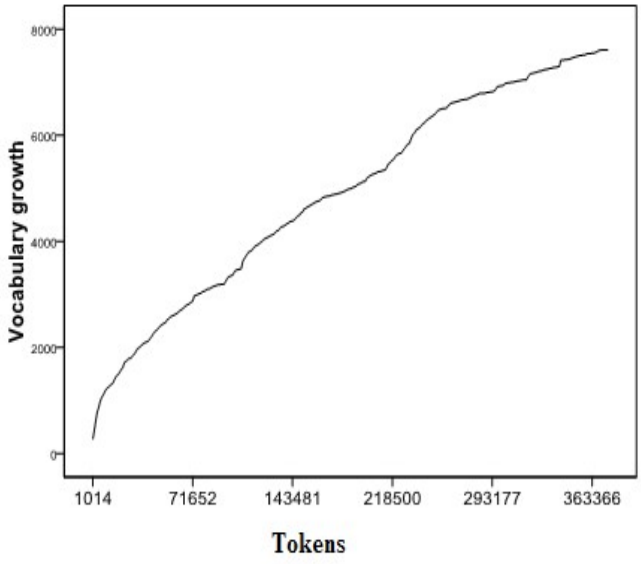

Figure 1. The vocabulary growth curve *V(N)* against the sample size *N* (Tokens) for the MCC

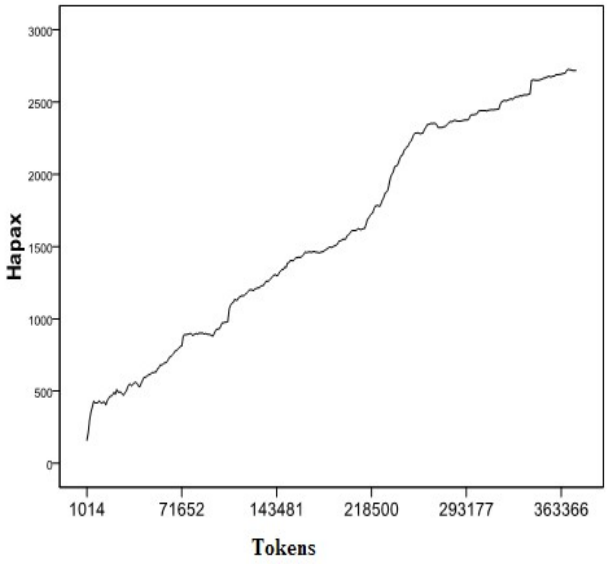

Figure 2. The growth curve of the number of tokens against the hapax legomena for the MCC

In order to study the vocabulary growth curve further, the hapax legomena were investigated. Figure 2 was drawn to study the relation between word tokens and hapax legomena. As the number of cumulative tokens increases, the hapax legomena's rates

of increase gradually slow down. In the MCC, the number of hapax legomena in the first 1,000-token text is 160, and the total number of hapax legomena is 2,719.

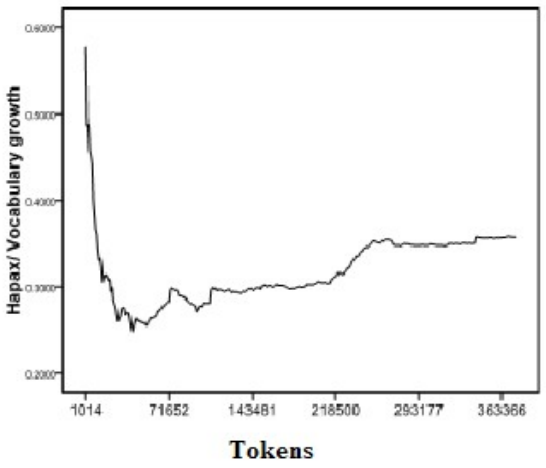

Figure 3. The curve of hapax/vocabulary (the ratio between one-time words and vocabulary sizes) and the number of tokens

In Figure 3, the *y*-axis stands for the number of tokens, while the *x*-axis represents the ratio between hapax legomena and vocabulary sizes. The ratios drop sharply at the beginning and spring a little to about 0.3. The ratio of the first text is 0.5955, but after the first 20 texts, the figures fall below 0.3. At the end, the ratios rise to about 0.4.

Table 2 lists the detailed data of text sizes, vocabulary growth, hapax legomena, and the hapax/vocabulary growth rate of the first twenty texts. The value of ratio in the first text is 0.5955, which means the one-time words occupy more than 59% of the tokens in the first 1000-token text. It decreases sharply to the seventh sample text from 0.5955 to 0.4355, which is just the opposite to the increasing trend of *V(N)* and *N* in Figure 3. Then, from the eighth text on, the ratio drops from 0.44 to 0.39.

Table 2

The statistical data of the first twenty texts, vocabulary sizes, and hapaxes

Text size	Vocabulary size	Hapax	Hap/Voc ratio
1,054	309	184	0.5955
2,092	440	233	0.5295
3,142	582	302	0.5189
4,192	621	287	0.4622
5,243	730	335	0.4589
6,298	820	360	0.439
7,333	923	402	0.4355
8,403	979	432	0.4413

9,525	1,123	496	0.4417
10,588	1,221	530	0.4341
11,623	1,260	533	0.423
12,672	1,291	541	0.4191
13,737	1,331	555	0.417
14,896	1,411	583	0.4132
15,937	1,433	565	0.3943
16,993	1,461	581	0.3977
18,044	1,492	593	0.3975
19,101	1,534	602	0.3924
20,167	1,583	615	0.3885
21,240	1,661	656	0.3949

3.2 Test of the Existing Vocabulary Growth Power Models

Mathematical models are extrapolated to predict the vocabulary growth pattern of a particular text. Power mathematical models, such as Herdan's model, Tuldava's model, and Brunet's model, are popular in the study of vocabulary growth patterns, and their expressions are much simpler than non-power models. Therefore, this section tests the goodness of the fitting and the residual errors for the three power models.

3.2.1 Test of Brunet's model

Figure 4 is the goodness of the fitting for the MCC according to Brunet's model as expressed in (1). The parameter β is the exponential of Brunet's mathematical model. In Brunet's model, the estimated values for α and β are 0.003 and 0.706, respectively.

In order to analyse the fit of these models, the determination coefficient R^2 is calculated. R^2 is a measure to assess the adequacy of a regression model; the value of R^2 varies from 0 to 1, which reveals the direct relation of the model's fit. The bigger the R^2 is, the better a model can reveal the empirical vocabulary increase. The value of R^2 of Brunet's model is 0.993.

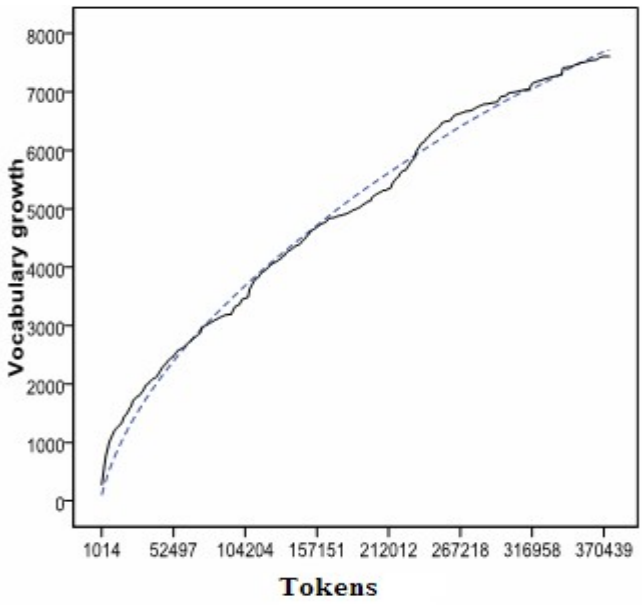

Figure 4. The growth curve of the observed values and the predicted values of Brunet's model. The curve represents the observed values; the dotted line is the predicted values

Computing the residual errors is another efficient approach to test the model fitting. Figure 5 exhibits the residual errors for the MCC of the Brunet's model. The dotted line is the line of the expected value 0, which means there is no difference between the predicted values and the observed values. Residual 0 stands for the residual errors from the test of Brunet's model. The residual errors are calculated as the sum of the observed values and the predicted values. If the values are positive, it means that the observed values are greater than the predicted values. If the values are negative, the observed values are smaller than the predicted values. In Figure 5, the residual errors have a wide dispersion, from about -300 to 500, and there are positive and negative values around the expected value 0.

After the analysis of the residual errors and the determination coefficient R^2, the conclusion is drawn that the observed values differ largely from the calculated value.

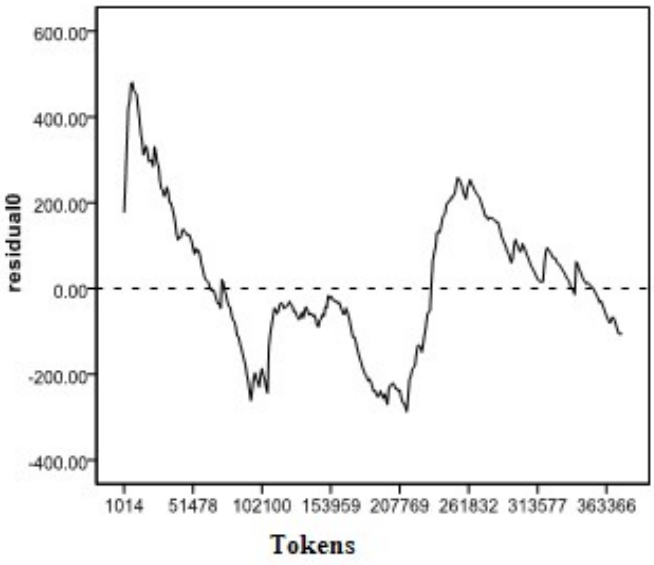

Figure 5. Residual errors of Brunet's model as a function against the sample size *N*. The curve represents the residual errors; the horizontal dotted line is the expected value 0.

3.2.2 Test of Tuldava's Model

Tuldava deduces a mathematical model of power function to describe the relation between *V(N)* and *N*. In the test of vocabulary growth pattern of Tuldava's model for the MCC, the values of estimated parameters α and β are 0.015 and 1.378, respectively. The determination coefficient R^2 is 0.995.

Figure 6 is the vocabulary growth curve from Tuldava's model for the MCC. The dotted line is the values predicted by Tuldava's model, and the curve stands for the observed values. In the first 1,000-token text, the predicted vocabulary size of Tuldava's model is 194 word types, and the observed vocabulary size is 277 word types. The predicted value is by 82 word types lower than the observed value. Generally speaking, the curve fluctuates around the dotted line.

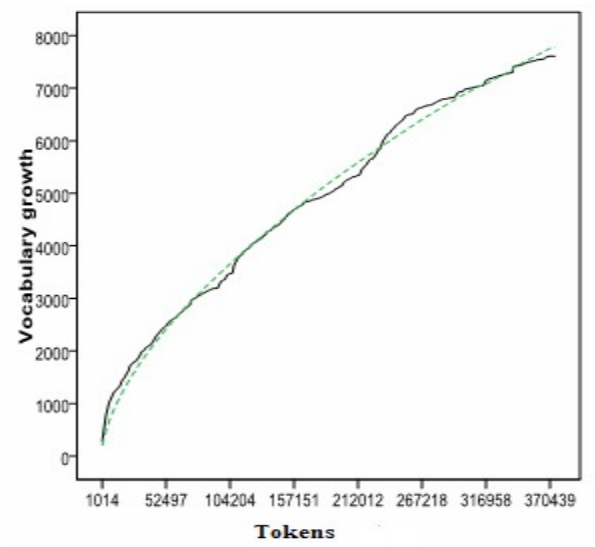

Figure 6. The growth curve of the observed values and the predicted values of Tuldava's model. The curve represents the observed values; the dotted line is the predicted values

Figure 7 is the residual errors between the predicted values and the observed values of Tuldava's model. Residual 1 represents the residual errors from the test of Tuldava's model. The dotted line is the line of the expected value 0. The curve is the dispersion of residual errors for the MCC of Tuldava's model. There are both positive values and negative values at different numbers of tokens, and these residual errors scatter in a range from -300 to 400, or so.

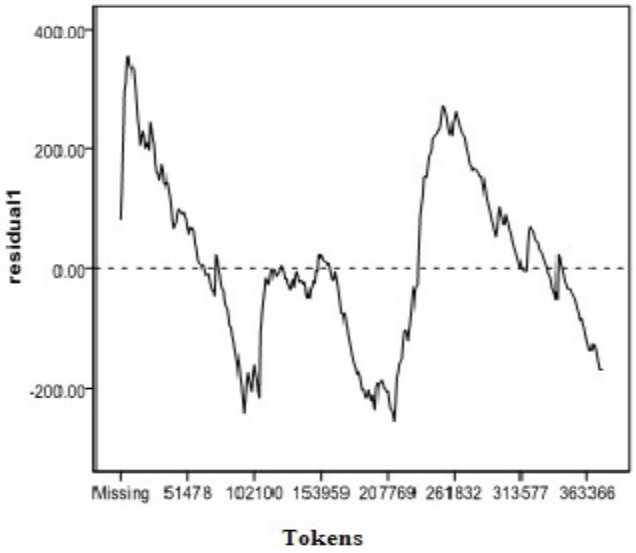

Figure 7. Residual errors of Tuldava's model as a function against the sample size *N*. The curve represents the residual errors; the horizontal dotted line is the expected value 0.

3.2.3 Test of Herdan's Model

Herdan's mathematical model of the power function is of great significance in the study of vocabulary growth patterns. His power expression is as (6) shows.

In the test of goodness of fit for the MCC of Herdan's model, the parameter α is estimated as 3.864 and the parameter β is 0.594. The value of R^2 is 0.995.

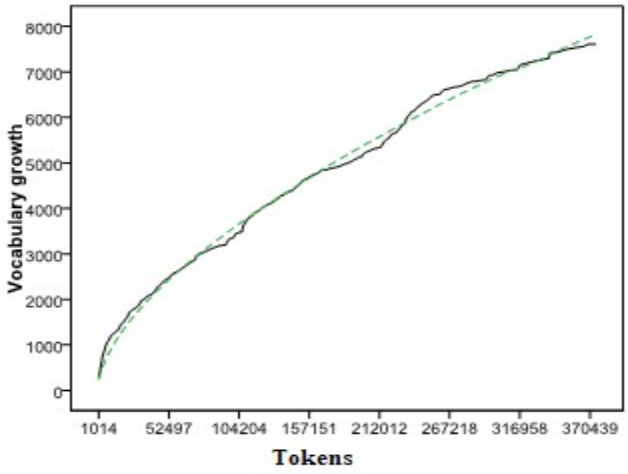

Figure 8. The growth curve of the observed values and the predicted values of Herdan's model. The curve represents the observed values; the dotted line is the predicted values

Figure 8 shows the relation between the predicted values and the observed values of the MCC. The dotted line represents the predicted values of Herdan's model. The curve is the observed vocabulary growth of the MCC. Vocabulary size of the first 1,000-token text is 277, while the predicted value of Herdan's model is 233.9. Generally speaking, the curve rises as the dotted line increases, with some deviation.

Figure 9 describes the residual errors between the predicted values and the observed values of Herdan's model. Residual 2 represents the residual errors from the test of Herdan's model. The dotted line represents the ideal line where the predicted value is totally equal to the observed vocabulary size. As is shown in the graph of residual errors, there is a fluctuation between -250 and 300. In the first text, the residual error is 43.2 word tokens, which is a fairly small distance. After it reaches its maximum of 305 word tokens, the distance between the predicted values from Herdan's model and the observed values are smaller and smaller. Then the observed values are lower than the predicted values, as the result is negative. Generally speaking, the residual errors fluctuate.

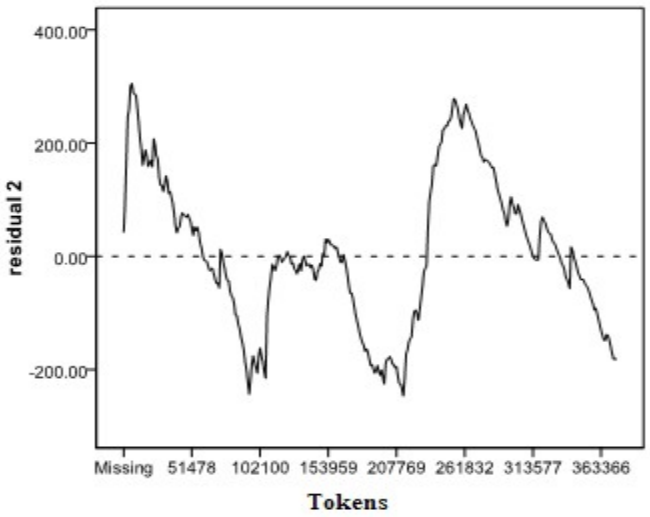

Figure 9. Residual errors of Herdan's model as a function against the sample size *N*. The curve represents the residual errors; the horizontal dotted line is the expected value 0.

3.3 Comparison of the Results of the Three Mathematical Models

This section summarizes the comparisons among three power mathematical models (Brunet's model, Tuldava's model, and Herdan's model) and endeavours to find out the most suitable vocabulary model for the MCC. The parameters in these models, the residual errors, and statistical data are listed in the following table.

Table 3

Parameters for three power mathematical models

Models	Expressions	Parameter α	Parameter β	R^2
Brunet's model	$V(N)=\alpha(\ln N)^{\beta}$	0.003	0.706	0.993
Tuldava's model	$V(N)=Ne^{-\alpha(\ln N)^{\beta}}$	0.015	1.378	0.995
Herdan's model	$V(N)=\alpha N^{\beta}$	3.864	0.594	0.995

Table 3 lists the expressions of three mathematical models (Brunet's model, Tuldava's model, and Herdan's model), and the value of parameters α and β for each expression. These parameters were calculated by the software SPSS to obtain the vocabulary size for each model. Table 4 displays the statistical data concerning the residual distribution. Compared with the mean value of residual errors for other mathematical models, Herdan's model's manifests the smallest one (7.87). What is more, the range of the difference has also been tested, and the range of Herdan's model is 552.33, which

is also the smallest one. Considering R^2, Brunet's model is 0.993. The R^2 for Tuldava's model and Herdan's model is 0.995, which is closer to the ideal value 1. Thus, the R^2 for Tuldava's model and Herdan's model fit better.

Table 4

The mean and range of residual errors of the three models

	Mean residual error	Range of residual errors
Brunet's model	16.98	767.11
Tuldava's model	12.08	610.45
Herdan's model	7.87	552.33

The observed vocabulary growths (*V(N)*) and the predicted vocabulary growths in the first 10 texts are shown in Table 5. The first column represents the text sizes of the first 10 texts. The second column is the observed vocabulary size. The last three columns stand for the predicted values for Brunet's model, Tuldava's model, and Herdan's model, respectively. In the first 1,000-token text, the observed vocabulary size is 277. The predicted vocabulary size for Brunet's model, Tuldava's model, and Herdan's model is 98.6, 194.1, and 233.9. The value of Herdan's model 233.9 is the closest one to the observed values among these three values. What is more, in all the ten texts, the predicted values of Herdan's model are closest to the observed values.

To sum up, Herdan's model is the most suitable model for the MCC compared with other existing mathematical models, not only in terms of the determination coefficient R^2 and the residual errors of the predicted values of models and the observed values of the MCC, but also due to the simplicity of its expression.

Table 5

The numbers of tokens and vocabulary sizes of the first ten texts for the MCC, and the predicted values of three models

Number of tokens	Vocabulary growth	Brunet	Tuldava	Herdan
1,014	277	98.6	194.1	233.9
2,024	444	193.1	307.3	352.6
3,023	617	277.7	399.7	447.4
4,035	778	356.6	482.2	531.0
5,048	867	430.4	557.3	606.6
6,053	977	499.5	626.2	675.6
7,065	1,046	565.7	691.3	740.5
8,078	1,091	628.9	753.0	801.8
9,083	1,146	689.2	811.3	859.6
10,088	1,200	747.3	867.1	914.9

4 Conclusion

This paper studies the inter-textual vocabulary growth pattern for the MCC of 372,266 tokens. The major findings are as follows:

The power expressions like Herdan's model, Tuldava's model, and Brunet's model are easy to compute and fit the MCC rather well. In the test of goodness of fit, the values of R^2 for Herdan's model, Tuldava's model, and Brunet's model are 0.995, 0.995, and 0.993, respectively. In addition, the mean values of their residual errors are respectively 7.08, 12.08, and 16.98, respectively. By comparing all the parameters and determination coefficient, Herdan's model has the smallest range and a higher R^2. What is more, the expression of Herdan's model is simple. Therefore, Herdan's model is considered to be the most suitable model for the Maritime Convention Corpus.

References

Brunet E. (1988). Une mesure de la distance intertextuelle: la connexion lexicale. *Le nombre et le texte. Revue informatique et statistique dans les sciences humaines* 24(1–4), 82–116.

Herdan, G. (1964). *Quantitative Linguistics*. London: Buttersworths.

Fan, F. (2006). A corpus-based study on inter-textual vocabulary growth. *Journal of Quantitative Linguistics* 13, 111–117.

Li, J. (2006). *Inter-textual vocabulary growth pattern for maritime English*. M. A Thesis. Dalian Maritime University.

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics* 3, 38–50.

Glottometrics 47. 2019, 66-82

A Corpus-Based Comparative Study of Lexis in Hong Kong and Native British Spoken English

Yaobin Yan[1]

Abstract. Based on the corpus of Hong Kong English and the one of native British English, the present study aims at characterizing the lexis in Hong Kong learners' spoken English. First, the study investigates the quantitative features of the lexis in terms of vocabulary size, mean word length, lexical density, and lexical coverage, and then moves on to the qualitative interpretation of the features, particularly from the perspectives of high frequency words, hapax legomena, inserts, informal words, contractions, and abbreviations.

Keywords: *spoken English vocabulary, vocabulary size, word length, lexical density, lexical coverage, high frequency word, hapax legomena*

1 Introduction

Vocabulary has been long seen as the essential part of language learning. The majority of teachers and learners are convinced that vocabulary knowledge constitutes quite an important part of competence in language teaching and learning. Wilkins (1996) states that "while without grammar little can be conveyed, without vocabulary nothing can be conveyed". Krashen and Terrell (1983) state that "vocabulary is of prime concern in foreign language settings because it plays a prominent role in classroom success". A great number of empirical studies have indicated that lexis has not got enough consideration from syllabus designers. Not until the second half of the 1980s had syllabus theorists begun to move attention to lexis. McCarthy (1984) states that coursebooks are suffering from the dominance of syntax, notions, and functions, at the expense of lexical development. Modern technologies have enabled us to set up huge computer corpora, which have permitted the compilation of much better word frequency lists, allowed more confident decisions on which frequency vocabulary to include, and also provided much broader data for concordance aims. Over the past two decades, spoken language has gained more importance in language teaching, with the emphasis on language for communication. Simpson (1998) estimates that a native English speaker may be exposed to the somewhat fantastic figure of a million words per day of spoken and written language combined. To meet the needs for communication, many different types of oral English exams have been carried out, and there are also mushrooming amounts of pedagogical publications specially designed for spoken English.

[1] Correspondence to Yaobin Yan, Shenyang City University, Shenyang, China. Email: yanyaobin1983@163.com

With a corpus-based approach, this study addresses the following questions: (1) How can we characterize the lexis in Hong Kong learners' spoken English both quantitatively and qualitatively? – (2) The results of previous studies have indicated that there exists a significant difference in the use of vocabulary between native spoken and written English. Can this type of difference be found in Hong Kong learners' English? – If so, is the difference reflected in the same manner? –

As to the typical previous studies in China and some European countries on the lexical characteristics of non-native learners' spoken English, Wen (2004) finds that Chinese college English majors tend to overuse and underuse some high-frequency adverbs. Wen (2003), too, finds that English compositions of advanced Chinese English learners display oral features, and they will decrease along with their second language proficiency development. Based on previous findings, Wen (2006) finds that certain patterns of changes exist in spoken vocabulary development by English majors through four years.

Zhao's (2008) study shows the following four main results. Firstly, the differences between Chinese EFL learners' spoken and written English are rather slight in terms of vocabulary size, lexical density, mean word length, and vocabulary increase ratio. Secondly, in Chinese EFL learners' spoken English, 800 words can cover 95% of the total tokens, while in native spoken English 3,000 words can reach the coverage rate of 95%. Thirdly, there are considerable differences in the use of inserts, contractions, and informal words between Chinese EFL learners' spoken English and native spoken English. Fourthly, there are a number of overgeneralized words in Chinese EFL learners' spoken English, and most of them are deviant structures of past tense. In *Automatic Profiling of Learner Texts*, Granger and Rayson (1998) find out that non-native speakers tend to overuse some linguistic features, such as most indefinite pronouns, most indefinite determiners, first and second personal pronouns, some auxiliaries and infinitives, and they also underuse some linguistic features, such as most prepositions, most subordinators, *-ly* adverbs, *-ing* nouns, and *-ed* participles. Petch-Tyson (1998) finds that advanced learners of four European nationalities produce from two to four times the number of spoken linguistic features than equivalent Americans do for an equal writing task. Cobb (2003) finds that Quebec learners' writings display the same common features as European learners' writings.

The insights gained from the studies above-mentioned are invaluable. However, little has been done on the study of lexical characteristics of Hong Kong learners' spoken English, especially in many different respects and from quantitative and qualitative perspectives, and this indicates the necessity of the present lexical study.

2 Methods and Materials

2.1 Research Methodology

The corpus-based approach is the guiding methodology in the present study. The corpus-based quantitative and qualitative analyses are used in this study. The corpora used in this research are the British National Corpus (BNC) and the Hong Kong Component of the International Corpus of English (ICEHK). The programmes used are Microsoft Visual Foxpro, which are used to decode the codes of the BNC and ICEHK corpora, break the corpora into words, and do the lemmatization, and other related processing. Based on the corpora, the contrastive interlanguage analysis approach is adopted as the concrete method for the present study.

2.2 Materials

BNC is a representation of 100 million words of contemporary spoken and written English, compiled between 1990 and 1994 and released in May 1995. BNC is made up of approximately 90% written data and 10% orthographically transcribed spoken data.

The International Corpus of English (ICE) contains 20 parallel sub corpora, each consisting of one million words of English used by adults older than 18 who have received formal education through the medium of English to at least the completion of secondary school, in countries like UK, USA, and Canada, as well as in India and Singapore, or regions like the Caribbean. The ICEHK corpus is the Hong Kong component of the ICE corpus. The ICEHK project was initiated at Hong Kong Polytechnic University in the early 1990s by Philip Bolt. The ICEHK corpus follows the common design of the ICE corpora, including 300 spoken texts and 200 written texts.

For the purpose of the present study, 170 spoken texts are randomly selected out of the corpus, containing 495,019 tokens, while the 200 written texts are composed of 440,332 tokens. The BNC spoken corpus contains around 10 million words. For the purpose of a reference corpus, 260 2000-word samples are selected, totalling 535,950 tokens. The BNC written corpus contains around 90 million words. For the purpose of a reference corpus, 280 2000-word samples are selected, totalling 500,021 tokens.

2.3 Tools and Procedures

To obtain the data necessary for the present study, the most popular statistical package for social science – SPSS and some Foxpro programmes – are applied. The procedures for data collection include decoding, tokenization, and lemmatization. The codes in the corpora are removed in the study, since it is a lexical study of spoken English. With the aid of the programmes in Microsoft Visual Foxpro, the samples of written and spoken texts are transformed into four lists of isolated words for further processing. Following tokenization, another process, lemmatization, is carried out. After lemmatization, the

frequencies of the word types can be computed so that linguistic information for further studies can be acquired efficiently.

3 Results and Discussion

This part begins with the quantitative analyses of the lexis in the spoken English of Hong Kong learners and the lexical features are further analyzed qualitatively.

3.1 Quantitative Analysis

3.1.1 Vocabulary Size

After the processes of tokenization and lemmatization, all the word types were retrieved from the four corpora in this study. Table 1 shows the types, tokens, and vocabulary size per ten million tokens for each corpus.

Table 1

Values for types, tokens, and vocabulary sizes in each corpus[1]

Corpus	Type	Token	Vocabulary size per tenth million tokens
BNCS	7,620	535,950	1,422
BNCW	18,394	500,021	3,679
HKS	8,804	495,019	1,779
HKW	15,157	440,032	3,442

From Table 1, the order of the vocabulary size per ten million tokens of the four corpora is as follows: BNCW > HKW > HKS > BNCS. From the results above three patterns can be found. Firstly, whether in BNC or in ICEHK, the vocabulary size of the written corpus is larger than that of the spoken corpus, which indicates that the spoken vocabulary is smaller than the written vocabulary. Secondly, compared with the difference between BNCS and BNCW in vocabulary size per ten million tokens, the difference between HKS and HKW is comparatively slight. To confirm this observation, the chi-square test is applied, and the results are given in Table 2.

As shown in Table 2, the values of χ^2 are 5382.27 and 2584.01 respectively for the comparisons between the two modes of BNC and those of ICEHK. Thirdly, the vocabulary size per ten million tokens of HKS is larger than the one of BNCS.

[1] BNCS refers to BNC spoken English, BNCW refers to BNC written English, HKS refers to Hong Kong spoken English, and HKW Hong Kong written English.

Table 2
Results of Chi-square test on vocabulary size

Comparison Group	Value for χ^2 test	P
BNCS vs. BNCW	5382.27	0.00
HKS vs. HKW	2584.01	0.00

3.1.2 Mean Word Length

Table 3 shows the values of mean word length in each corpus involved in comparison.

Table 3
Values of mean word length in each corpus (letters)

BNCS	BNCW	HKS	HKW
6.11	6.85	6.92	6.95

Among the four corpora, the mean word length of HKW is the greatest, and the mean word length of BNCS is the smallest, with the mean word lengths of HKS and BNCW standing in between. Through comparison between the corpora, the following three patterns can be perceived. Firstly, whether in BNC or ICEHK, the mean word length of the written corpus is greater than the one of the spoken corpus. Secondly, compared with the difference between BNCS and BNCW, the difference between HKS and HKW is comparatively slight. To confirm this observation, the one-way ANOVA test was conducted, on the basis of this raw data, and the results are given in Table 4.

As Table 4 shows, the *p*-value for the difference in mean word length between BNCS and BNCW is 0.000, which is much lower than 0.05, the pre-set significant value. So, we can claim that BNCS is significantly different from BNCW in terms of the mean word length. However, the examination of the difference between HKS and HKW indicated that the *p*-value (0.962) is much greater than 0.05, which indicates that although there is a difference between HKS and HKW, yet the difference is not statistically significant at the level of 5%.

Thirdly, the words in HKS are longer than those of BNCS. The second and third patterns suggest that Hong Kong spoken English displays characteristics of the written style, the difference of which has been confirmed to be significant at the significance level of 5%, as shown in Table 4.

A non-native speaker cannot easily achieve the proficiency of a native speaker during the process of second language acquisition (SLA). Consequently, it can be commonly believed that the mean word length of a non-native speakers' corpus is usually shorter than the one of a native speakers' corpus. However, the result here is contrary to our common sense. To this unusual phenomenon, we hypothesize that Hong Kong

learners of English use exclusively some long words. In order to test our hypothesis, we examined the BNCS-only words and HKS-only words. Table 5 shows the distribution of the length of words occurring exclusively in BNCS and HKS.

Table 4

Results of one-way ANOVA test on the mean word length

(I)CORPUS	(J)CORPUS	Mean Difference(I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
bncs	bncw	-.75*	0.049	0.000	-0.87	-0.62
	hks	-.82*	0.049	0.000	-0.94	-0.69
	hkw	-.84*	0.049	0.000	-0.97	-0.71
bncw	bncs	.75*	0.049	0.000	0.62	0.87
	hks	-.07*	0.049	0.479	-0.2	0.06
	hkw	-.09*	0.049	0.221	-0.22	0.03
hks	bncs	.82*	0.049	0.000	0.69	0.94
	bncw	.07*	0.049	0.479	-0.06	0.2
	hkw	-.02*	0.049	0.962	-0.15	0.1
hkw	bncs	.84*	0.049	0.000	0.71	0.97
	bncw	.09*	0.049	0.221	-0.03	0.22
	hks	.02*	0.049	0.962	-0.1	0.15

Table 5

Distribution of length of words only in BNCS and HKS

	3-letter words	4-letter words	5-letter words	6-letter words	7-letter words	8-letter words	9-letter words	10-letters or more
HKS-only	844	794	3,762	2,498	2,644	2,929	2,334	5,310
BNCS-only	3,049	1,795	2,267	1,946	1,236	902	590	700

The longest word in BNCS-only is of 17 letters (“multi-millionaire”), while in HKS-only, the longest word (“technology-intensive”) is of 20 letters. Table 5 shows that for words of 4 or fewer than 4 letters, the total frequencies in HKS-only are lower than their counterparts in BNCS-only. For words of 5 or more than 5 letters, the total frequencies in HKS-only are higher than their counterparts in BNCS-only. The most obvious difference lies in the words of 10 or more than 10 letters, which occur 5,310 times in HKS-only and 700 times in BNCS-only. The data above show that Hong Kong

English learners tend to overuse long words of 10 or more than 10 letters, which contributes to the phenomenon that the mean word length of HKS is longer than the one of BNCS.

3.1.3 Lexical Density

As proposed by Nation (1990), TTR is an important index of lexical complexity and text difficulty. A larger TTR means that the vocabulary in the text is more complex, which implies a more formal style.

Table 6
TTRs of BNCS, BNCW, HKS and HKW

BNCS	BNCW	HKS	HKW
0.0142	0.0368	0.0178	0.0344

From the results in Table 6, three patterns can be observed. Firstly, whether in BNC or ICEHK, the value for the lexical density of the written corpus is higher than that of the spoken corpus. Secondly, compared with the difference between BNCS and BNCW in lexical density, the difference between HKS and HKW is comparatively slight. To confirm it, the proportional test is applied, and the results are displayed in Table 7.

Table 7
Results of the proportional test on TTR

HKS vs. HKW	$Z=50.8332$ ($p < 0.01$)
BNCS vs. BNCW	$Z=73.3638$ ($p < 0.01$)

It can be inferred that in terms of lexical density, the difference between HKS and HKW is slighter compared to the one between BNCS and BNCW. Thirdly, the lexical density of HKS is larger than the one of BNCS. The second and third patterns above indicate that Hong Kong learners' spoken English has features of a written style to some degree.

3.1.4 Lexical Coverage

The results of lexical coverage of the corpora BNCS, BNCW, HKS, and HKW are listed in Table 8.

From the observation of the figures shown in Table 8, the following patterns can be seen. Firstly, whether in BNC or ICEHK, the values for the lexical coverage in the spoken corpus are greater than their counterparts of the same number of types in the

written corpus. Secondly, through the comparison between the same mode, the values for the lexical coverage in HKS are smaller than their counterparts of the same number of types in BNCS, while it is not the same case in HKW and BNCW. This indicates the vocabulary size in HKS is larger than that in BNCS. Furthermore, proportional test was conducted to test the difference between the two modes of both BNC and ICEHK on lexical coverage.

Table 8
Lexical coverage in BNCS, BNCW, HKS and HKW

Top words	Coverage in BNCS	Coverage in BNCW	Coverage in HKS	Coverage in HKW
25	48.27%	37.37%	39.58%	36.16%
50	62.26%	44.13%	50.73%	42.45%
100	73.53%	51.09%	61.94%	49.77%
200	81.81%	58.04%	71.48%	57.22%
500	89.87%	68.85%	82.14%	68.96%
1,000	94.28%	77.94%	89.16%	78.64%
1,500	96.19%	82.99%	92.64%	83.96%
2,000	97.27%	86.28%	94.76%	87.42%
2,500	97.97%	88.63%	96.14%	89.82%
3,000	98.47%	90.38%	97.10%	91.61%

Table 9
Results of proportional tests on lexical coverage

Top words	Z Score between BNCS and BNCW ($p < 0.01$)	Z Score between HKS and HKW ($p < 0.01$)
25	111.99	33.99
50	184.91	80.02
100	235.99	118.38
200	264.67	144.16
500	265.79	148.45
1,000	242.44	139.27
1,500	221.95	131.61
2,000	205.60	125.82
2,500	192.16	120.84
3,000	181.49	116.31

As the figures in the same line in Table 9 indicate, the difference between HKS and HKW is less obvious than the one between BNCS and BNCW in lexical coverage of the same number of word types.

3.1.5 Summary of the Quantitative Analysis

We have carried out some quantitative studies on the lexical characteristics of Hong Kong learners' spoken English. The following two patterns have been found. Firstly, the difference between Hong Kong learners' spoken and written English is comparatively slight compared to the one between native spoken and written English. Secondly, the values for vocabulary size per ten million tokens, mean word length, and TTR of HKS are larger than their counterparts of BNCS, which indicates that there are some written style features in Hong Kong learners' spoken English. For further scientific exploration of the lexical characteristics of Hong Kong learners' spoken English, some further qualitative studies will be conducted in the following sections. We will mainly focus on high frequency words, hapax legomena, and the spoken-only words.

3.2 Qualitative Analysis

3.2.1 High Frequency Words

The results of the top 50 words in BNCS, BNCW, HKS, and HKW are listed in Table 3.10.

Table 10
Top 50 words in BNCS, BNCW, HKS, and HKW

BNCS	be(35,143) I(22,627) you(17,095) it(16,738) the(14,278) not(13,190) do(12,089) that(12,047) and(11,469) a(10,665) have(10,510) to(9,784) yeah(7,183) get(6,964) he(6,127) in(5,913) they(5,912) oh(5,858) no(5,514) of(5,508) go(5,307) what(4,861) well(4,731) we(4,618) say(4,595) she(4,391) there(4,315) know(4,289) on(4,140) one(4,024) but(3,572) yes(3,552) this(3,447) think(3,395) will(3,379) so(3,282) for(3,007) er(2,643) like(2,608) all(2,536) would(2,436) just(2,334) up(2,294) then(2,273) with(2,251) right(2,219) if(2,209) me(2,203) them(2,142) see(2,063)
BNCW	the(35,370) be (20,333) of (17,058) and(14,695) to(13,826) a(13,781) in(10,932) have(6,462) that(5,593) for(4,880) it(4,734) on(4,059) with(3,758) I(3,689) not(3,554) he(3,510) this(3,285) by(3,050) as(2,980) at(2,761) you(2,525) his(2,453) but(2,357) do(2,296) from(2,291) or(2,119) they(2,077) which(2,026) she(1,746) her(1,635) say(1,537) there(1,530) their(1,477) one(1,446) will(1,352) we(1,259) all(1,216) would(1,213) can(1,201) make(1,149) when(1,131) who(1,088) more(1,074) if(1,058) use(1,024) some(1,001) year (970) other(964) go(957) up(945)
HKS	the(25,290) be(18,396) I(15,298) to(13,978) and(13,312) you(12,296)

	a(10,230) of(9,470) in(8,925) that(8,815) it(8,677) have(6,549) yeah(6,009) so(4,674) but(4,328) do(4,294) for(4,152) we(4,121) they(4,089) this(4,027) think(3,377) not(3,366) know(3,001) what(2,896) one(2,847) on(2,747) he(2,649) yes(2,641) or(2,625) like(2,604) go(2,529) there(2,438) no(2,403) will(2,273) very(2,244) because(2,205) can(2,116) as(2,095) then(2,036) with(2,033) oh(2,015) say(1,970) if(1,961) at(1,924) just(1,899) okay(1,828) about(1,817) my(1,773) from(1,666) all(1,645)
HKW	the(30,024) be(17,884) of(14,509) to(13,466) and(11,926) a(11,110) in(9,694) for(5,163) have(5,038) that(4,337) I(4,330) it(3,820) as(3,326) on(3,122) with(2,997) you(2,801) by(2,585) this(2,507) from(2,178) not(2,112) or(2,006) at(1,907) will(1,741) say(1,683) he(1,653) can(1,641) but(1,530) they(1,435) we(1,364) do(1,309) my(1,285) which(1,279) one(1,248) their(1,231) more(1,197) your(1,097) there(1,090) if(1,070) conjurer(1,061) his(1,060) so(1,010) make(986) she(985) her(974) year(954) all(950) when(940) use(937) time(916) would(903)

Note: the numbers in the parentheses indicate respective frequencies of the preceding words.

As the results in Table 10 show, the most noticeable difference between the high frequency words of BNCS, BNCW, HKS, and HKW is the rank-1 word. HKS and HKW have the same one — the determiner *the*. Nouns appear much less common in native spoken English than in native written English. The difference in the rank-1 words indicates that Hong Kong learners' spoken English show written style features to some extent.

31 words are shared by BNCS's and BNCW's top 50 words, and among them there are no inserts. 19 words appear in BNCS, while not in BNCW. They are classified as follows: 1 inserts: "yeah", "oh", "well", "er", "right"; 2 lexical verbs: "get", "know", "think", "live", "mean", "see"; 3 negator: "no"; 4 *wh*-word: "what"; 5 conjunction: "so"; 6 adverbials: "just", "yes", "up", "then"; 7 pronoun: "me". Compared with BNC, 35 words are shared by HKS's and HKW's top 50 words, and among them, there are no inserts. 15 words appear in HKS, while not in HKW. They are classified as follows: 1 inserts: "okay", "yeah", "oh"; 2 negator: "no"; 3 lexical verbs: "think", "know", "like", "go"; 4 preposition: "about"; 5 *wh*-word: "what"; 6 conjunction: "because"; 7 adverbials: "yes", "very", "then", "just". In the top 50 words of the four corpora, BNCS has 5 inserts, while HKS has 3 inserts. Inserts are characteristic of conversations. The total frequency of inserts is much higher in the BNCS top 50 words than in the HKS top 50 words.

3.2.2 Hapax Legomena in BNCS and HKS

Altogether 2,472 hapaxes appear in BNCS, while 2,726 in HKS. 68 of the hapaxes are

compounds in BNCS, while that number for HKS is 118. That is, there are much more compounds in the hapaxes of HKS than in the hapaxes of BNCS. The examples are in Table 11.

Table 11

Examples of compounds in the hapaxes of BNCS and HKS

Corpus	Examples of compound words from hapaxes
BNCS	Backstage, bedside, childbirth, dinnertime, dustman, floorboard, guesswork, etc.
HKS	Anchorman, battleground, bottleneck, bookshop, coastline, decision-making, doorstep, filmmaker, freeman, guestroom, guidebook, hindsight, hotline, lifetime, moviemaking, etc.

Hong Kong learners' spoken English is strikingly different from native spoken English in using compound words, especially in the number of compound words, just as the above data and examples represent.

3.2.3 The Composition of the Spoken-only Words

The top 50 spoken-only words of the BNCS and HKS are listed in Table 12.

Table 12

Top 50 Spoken-only words of BNCS and HKS

BNCS	ooh(651) wan(159) aha(154) mhm(125) quid(97) bloke(82) caravan(78) bugger(75) telly(74) darling(48) teddy(46) croquet(46) whoops(41) ginger(41) lad(40) sod(39) tiny(38) jumper(38) sponge(37) pudding(37) petrol(36) fridge(34) bastard(34) peg(31) nowt(31) dice(31) bean(31) nick(29) hoop(29) cottage(29) onion(28) gracious(27) thingy(26) polish(26) daft(26) butter(26) tomato(25) sock(25) nerve(24) lemon(24) graham(24) fiver(24) dawn(24) asleep(24) wicked(23) spoon(23) oops(23) muck(23) marg(23) feather(23)
HKS	yah(401) culture(154) executive(132) legislature(111) motion(85) growth(84) mandarin(82) deputy(81) tung(75) laughter(64) gonna(61) administration(61) Asian(60) patten(59) humanity(56) database(53) barrier(49) reserve(48) provisional(48) inflation(48) supreme(47) ensure(47) democratic(47) mainland(46) legislator(46) semester(45) lecturer(45) solution(43) urban(43) status(42) gong(42) unidentified(40) overseas(40) ordinance(40) provision(39) hero(39) declaration(39) concept(39) transition(38) traditional(38) linguistic(38) focus(38) cultural(38) freedom(37) transaction(36) preparatory(36) influence(36) directly(36) putonghua(35) impact(35)

Note: the numbers in the parentheses indicate respective frequencies of the preceding words.

3.2.3.1 Inserts in Spoken-only Words

Inserts are more marginal than lexical words and function words. However, they are quite important in spoken language. In order to describe spoken English adequately, we shouldn't disregard them.

(1) Interjections

Table 13

Interjections in the top 50 spoken-only words of BNCS and HKS

Rank	Interjections in BNCS	Frequency	Interjections in HKS	Frequency
1	Ooh	651	Yah	401
2	Aha	154		
3	Whoops	41		
4	Oops	23		
Total frequency		869		401

As Table 13 shows, there are 4 interjections in the top 50 spoken-only words of BNCS, while there is only one interjection appearing in the top 50 spoken-only words of HKS. The total frequencies of the interjections from BNCS and HKS are drastically different, with the former being much higher than the latter.

(2) Response forms

Table 14

Response forms in the top 50 spoken-only words of BNCS and HKS

Rank	Responses in BNCS	Frequency	Responses in HKS	Frequency
1	Mhm	125		

As Table 14 shows, there is only one response form in the top 50 spoken-only words of BNCS, while no response form exists in the top 50 spoken-only words of HKS. The above analysis shows that Hong Kong learners tend to underuse inserts in their spoken English.

3.2.3.2 Informal Words in Spoken-only Words

Altogether, there are 6 informal words in the top 50 spoken-only words of BNCS, while none in the top 50 spoken-only words of HKS.

Table 15
Informal words in the top 50 spoken-only words of BNCS

Rank	Informal Words in BNCS	Formal Words	Frequency
1	Quid	Pound	97
2	Bloke	Person, man	82
3	Lad	Young man	40
4	Fridge	Refrigerator	34
5	Nowt	Nothing	31
6	Marg	Margarine	23
Total Frequency			307

These examples and data in Table 15 show that Hong Kong learners seldom use informal words in their spoken English.

3.2.3.3 Contractions and Abbreviations in Spoken-only Words

Mastering contractions and abbreviations can make for better comprehension of authentic communication. Through the comparison between the top 50 spoken-only words of BNCS and HKS, only one form of contraction – *gonna* (going to) – appears in HKS with the frequency of 61. Besides, there are different forms of abbreviations in BNCS and HKS spoken-only words that do not belong to the top 50 spoken-only words. The examples are given in Table 16.

Table 16
Abbreviations in BNCS and HKS

BNCS	Ave (avenue), awa (away), edu (education), choc (chocolate), cig (cigarette), frig (refrigerator), kil (kilometre), med (medicine), mitt (mitten), porn (pornography), etc.
HKS	Abo (aborigine), auto (automobile), co-ed (co-education), cont (contents), demo (demonstration), expo (exposition), math (mathematics), memo (memorandum), oft (often), op (operation), etc.

In addition to *gonna*, the examples in Table 16 have indicated that contractions and abbreviations in Hong Kong learners' spoken English are quite different from those in native spoken English.

3.2.3.4 The Phenomena of Overgeneralization in HKS-only Words

In HKS-only words, there are altogether 2 words (with the total frequency of 3) reflecting the phenomena of overgeneralization. Zhao's (2008) study shows there are some

overgeneralized words in SECCL-only words, and most of them are deviant structures of the past tense created by the Chinese EFL learners. The data and the phenomena of overgeneralization below indicate that Hong Kong learners' mastery of the past tense inflection of irregular verbs is better, compared with Chinese EFL learners.

Table 17
Overgeneralization in HKS-only words

Rank	Words Reflecting Overgeneralization	Frequency
1	Childrens	2
2	Writed	1

3.2.3.5 The Composition of the Long HKS-only Words

As the figures in Table 5 show, through the comparison in the distributions of word lengths between HKS-only words and BNCS-only words, there are 5,310 words of 10 or more than 10 letters in HKS-only, while there are only 700 words of 10 or more than 10 letters in BNCS-only. As has been covered in 3.1, there are some written style features in Hong Kong learners' spoken English. For further investigations of the two phenomena above, we will discuss two dimensions of the characteristics of the long words in HKS-only in the following two sections.

(1) The Formal Words from Long Words in the HKS-only Wordlist

Table 18
Examples of the formal words from the HKS-only wordlist

Word	Frequency	Word Length	Synonym
Assignment	34	10	Duty, task
Expenditure	33	11	Cost, fee, fare
Unidentified	40	12	Unknown
Significantly	7	13	Vitally, crucially
Recommendation	20	14	Advice, suggestion
Notwithstanding	3	15	Despite
Responsibilities	5	16	Task, duty

Table 18 shows some examples of long formal words with their frequencies, word lengths, and corresponding shorter synonyms in the HKS-only wordlist. This is a reflection of the written style features in Hong Kong learners' spoken English.

(2) The Word Formation Processes of Long Words in the HKS-only Wordlist

Table 19
Examples of the main word formation processes of the long HKS-only Words

Com-pounding	birthplace (1), bottleneck (1), headmaster (2), expressway (1), film-making (2), goal-keeper (1), air-condition (1), fair-minded (1), grand-stand (1), etc.
Derivation	(adjectives) nationwide (3), multipurpose (1), controversial (17), irre-sponsible (4), post-secondary (8), retrospective (2), respiratory (6), ex-tracurricular (3), etc.
	(adverbs) adequately (2), constitutionally (2), etc.
	(nouns) accordance (10), complexity (2), adjustment (14), ex-girlfriend (2), co-existence (2), sub-committee (5), fundamentalist (2), roadwor-thiness (7), etc.

Note: the numbers in the parentheses indicate respective frequencies of the preceding words

As the examples in Table 19 show, the main word formation processes in the long HKS-only words are compounding and derivation. The examples above are concrete and they mainly cover three parts of speech – nouns, adjectives, and adverbs. Besides, the long words are formed from different perspectives.

3.2.4 Summary of the Qualitative Analysis

According to the analyses from 3.2.1 to 3.2.3, the following features have been revealed. Firstly, the difference in the rank-1 words of the top 50 words in BNCS, BNCW, HKS, and HKW shows Hong Kong learners tend to show some written style features in their spoken English. Secondly, there are more compounds in the hapaxes of HKS than in the hapaxes of BNCS. Thirdly, Hong Kong learners tend to underuse inserts and informal words in their spoken English. Fourthly, there is a significant difference between Hong Kong learners' spoken English and native spoken English in the use of contractions and abbreviations. Fifthly, Hong Kong learners' mastery of the past tense inflection of irregular verbs is better, compared with Chinese EFL learners. Sixthly, in the long HKS-only words, some formal words have shorter and simpler synonyms to be replaced by, and their main word formation processes are compounding and derivation.

4 Conclusion

The major findings are summarized as follows: (1) In terms of vocabulary size per ten million tokens of BNCS, BNCW, HKS, and HKW, the value for BNCW is larger than

the one for HKW, while the value for BNCS is smaller than the one for HKS. (2) The mean word length of HKS is greater than the one of BNCS. (3) As to the lexical density of BNCS, BNCW, HKS, and HKW, the value for BNCW is larger than the one for HKW, while the value for BNCS is smaller than the one for HKS. (4) The values for the lexical coverage in HKS are smaller than their counterparts of the same number of types in BNCS, while it is not the same case in HKW and BNCW. (5) In the top 50 words of the four corpora in comparison, BNCW, HKS, and HKW share the same rank-1 word — the determiner *the*. HKS has 3 inserts, while BNCS has 5 inserts in the top 50 words. (6) HKS has more hapaxes than BNCS, and there are more compounds in the hapaxes of HKS than in those of BNCS. (7) Hong Kong English learners tend to underuse inserts and informal words, and there is a significant difference between Hong Kong learners' spoken English and native spoken English in the use of contractions and abbreviations. (8) Hong Kong learners' mastery of past the tense inflection of irregular verbs is better, compared with Chinese EFL learners. (9) Small proportions of the long HKS-only words are quite formal and can have shorter and simpler synonyms to be replaced by. The main word formation processes of the long HKS-only words are compounding and derivation.

After a systematic analysis, the following conclusions can be drawn. Firstly, compared with the differences between BNCS and BNCW, the differences between Hong Kong learners' spoken and written English are slight. Secondly, there are some written style features in Hong Kong learners' spoken English. Thirdly, there are differences between Hong Kong learners' spoken English and native spoken English in some conversation-specific characteristics. Therefore, more teaching materials on authentic daily conversations in English-speaking countries should be introduced and adopted. Fourthly, the phenomena of overgeneralization exist in HKS. Both teachers and learners should pay enough attention to linguistic irregularities during the teaching and learning processes. Fifthly, Hong Kong learners' vocabulary size in spoken English is quite large. Therefore, teachers and students should improve themselves in vocabulary teaching and learning, especially in the practical use of word formation processes.

The pedagogical outcomes on English teaching in mainland China are as follows. Firstly, since the difference between Hong Kong learners' spoken English and written English is not as obvious as the one between native spoken English and written English, a greater importance should be attached to the oral English teaching and learning. Secondly, since Hong Kong learners' spoken English show some written style features, an introduction to variations across speech and writing is suggested in classroom English teaching. Thirdly, more and more authentic spoken materials ought to be introduced and adopted for teaching to make learners aware of some specific characteristics of spoken English. Fourthly, teachers and learners should pay enough attention to the memory reinforcement and practical use of linguistic irregularities during the process of foreign language teaching and learning. Besides, improvements need to be made in vocabulary acquisition and learning to enlarge learners' vocabulary – for instance, by the use of compounding.

Due to the limitation in space and focus of the paper, only some lexical characteristics have been dealt with in part three. Some other lexical characteristics – such as hedges, small words, vocabulary increase ratio, and so on – should be studied in the future studies. The degree of written characteristics of HKS needs to be studied further. Moreover, much more work needs to be done to give teachers and learners some insights into the scientific study of a second language.

References

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review*, *59*(3), 393-424.

Krashen, S. & Terrel, T. (1983). *The Natural Approach: Language Acquisition in the Classroom*. Oxford: Pergamon.

Granger, S., & Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (ed.), *Learner English on Computer*. London & New York: Longman. pp. 119–131.

McCarthy, M. (1984). A new look at vocabulary in EFL. Applied Linguistics 5(1), 12–22.

Nation, P. (1990). *Teaching and Learning Vocabulary.* New York: Newbury House.

Petch-Tyson, S. (1998). Writer/Reader Visibility in ESL Written Discourse. In S. Granger (ed.), *Learner English on Computer*. London & New York: Longman.

Simpson, J. (1998). The New Vocabulary of English. In E. G. Stanley and T. F. Hood. *Words.* Cambridge: D. S. Brewer, 143–152.

Wen Qiufang, Ding Yanren & Wang Wenyu (2003). Features of oral style in English compositions of advanced Chinese EFL learners: An exploratory study by contrastive learner corpus analysis. *Foreign Language Teaching and Research* 7, 268–274.

Wen Qiufang & Ding Yanren (2004). A Corpus-based Analysis of Frequency Adverbs Used by Chinese English Majors. *Modern Foreign Languages* 5, 150–156.

Wen Qiufang (2006). Patterns of Change in Speaking Vocabulary Development by English Majors across Four Years. *Computer-assisted Foreign Language Education* 8, 3–8.

Wilkins. D. (1996). ESL vocabulary learning in a TOEFL preparation class: A case study. *Canadian Modern Language Reviews* 53(1), 97–119.

Zhao Yanxue. (2008). *Lexical Characteristics of Chinese EFL Learners' Spoken English.* Dalian: Dalian Maritime University Press.

Glottometrics 47, 2019, 83-103

A Study on the Subjectival Position and the Syntactic Complexity in Spoken English

Pianpian Zhou[1]

Abstract. The sentence is considered the key unit of syntax. In quantitative linguistics, there are many ways to probe the inter-relationships among constituents of sentences, such as length, complexity, position and frequency, etc. The subject of the sentence is also an important constituent. It usually works as an unmarked theme, which is the point of departure of the message. This paper is corpus-based and employs both quantitative and qualitative methods, aiming to study the relationship between the subjectival position and the sentential syntactic complexity in the spoken part of The British Component of the International Corpus of English (ICE-GB). The result of the study shows that the phrasal syntactic function elements (PSFEs) in ICE-GBS serve 43 different sentential syntactic functions and the ten most frequent PSFEs account for 91.61% of the total. The sentential syntactic complexities in ICE-GBS range from 1 to 126. The number of sentences increases along with the sentential syntactic complexity until reaching the peak, and then begins to decrease. The number of sentential structural variations increases along with the sentential complexity until reaching the peak. Then, it begins to decrease. In ICE-GBS, the sentential subjects appear in 43 different positions in the sentences, with the predominant position of 1. The sentential subjectival position can indicate the sentential syntactic complexity – that is, the later the subject appears, the more syntactically complex the sentence.

Keywords: *subjectival position, syntactic complexity, quantitative linguistics, corpus, spoken English, model*

1 Introduction

Among syntactic functions, the subject usually works as an unmarked theme, which is the point of departure of the message (Halliday, 1994). The information that it conveys is often thought as given, while the following parts show something new. The inclination to explain brand new information in detail implicates that the constituents after the subject are likely to be long. This phenomenon echoes the rule of end-weight. Since the SVO structure is an overwhelming construction in English sentences, it seems that the

[1] Correspondence to Pianpian Zhou, New Oriental Education & Technology Group, Hefei, China. Email: susiezhoupianpian@foxmail.com

position of subject is connected to the length of the sentence, therefore the complexity of the sentence.

Spoken language is always of interest in the linguistic studies. Many distinctions exist between spoken and written forms. This indicates that the results of studies about the subjectival position and the sentence complexity in spoken and written language might be different. It is reasonable to explore the similarities and differences, and contribute thus to further understanding of natural English language.

This paper concentrates on exploring the distribution of subjectival positions in ICE-GBS and calculating the complexity of sentences by counting the number of phrasal syntactic functional elements (PSFEs), such as nominal groups working as subjects, since the corpus has already been syntactically tagged. Finding out the law of relations between these two concepts is the final aim.

2 Methods and Materials

2.1 Corpus Used in the Study

The International Corpus of English (ICE) is a set of corpora representing varieties of English from around the world. The British Component of the International Corpus of English (ICE-GB) is just one among many ICE components. Like all the ICE corpora, ICE-GB consists of a million words of spoken and written English and adheres to the common corpus design. 200 written and 300 spoken texts make up the million words. Every text is grammatically annotated, permitting complex and detailed searches across the whole corpus.

The majority of texts in each component of ICE are derived from spoken data. ICE corpora contain 60% (600,000 words) of orthographically transcribed spoken English, since the father of the project insists on the primacy of the spoken word. ICE-GBS is made up of 180 dialogues and 120 monologues.

In this paper, only the spoken part, that is the ICE-GBS, will be used. Table 1 shows the major syntactic functions and syntactic classes in ICE-GBS.

Table 1
Major syntactic tags in ICE-GBS

A	*adverbial*	MVB	*main verb*
AJPO	*adjective phrase postmodifier*	NOOD	*notional direct object*
AJPR	*adjective phrase premodifier*	NOSU	*notional subject*
AVPO	*adverb phrase postmodifier*	NPHD	*noun phrase head*

AVPR	*adverb phrase premodifier*	NPPO	*noun phrase postmodifier*
CF	*focus complement*	NPPR	*noun phrase premodifier*
CJ	*conjoin*	OD	*direct object*
CO	*object complement*	OI	*indirect object*
CS	*subject complement*	OP	*operator*
CT	*transitive complement*	PARA	*parataxis*
DEFU-NC	*detached function*	PC	*prepositional complement*
DISMK	*discourse marker*	PMOD	*prepositional modifier*
DT	*determiner*	PREDGP	*predicate group*
DTCE	*central determiner*	P	*prepositional*
DTPO	*determiner postmodifier*	PROD	*provisional direct object*
DTPR	*determiner premodifier*	PRSU	*provisional subject*
DTPS	*postdeterminer*	SBMO	*subordinator phrase modifier*
FNPPO	*floating NP postmodifier*	SU	*subject*
FOC	*focus*	SUB	*subordinator*
INDET	*indeterminate*	TAGQ	*tag question*
		VB	*verbal*

2.2 Tools for Data Processing

To process the corpus, several instruments were employed to get the data required. These tools concluded Perl, SPSS, and Visual FoxPro.

Perl, which is the short form of Practical Extraction and Reporting Language, is a computer programming language designed for multitasking and real-time programming. It is a family of high-level, general-purpose, interpreted, dynamic programming languages. Perl 5.10 was employed in this paper.

SPSS is the software name standing for Statistical Package for the Social Sciences. It is a widely used program for statistical analysis in social science. It is also employed in many other fields widely. SPSS has also a number of functions to summarize and display data in the form of tables and graphs. In the present study, SPSS 16.0 was used.

Visual FoxPro is a data-centric object-oriented procedural programming language produced by Microsoft. It is a full-featured, dynamic programming language that does not require the use of an additional general-purpose programming environment. FoxPro 9.0 assisted this study.

2.3 Procedures for Data Collection

Since ICE-GBS is a corpus which is heavily coded, the primary step was to get the codes necessary for this paper and to remove other codes which signify the boundary and part of speech for each word – such as codes for speech turns, pausing and paralinguistic features like laughter in spoken texts, paragraphs, sections, headings, and

meta-textual information about the source of encoding of individual texts. The only codes needed here were those on the level of syntax. This work was finished by Perl and FoxPro.

After getting the original syntactical codes, the number and distribution of phrasal syntactic elements were found by using the aforementioned software. Generally, there are three ways to measure the complexity of a sentence. The first one is based on the number of clauses. The second one is by counting the number of phrases, and the last one by calculating the total number of words. In this study, the second method was employed while the numbers of phrasal syntactic functional elements such as subjects, objects were counted, since the subject is the syntactic function of a phrase. Then, variations of subjectival positions and their distributions were listed. In this paper, the subjectival position refers to the position of the first subject of the entire sentence as captured from the beginning of the sentence. Some mathematical models were applied in the study. The detailed figures and tables were drawn by SPSS.

3 Results and Discussion

3.1 Syntactic Structures and Complexity in ICE-GBS

This section introduces different syntactic structures in ICE-GBS and their complexities.

3.1.1 Number and Distribution of Variations

The total number of sentences in ICE-GBS is 60,866, among which 18,705 are non-clause sentences. For example, the sentence, the syntactic tags of which are ELE NPPO PC ELE NPPO, is a non-clause sentence. In this study, these non-clause sentences are excluded, and only the complete sentences are used as data source. The actual number of sentences processed is 42,161. These sentences consist of 508,215 phrasal elements which serve 43 different syntactic functions. Figure 1 displays the distribution of these PSFEs.

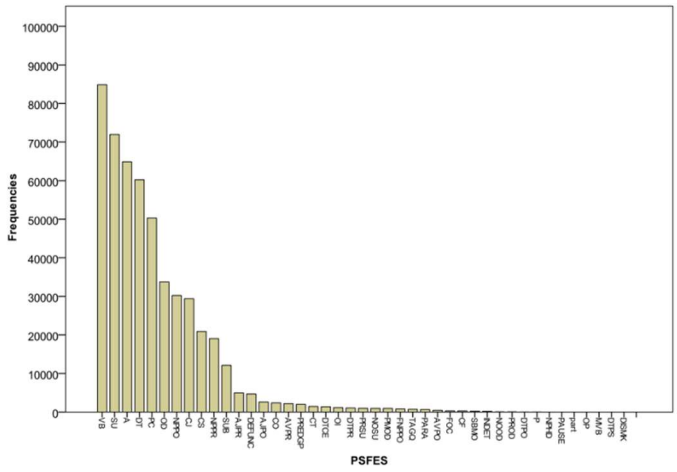

Figure 1. The Distribution of PSFEs

As the figure shows, some of the elements are frequent, while others are rarely examined. The ten most frequent PSFEs are VB, SU, A, DT, PC, OD, NPPO, CJ, CS and NPPR. Their frequencies and percentages can be shown in Table 2.

Table 2
Ten most frequent PSFEs

PSFEs	Frequencies	Percentages
VB	84,852	16.70%
SU	71,993	14.17%
A	64,882	12.77%
DT	60,229	11.85%
PC	50,296	9.90%
OD	33,734	6.64%
NPPO	30,228	5.95%
CJ	29,386	5.78%
CS	20,918	4.12%
NPPR	19,038	3.75%
Total	465,556	91.61%

According to Table 2, it is clear that these ten elements account for 91.61% of all the 508,215 structures in ICE-GBS, with all the left elements accounting for only 8.39%.

PSFEs served as subjects are the second most frequent elements, which implicates that the study on subjectival positions is well-justified and the results are probably representative.

3.1.2 Distribution of Syntactic Complexity

The complexity of a sentence is the number of PSFEs in it. For instance, the syntactic complexity of a sentence, the syntactic tag of which is SU VB CS AJPR, is 4. The complexities of all the sentences in ICE-GBS, and the frequencies of sentences with such syntactic complexities are shown in Table 3.

The mean sentential syntactic complexity is calculated by dividing the total number of the PSFEs, 508,215, by the total number of sentences, 42,161. The value is 12.05. According to Table 3, the median of the sentential syntactic complexity is 9, and the mode is 4. The sentential syntactic complexities range from 1 to 126. However, the first quartile of the total number of sentences, 11,355, consists of sentences with syntactic complexities from 1 to 5. Those with syntactic complexities between 6 and 18 constitute the inter-quartile, totalling 20,601, while those with syntactic complexity of 19–126 constitute the last quartile, totalling 10,205. Out of all the 42,161 sentences, only one has the syntactic complexity of 126, and 331 have the complexity of 1. The sentence with a complexity of 126 is as follows:

> A CJ SUB A PC SU A VB OI OD DT DEFUNC CJ SUB SU VB A PC DT DTCE CJ SU VB A OD DT SU VB A A PC NPPO PC DT CJ A PC VB OD DT SU VB CS AJPR AJPO PC DT SU VB A A A SUB SU VB CS DT NPPR AJPR AJPR SU VB OD DT NPPO SUB SU DT NPPO VB OD PREDGP CJ VB A OD NPPR CJ VB OD CJ A VB OD A PC CJ VB CT SU VB A PC DT A PC DT CJ VB OD DT NPPR CO A PC A VB OD NPPO PC DT NPPR DT DTCE NPPO PC CJ DT NPPO PC CJ DT NPPO PC CJ NPPO

Table 3
Complexity and Frequency of Sentences

Complexity	Frequency	Complexity	Frequency
1	331	49	41
2	1,932	50	30
3	2,950	51	37
4	3,141	52	30
5	3,001	53	25
6	2,914	54	21

7	2,598	55	32
8	2,202	56	20
9	1,959	57	22
10	1,726	58	18
…	…	…	…
47	50	126	1
48	52		

3.1.3 Relationship between Complexity and Frequency of Sentences

On the data in Table 3, the relationship between the complexity of syntactic structures and the number of sentences with such complexity can be displayed. The frequency of sentences increases along with the sentential syntactic complexity, until reaching the peak of 3,141 at the complexity of 4; then, it begins to decrease. Sentences rarely appear with the complexity larger than 65 appear rarely.

The reparametrized Tuldava's model (Tuldava, 1980) can describe the relationship between frequency of syntactic structures and their complexity. The original Tuldava's model is used for the description of the relationship between text length and vocabulary size. It is shown below:

$$V = Ne^{-\alpha(\ln N)^{\beta}} \tag{1}$$

(1) is adjusted by adding a parameter γ to it, to fit the ratio data:

$$V = \gamma Ne^{-\alpha(\ln N)^{\beta}} \tag{2}$$

(2) can describe the relationship between the syntactic sentential complexity and the corresponding frequency of sentences. In the model, V stands for the frequency of sentences and N for the complexity, while α, β, γ are parameters. After the processing of data in SPSS, the model fits the observed values very well, with R^2= 0.98, α = 2.6747, β =1130.5992 and γ = 0.1906. Figure 2 is the model fit. The solid line is the model fit and small circles represent the empirical values.

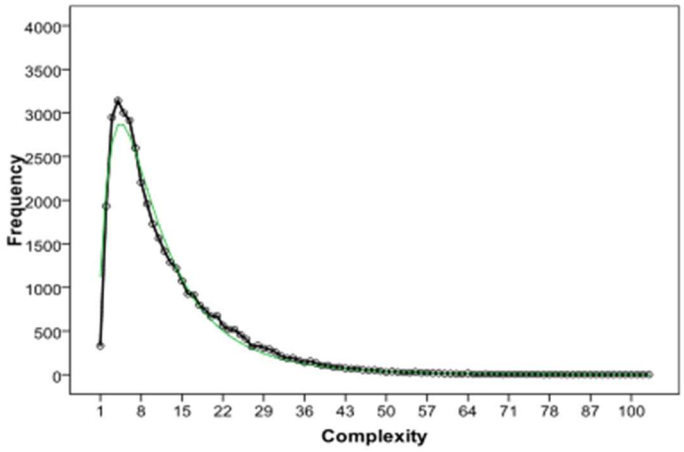

Figure 2. The relationship between complexity and frequency of sentences

3.1.4 Relationship between Complexity and Structural Variations

Table 4 shows the sentential syntactic complexity and the corresponding sentential structural variation in ICE-GBS.

Table 4
Syntactic complexity and structural variation

SC	SV	SC	SV
1	9	49	41
2	36	50	30
3	134	51	37
4	342	52	30
5	710	53	25
6	1,151	54	21
7	1,598	55	32
8	1,784	56	20
9	1,815	57	22
10	1,701	58	18
…	…	…	…
47	50	126	1
48	52		

It is clear from Table 4 that the number of sentential structural variations increases along with the sentential complexity until reaching the peak of 1,815 with the complexity of 9. Then, it begins to decrease. There are only nine variations with the complexity of 1, and only one variation with the complexity of 126. A large number of variations appear when the complexity is from 6 to 17, while very few variations occur when the complexity is larger than 64. The structures of the 36 sentences whose syntactic complexity is 2 are displayed in Table 5.

Table 5
Variations with the complexity of 2

1	SU VB	19	VB DEFUNC
2	VB A	20	CS VB
3	SU CS	21	VB CO
4	VB OD	22	A A
5	CS SU	23	ELE PC
6	VB SU	24	A SUB
7	SU A	25	A PC
8	VB CS	26	PRSU VB
9	A VB	27	A DT
10	A SU	28	SU OD
11	VB OI	29	SU PARA
12	SUB SU	30	SUB VB
13	SU DT	31	OD DT
14	SU SU	32	SU NPPR
15	OD SU	33	A AVPO
16	SU DEFUNC	34	INDET SU
17	DEFUNC VB	35	SUB A
18	SU NPPO	36	PU INDET

The Nemcová and Serdelová's (2005) model can be used to define the relationship between the sentential syntactic complexity and sentential structural variations. It was originally employed to describe the relationship between the number of synonyms (y)

of a word and the length of the word in syllables (x):

$$y = ax^b e^{cx} + 1 \tag{3}$$

(3) is a special case of Wimmer & Altmann (2005). The model can be adopted to define the relationship between sentential complexity and structural variations as follows, with SV standing for structural variations and SC standing for syntactic complexity, while a, b, c are parameters.

$$SV = aSC^b e^{cSC} + 1 \tag{4}$$

The model fit the observed data very well, with $R^2 = 0.959$, $a = 39.858$, $b = 2.789$, $c = -0.274$. Figure 3 shows the fit. The smooth dotted line is the model fit, and the small circles are the empirical values.

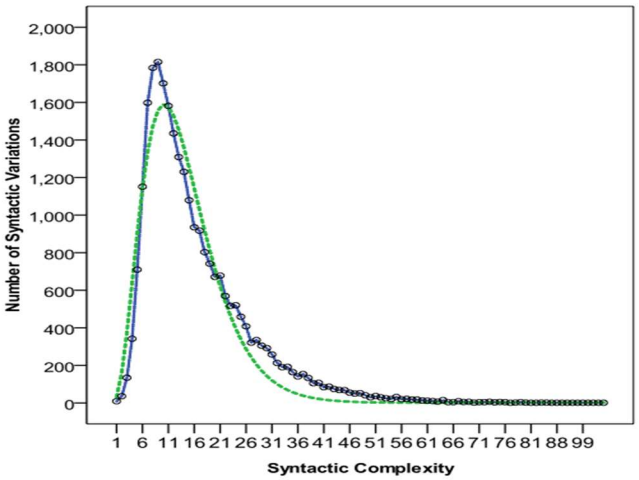

Figure 3. The relationship between complexity and structural variation

3.1.5 Comparison between Results of ICE-GBS and ICE-GBW

According to the data from Fan et al. (2013) and this study, similarities and differences between sentential complexities of ICE-GBS and ICE-GBW are presented.

In both ICE-GBS and ICE-GBW, the ten most frequent PSFEs are VB, SU, A, DT, PC, OD, NPPO, CJ, CS, and NPPR. They account for over 90% of all the syntactic function elements in both parts, while the other elements rarely appear. Both ICE-GBS and ICE-GBW have a wide range of syntactic complexity. Sentences with complexities from 1 to 26 account for the major part, while sentences with complexities larger than 26 are not common. In ICE-GBS and ICE-GBW, the relationship between sentential syntactic complexity and sentential structural variations within a number of different

structures can be displayed by the same model. Both of the data fit the model very well. The number of sentences with different sentential structures increases along with the sentential syntactic complexity to a certain point, and then begins to decrease.

There are 36 different syntactic function tags in ICE-GBW while 43 kinds of PSFEs exist in ICE-GBS. DB is only found in the written part while DISMK, DTPS, MVB, NPHD, OP, P, PARA, and PAUSE can only be examined in the spoken part. The most frequent syntactic function in ICE-GBS is VB, while the one in ICE-GBW is DT. The syntactic complexity of ICE-GBS ranges from 1 to 126 while the one of ICE-GBW ranges from 1 to 95. This might be attributed to pauses and delays in spoken texts. The median, mode, and mean sentential complexity of ICE-GBS are 9, 4, and 12.02 respectively, smaller than the ones of ICE-GBW, which are 17, 14, and 18.60. This implicates the brevity of spoken language and the fact that written language tends to be elaborated descriptions. The difference is explicitly shown in Figure 4. The solid rectangle represents the data of ICE-GBS, and the rectangle with slashes shows the data of ICE-GBW.

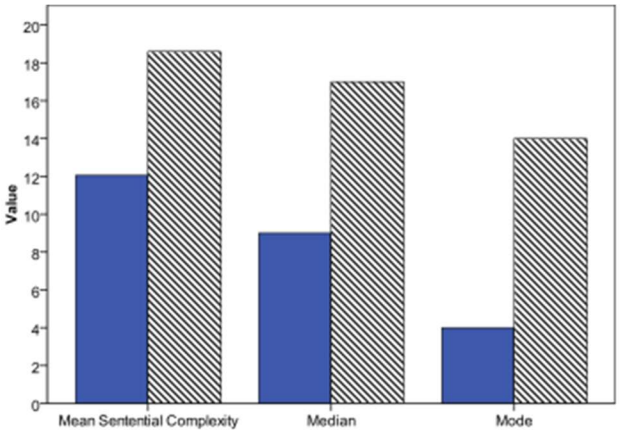

Figure 4. Differences between complexities in ICE-GBS and ICE-GBW

3.2 Subjectival Positions and the Distribution

This section introduces subjectival positions in ICE-GBS and their distributions.

3.2.1 Number of Subjectival Positions and Distribution

There are 34 different sentential subjectival positions in the 42,161 sentences of ICE-GBS. There are 1,714 sentences in which there is no subject. They are marked with the position of 200, and are not included in the calculation. The rightmost position in all the sentences is 43. Generally, the number of the sentences with the corresponding subjectival position decreases along with the increase of the sentential subjectival position. The sentential subjects of most sentences are in position 1. The number of sentences with the position of 1 is 28,059, accounting for 66.55% of all the sentences in ICE-GBS.

The number of sentences whose sentential subjectival position is larger than 10 is only 361, accounting for only 0.86%. Table 6 shows the sentential subjectival position and the number of sentences with such position.

Table 6
Subjectival Position and Number of Sentences

Subjectival Position	Number of Sentences
1	28,059
2	6,051
3	2,603
4	1,241
5	648
6	477
7	406
8	264
9	159
10	108
…	…
33	2
43	1

The syntactic tags of the sentence which has a subjectival position of 43 are as follows:

PU CJ VB OD A PC CJ PC VB OD NPPR NPPR A A CJ PC NPPR CJ PC NPPR NPPO PC DT CJ VB A OD NPPR CJ VB OD NPPR CJ A PC A PC VB A A SUB VB SU DT NPPR DEFUNC AJPO SUB SU VB AJPO PC CJ DT NPPO PC DT CJ DT NPPR NPPO PC DT NPPR

3.2.2 Relationship between Subjectival Position and Structural Variations

The relationship between the subjectival position and the sentential structural variations in a number of different structures can be invesitigated. Table 7 displays the subjectival position and the corresponding sentential structural variations in ICE-GBS.

Table 7
Subjectival position and structural variations

Subjectival Position	Structural Variations
1	15,755
2	4,661
3	2,335
4	1,176
5	620
6	472
7	404
8	264
9	159
10	108
…	…
33	2
43	1

According to Table 7, the number of sentential structural variations generally decreases as the subjectival position increases. There are 15,755 variations with the subject position of 1, and only one variation with the position of 43. A large number of variations appear when the subjectival position is from 1 to 10, while very few variations occur when the position is larger than 20. 20 structural variations of the 44 sentences the subjectival position of which is 14 are displayed in Table 8 as follows.

Table 8
20 Variations with the position of 14

1	A A CJ PC DT CJ PC DT A A DT CJ CJ SU DT NPPR VB A VB OD DT
2	A A PC A DEFUNC DT NPPR A PC DT NPPO PC DT SU DT NPPO PC DT VB A VB OD NPPO PC DT
3	A PC CJ VB A AVPR CJ A VB OD DT NPPO PC SU VB CS AJPO A VB A PC NPPR AJPR AVPR
4	A PC DT DTPR NPPO PC NPPO PC DT NPPR NPPO A PC SU CJ DT NPPR CJ VB SU DT DTPO VB CO PC AJPR

5	A PC DT NPPO NPPO PC DT NPPO PC DEFUNC DT DTCE DT SU VB A PC A PC TAGQ SU
6	A PC DT NPPR NPPO PC NPPO PC NPPO PC DT DTPR AVPR SU VB A CS DT NPPO PC DT NPPR NPPR
7	A PC DT VB FOC NPPO PC NPPO PC NPPO PC DT CF SU VB CT SU PREDGP CJ VB CS CJ VB OD
8	A PRSU VB CS NOSU VB OD CJ DT NPPO VB A A SU DT NPPO PC
9	CJ A VB A PC DT NPPO VB A A PC DT CJ SU A VB
10	CJ PRSU VB CS A PC DT NPPO PC NOSU VB OD SUB SU VB CS CJ A SUB SU DT VB CS SU VB
11	CJ SUB VB OD VB CS DT CJ VB A PC NPPO NPPO SU VB CS DT NPPR
12	CJ VB OD A A PC CJ VB OD A AVPO PC NPPO SU A PC DT VB OD DT DTPR
13	DEFUNC DT DT NPPO PC VB OD A PC DT VB OD NPPO SU VB
14	PRSU A VB A OD DT DTPR AVPR NOSU VB OD A TAGQ SU
15	PRSU VB CS A PC DT NOSU SUB A PC DT NPPO PC SU DT VB A CS AJPO VB PARA SU VB A CS DT
16	PRSU VB CS NOSU VB CS AJPO VB A PMOD PC A SUB SU VB OD NPPO VB A
17	PRSU VB CS NOSU VB OD A A VB OD CO AJPO SUB SU VB
18	PRSU VB CS NPPO PC DT NOSU VB A PC DT OD SUB SU VB OD DT
19	VB A PC DT NPPO PC DT DEFUNC DT A PC DT NPPR SU NPPR DT
20	VB OD DT NPPO PC DT A PC CJ DT CJ A SUB SU VB OD DT

Altmann (1980) proposes that the length of a component is inversely proportional to the length of its superstructure according to:

$$y = ax^{-b}e^{-cx} \tag{5}$$

(5) can be adopted to define the relationship between subjectival positions and the syntactic variations in a number of different structures. The model is as follows:

$$y = ax^{-b} \tag{6}$$

(6) can be adopted to describe the data in this study:

$$SV = aSP^{-b} \tag{7}$$

In (7), *SV* represents the syntactic variations in a number of different structures, and *SP* stands for the subjectival position, with *a*, *b* serving as parameters. The model fit can be shown in Figure 5. The solid line represents the model fit, and the small circles stand for the empirical values.

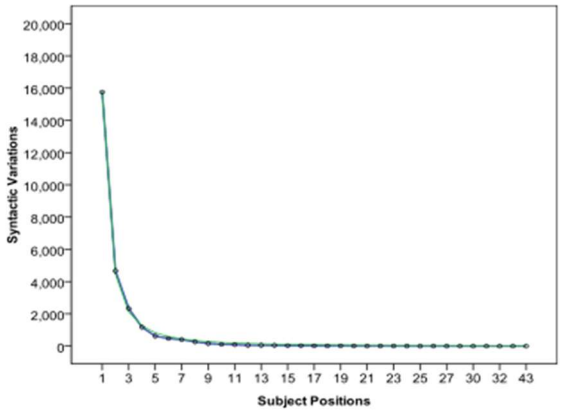

Figure 5. The relationship between subjectival position and structural variations

The two lines in the figure almost overlap completely, which shows that the observed data fit the model very well, with $R^2 = 0.999$, $a = 1$, $b = 0.294$.

3.2.3 Comparison between Results of ICE-GBS and ICE-GBW

In both ICE-GBS and ICE-GBW, the subject of most sentences is in the position of 1. Sentences with the subject position of 1 account for 66.55% of all the sentences in ICE-GBS, while they account for 62% in ICE-GBW. Sentences the sentential subjectival positions of which are larger than 10 account for only 0.86% in ICE-GBS, and they account for only 1.55% in ICE-GBW. No matter the form is spoken or written, the number of sentences with different subjectival positions generally decreases along with the subjectival position increase. ICE-GBS and ICE-GBW have a similar range of subjectival positions, with 1 to 34 in the spoken part and 1 to 32 in the written part. In ICE-GBS and ICE-GBW, the numbers of sentences with different sentential structures decrease along with the increase of the subjectival position.

There are 32 different subjectival positions in ICE-GBW while 34 kinds of subjectival positions exist in ICE-GBS. Sentences with the positions of 29, 32, 33, and 43 are only found in the spoken part, while those of 44, and 46 can only be found in the written part.

3.3 Complexity in Different Subjectival Positions and the Relationship between them

This section introduces the distribution of the sentential syntactic complexity in different subjectival positions and the relationship between them.

3.3.1 Distribution of PSFEs with Different Subjectival Positions

The number of PSFEs on a corresponding subjectival position is counted to help the calculation of the mean sentential syntactic complexity with different subjectival positions, which can be obtained through dividing the number of PSFEs linked to a particular subjectival position by the number of sentences with such position. Table 9 shows the distribution.

Table 9
Distribution of PSFEs with different positions

Subjectival Position	Number of PSFEs
1	245,767
2	79,540
3	41,973
4	21,827
5	10,755
6	9,023
7	7,895
8	5,359
9	3,746
10	2,643
…	…
33	111
43	66

3.3.2 Distribution of Complexity with Subjectival Position

This section displays the distribution of the sentential complexity with the sentential subjectival position. Since the number of sentences the subjectival positions of which are larger than 10 is very small, only the distribution of sentential complexity with the sentential subjectival positions from 1 to 10 will be shown in this part. Figure 6 shows the distribution of syntactic complexity with the subjectival positions from 1 to 10. Panel A to panel J represent position 1 to position 10 respectively. The *x* axis represents

the syntactic complexity in a number of PSFEs, and the *y* axis represents the number of sentences with such complexities.

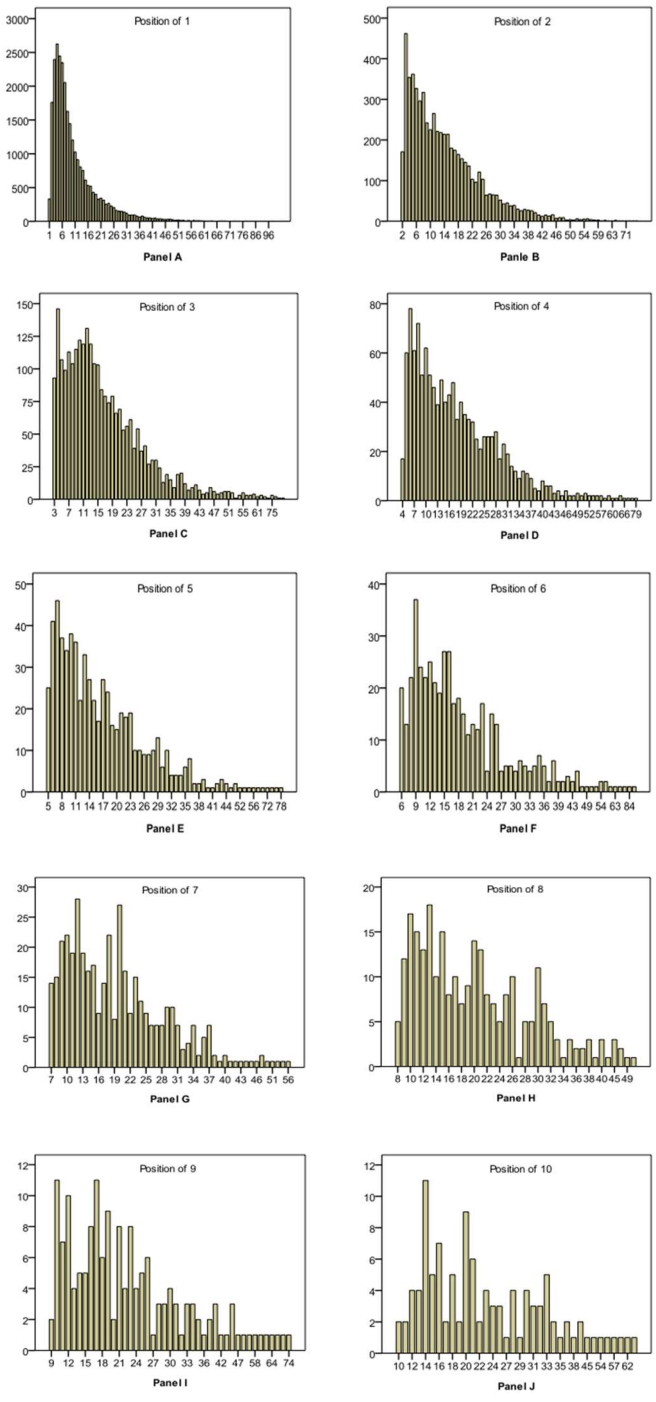

Figure 6. Distribution of the syntactic complexities with subjectival positions from 1 to 10

The general distribution patterns of the sentential syntactic complexities with the sentential subjectival positions from 1 to 10 in ICE-GBS are similar, which are skewed with a long right tail.

3.3.3 Relationship between Syntactic Complexity and Subjectival Position

The mean sentential syntactic complexity of the sentences with the same sentential subjectival position is computed in order to study the relationship between sentential syntactic complexity and sentential subjectival position. It can be calculated on the basis of the data from Table 6 and Table 9. The results are shown in Table 10.

Table 10
Subjectival position and mean sentential complexity

Subjectival Position	Mean Sentential Complexity
1	8.7589
2	13.1446
3	16.1241
4	17.5866
5	16.5941
6	18.9119
7	19.4409
8	20.2917
9	23.5472
10	24.4537
…	…
33	54.5000
43	64.0000

The mean sentential syntactic complexity is a function of the sentential subjectival position. This relationship is linear and can be expressed with the following linear regression model, with α, β as parameters:

$$y = \alpha + \beta x \tag{8}$$

(8) can be adopted to describe the relationship between mean sentential syntactic complexity (*MSSC*) and subject position (*SP*) as follows:

$$MSSC = \alpha + \beta SP \tag{9}$$

The model fit is shown in figure 7.

The small circles are the observed values of sentential syntactic complexity and the straight line represents the model fit. The data fit the model well, with $R^2 = 0.923$, $\alpha = 1.133$, $\beta = 11.046$. Generally, the sentential syntactic complexity increases alongside with the increase of the corresponding sentential subjectival position. In other words, the later the subject appears, the more syntactically complex the sentence.

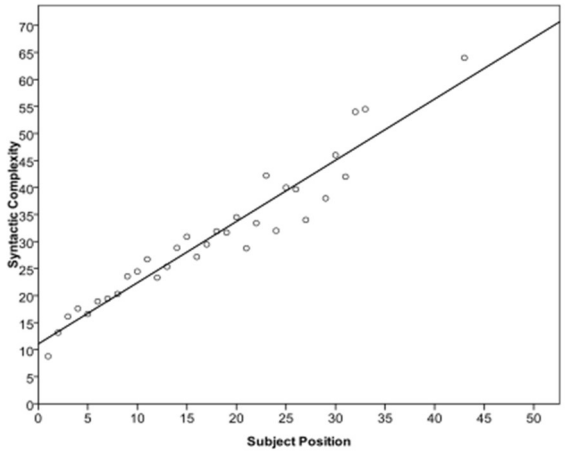

Figure 7. The linear regression model for the relationship between subjectival position and sentential syntactic complexity

3.4.4 Comparison between Results of ICE-GBS and ICE-GBW

In both ICE-GBS and ICE-GBW, the number of PSFEs generally decreases as the sentential subjectival position increases. The major part of PSFEs in sentences appears with the subjectival position of 1. The distribution of sentential complexity with the sentential subjectival positions from 1 to 10 is similar in both ICE-GBS and ICE-GBW. The patterns in the figures are skewed, tailing out to the right. No matter whether the form is spoken or written, the mean sentential syntactic complexity increases along with the motion of subjectival position to the right of the sentence. The data on the mean sentential syntactic complexity and subjectival position in both ICE-GBS and ICE-GBW fit the same linear regression model well.

The numbers of sentences with the subjectival positions from 8 to 10 in ICE-GBS is lower than the one in ICE-GBW, this has been found out through comparing the last three panels of the ten successive panels which show the distribution of sentential complexity with the sentential subjectival positions of 1 to 10. The values of mean sentential syntactic complexity in ICE-GBS and ICE-GBW are very different although they share the same model. The difference can be shown in Figure 8. The *x* axis represents the sentential subjectival position, and the *y* axis stands for the mean sentential syntactic complexity. The dotted line with small circles represents the data of ICE-GBW, and the

solid line with small diamonds stands for the data of ICE-GBS.

According to the figure, it is clear that the mean sentential syntactic complexity with the same subjectival position in ICE-GBW is generally higher than that in ICE-GBS, which implicates that the sentences in written texts of ICE-GB are more complex than those of spoken texts.

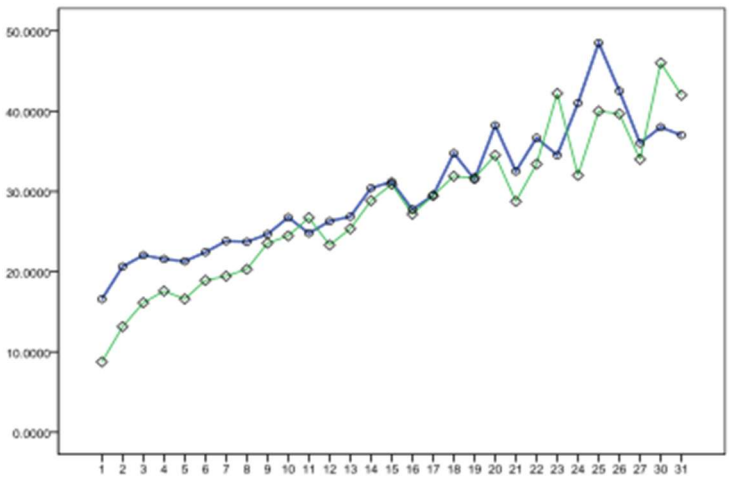

Figure 8. The mean sentential syntactic complexity in ICE-GBS and ICE-GBW

4. Conclusion

In this corpus-based study, the data show that the frequency of sentences increases along with the sentential syntactic complexity until reaching the peak; then, it begins to decrease; the number of sentential structural variations increases along with the sentential complexity until reaching the peak, and then it begins to decrease; there is a certain relationship between the sentential subjectival position and the structural variations in a number of different syntactic structures; the number of sentential structural variations generally decreases as the subjectival position moves to the right of the sentence. The number of PSFEs generally decreases with the increase of the sentential subjectival position. The relationship can also be applied to ICE-GBW. The mean sentential syntactic complexity with the same subjectival position in ICE-GBW is generally higher than the one in ICE-GBS.

Generally, in both ICE-GBS and ICE-GBW, sentential subjectival position can indicate sentential syntactic complexity. That is, the later the subject appears, the more syntactically complex the sentence. The result shows the reliability of the subjectival position as an indicator of sentence complexity. Due to time and space limitations, however, the investigation concentrates on the relationship between subjectival position and sentential syntactic complexity. Positions of other constituents might be studied, and other corpora can be employed in further research.

References

Altmann, G. (1980). Prolegomena to Menzerath's law*, Glottometrika* 2, 1–10.

Fan, F., Yu, Y., Wang, H. (2013). Subjectival position and sentential syntactic complexity In R. Köhler, G. Altmann (eds). *Issues in Qunatitative linguistics 3,* Studies in Quantitative Linguistics 13. Ram-verlag, 137-149.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar*. London: Edward Arnold.

Nemcová, E., Serdelová, K. (2005). On Synonymy in Slovak, In: Altmann, G, Levickij, V., Perebyinis, V. (eds.), *Problems of Quantitative Linguistics*. Chernivtsi: Ruta, 194–209.

Tuldava, J. (1980). K voprosu ob analitičeskom vyraženii svjazi meždu ob'emom slovarjai ob'emom teksta. In: *Lingvostatistika i kvantitativnye zakonomernosti teksta.* Tartu: Učenye zapiski Tartuskogo gosudarstvennogo universiteta, 549, 113–144.

Wimmer, G. & Altmann, G. (2005). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the science of language: Word length and related issues*. Boston: Kluwer, 93–117.

Glottometrics 47, 2019, 104-122

A Comparative Study on NP Length, Complexity and Pattern in Spoken and Written English

Yujia Zhu[1]

Abstract. Noun phrase (NP) is considered one of the most important phrasal categories and has received attention from generations of linguists. However, most linguists study NP using the qualitative approach and focus on NP patterns. This paper aims to compare NPs in spoken and written English from three different aspects – length, complexity, and pattern. Both qualitative and quantitative approaches were used in the study. ICE-GB-S and ICE-GB-W were used as the sources. The results show that firstly, the mean of NP length in ICE-GB-S is much shorter than that in ICE-GB-W. Secondly, the optimum mathematical model for the distribution of NP length in ICE-GB-S is $F_l = aL^b$, and the one for ICE-GB-W is $F_l = a + b/L$. Thirdly, the mean of NP complexity is smaller than in ICE-GB-W. Fourthly, the optimum mathematical model for distribution of NP complexity in ICE-GB-S is $F_c = a + b/C$, and the one for ICE-GB-W is $F_c = aC^b$. Fifthly, NPs with a prepositional phrase or a clause as the post-modifier are usually with determiners in both spoken and written English, and prepositional phrases are used more frequently than clauses as post-modifiers. Sixthly, NPs with the complexity of 4 have the greatest number of different patterns in both ICE-GB-S and ICE-GB-W. Finally, the same mathematical model can be used to describe the relationship between complexity and pattern in both ICE-GB-S and ICE-GB-W, just with different parameters only. The mathematical model is $P = a C^b e^{cC} + 1$.

Keywords: *Length, complexity, pattern, corpus, mathematical model*

1 Introduction

Noun phrase (NP) is considered as one of the most important phrasal categories (Quirk, 1985) and has received attention from generations of linguists. However, most linguists study NP using the qualitative approach and focus on NP patterns (Dan & Adrian, 2004). With the development of quantitative linguistics, an increasing number of linguists explore NP from the quantitative perspective.

This paper aims to compare NPs in spoken and written English from three different aspects – length, complexity and pattern. Both qualitative and quantitative approaches were used in the study. ICE-GB-S and ICE-GB-W were used as the data source.
The study is significant in the following three aspects.

Firstly, it would contribute to the comparative study between spoken and written

[1] Correspondence to Yujia Zhu, Lanhai Group (Holding) Co., Ltd., Shanghai, China. Email: zhuyujia@lanhaigroup.com.cn

English by providing a quantitative support for it. Secondly, this study may shed light on the quantitative research on NP. It aims to use quantitative approaches to explore the properties of NP and find mathematical models to describe the distribution of length, complexity, and relation between complexity and patterns. Thirdly, this study would also make contributions to pedagogy. It may help English learners to know more about difference in using NP in between spoken and written English.

2 Methods and Materials

2.1 Corpus Used in the Study

The corpus used in this study is The International Corpus of English (ICE).

Specifically speaking, it is ICE of Great Britain (ICE-GB), a major component of ICE. ICE-GB is a 1-million-word corpus of contemporary British English, which was first released in 1998. Together with the dedicated retrieval software, ICE Corpus Utility Program (ICECUP), ICE-GB is an unprecedented resource for the research and education of English grammar in universities and schools all over the world.

ICE-GB contains both spoken and written British English from the 1990s, including 200 written and 300 spoken texts. Every text in the ICE-GB is grammatically annotated, which makes available the extraction of NPs by computer available. Moreover, ICE-GB includes data of spoken and written English, in ICE-GB-S and ICE-GB-W respectively. The separation of the genre makes the comparison of the NPs of spoken and written English possible. In addition, it has been fully checked by linguists at various stages and strategies – by, for instance, post-checking and cross-sectional error-based searches –, which guarantees the accuracy of the research.

2.2 Tools for Data Collection and Analysis

To explore the NPs in both spoken and written English, two important tools – Perl and SPSS – were used in the study. Perl was used for extracting all NPs efficiently from ICE. SPSS was used to calculate and find the mathematical model that best captures the relationship best.

2.3 Process of Data Collection

In this study, ICE was used as the data source. The data of NPs in spoken English was obtained from ICE-GB-S, and data in written English was gained from ICE-GB-W.

NP length in this study was measured on the basis of the number of words. Complexity was obtained by calculating the immediate components of the NP.

Firstly, with the aid of FoxPro, all NPs were extracted from ICE-GB. In ICE-GB, all the NPs had an NP head symbolized by the tag NPHD, and a post-modifier symbolized by the tag NPPO.

Secondly, with the help of Perl, nothing but the raw texts in the curly bracket were kept to calculate the length of NPs. The whole process was complex because there were always some NPs nested in NPs. Therefore, a loop was used in the programming in order to guarantee the NPs are extracted completely.

Thirdly, the complexity and patterns of NPs were obtained with the aid of Perl, too. For this time, the syntactic tags were kept for this time. The syntactic tags needed in calculation were the tags with one space right behind the beginning of lines with the NP in them. The tags after the comma were also kept because they would be used to obtain the patterns of NPs. Then the total number of spaces in the NP was obtained and the complexity was the number of spaces plus one, which is the immediate constituents of the NP.

Fourthly, with the help of SPSS, descriptive statistics of NP length and complexity were obtained and analysed thoroughly. Moreover, mathematical models were obtained by using regression in SPSS to capture the distribution of NP length and NP complexity in ICE-GB-S and ICE-GB-W respectively. Both the descriptive statistics and mathematical models were compared with each other.

Fifthly, the distributions of NP patterns in ICE-GB-S and ICE-GB-W were obtained and compared. A mathematical model was gained to capture the relationship between the NP complexity and the number of NP patterns with the corresponding complexity.

3 Results and Discussion

3.1 Comparison of NP length between ICE-GB-S and ICE-GB-W

With the help of Perl, lengths of both ICE-GB-S and ICE-GB-W were obtained. There are totally 185,469 NPs and 132,563 NPs in ICE-GB-S and ICE-GB-W, respectively. The lists of them were obtained and compared. Further analysis was done on them by means of Excel and SPSS.

3.1.1 Descriptive Statistics of NP Length

First, before descriptive statistics of NP length were obtained, a test was done first to check whether there was a significant difference between NP length in spoken and written English.

Firstly, a 1-Sample Kolmogorov-Smirnov (K-S) Test was done to check whether the distribution of the data was normal, determining which test should be used to test the existence of a significant difference.

The significance of the 1-Sample K-S test on both ICE-GB-S and ICE-GB-W is 0, smaller than 5 percent, which means that neither of them is normally distributed. Since the two groups of data are numeric and independent, and neither of them is normally

distributed, Mann-Whitney U test was used to examine the existence of the significant difference between them.

It is shown in the result of the Mann-Whitney U test that the absolute value of Z is 79.12, much larger than 1.96, and *p*-value is 0. Both of them mean that there is a significant difference between NP length in ICE-GB-S and ICE-GB-W.

Then, the descriptive statistics of NP length in ICE-GB-S and ICE-GB-W were obtained. There are altogether 185,469 NPs in ICE-GB-S and 132,563 NPs in ICE-GB-W. NP length in ICE-GB-S ranges from 1 to 95, while the one in ICE-GB-W ranges from 1 to 88.

The mean of NP length in ICE-GB-S is only 2.74, much shorter than that in ICE-GB-W, which is 3.44. It means that NPs in spoken English are usually shorter than those in written English. Moreover, the deviation of NP length in ICE-GB-W, which is 4.37, is much larger than the one in ICE-GB-S, which is 3.90. It means that NP length varies more in written English than in spoken English.

3.1.2 Distribution of NP Length

The data of NP length in both ICE-GB-S and ICE-GB-W were aggregated with the help of Perl.

There are altogether 185,469 NPs in ICE-GB-S. It can be calculated that NPs with the length of 1 account for 52.76% of all NPs, with the frequency of 97,856. NPs with the lengths of 1 and 2 together take up 72.47% of all NPs, with a total frequency of 134,409. The frequency of NPs with the length smaller than or equal to 3 is 150,245, accounting for 81.01% of NPs. NPs with the length smaller than or equal to 5 altogether comprise 89.38% of all NPs in ICE-GB-S. It means that most of the NPs in spoken English are short NPs with the length smaller than or equal to 5. The long NPs are rarely used in spoken English.

A figure of the NP length distribution in ICE-GB-S was drawn, as the following shows. Since NPs with the length smaller than or equal to 10 account for 96.13% of all NPs, NPs with the length longer than 10 were excluded from the figure below.

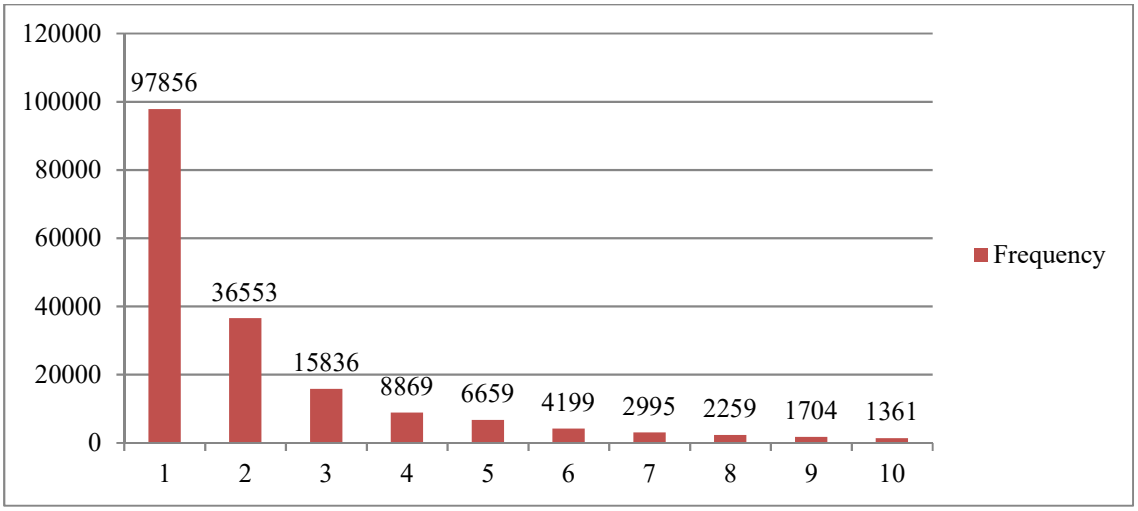

Figure 1. Distribution of NPs with length smaller than or equal to 10 in ICE-GB-S

It is shown in Figure 1 that the frequencies of NPs decrease as the NP length increases. In other words, the shorter the NP is, the more frequently it is used in spoken English. It is the same as the distribution of sentence length in spoken English. According to Li (2009), sentences with one word only are used most frequently in spoken English. The frequencies of sentences decrease as the sentence length increases.

To compare NP lengths between spoken and written English, the distribution of NP length in ICE-GB-W was obtained as well. There are altogether 132,563 NPs in ICE-GB-W. Differently from ICE-GB-S, the frequency of NPs with the length equal to 1 in ICE-GB-W is 51,287, only accounting for 38.69% of all NPs in ICE-GB-W, much smaller than that in ICE-GB-S, where it covers 52.76%. Moreover, the frequency of NPs with the lengths of 1 and 2 altogether compose 62.32% of NPs in ICE-GB-W, which is approximately 10% less than in ICE-GB-S, where it is 72.47%. It is also the same with NPs with the length smaller or equal to 3. NPs with the length of 1, 2, and 3 together make up 72.88% of all NPs in ICE-GB-W, which is much less than those of ICE-GB-S, which form 81.01%. However, the difference is getting increasingly smaller. The frequencies of NPs with the length smaller than or equal to 5 in ICE-GB-W is 111,491, accounting for 84.10% – only about 5% smaller than that of ICE-GB-S, which is 89.38%. The frequencies of NPs smaller than or equal to 10 form 94.09%, approximately 2% smaller than the ones of ICE-GB-S, which make up 96.13%.

In order to analyze the frequencies of NP lengths in ICE-GB-W further, a figure of NPs in ICE-GB-W was drawn, as follows.

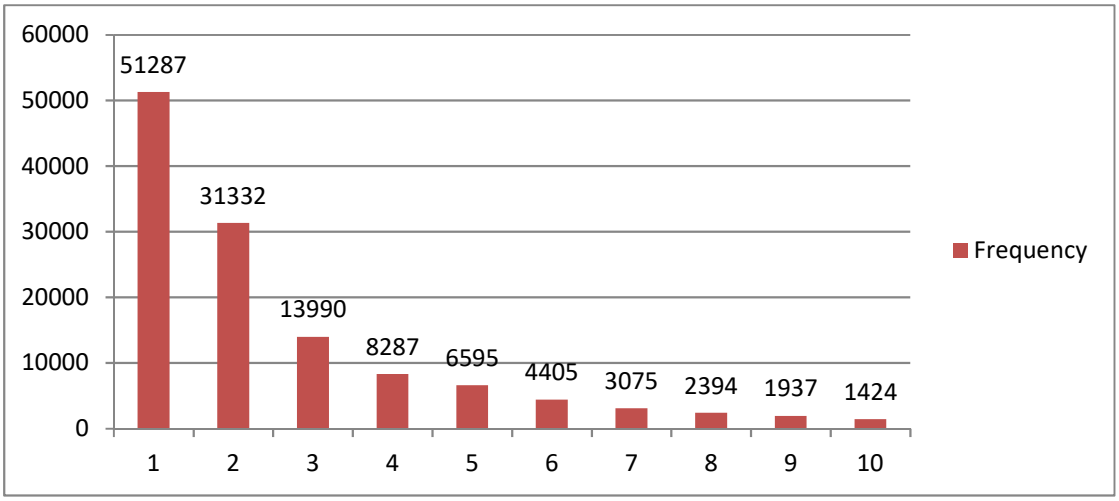

Figure 2. Distribution of NPs with the length smaller than or equal to 10 in ICE-GB-W

It is shown in Figure 2 that the frequency of NPs in ICE-GB-W decreases as the length increases, too.

The figure of NPs with the length shorter than or equal to 10 in both ICE-GB-S and ICE-GB-W was obtained as well. Since the total number of NPs in ICE-GB-S and ICE-GB-W was different, the percentage, instead of the frequency, was used in the figure.

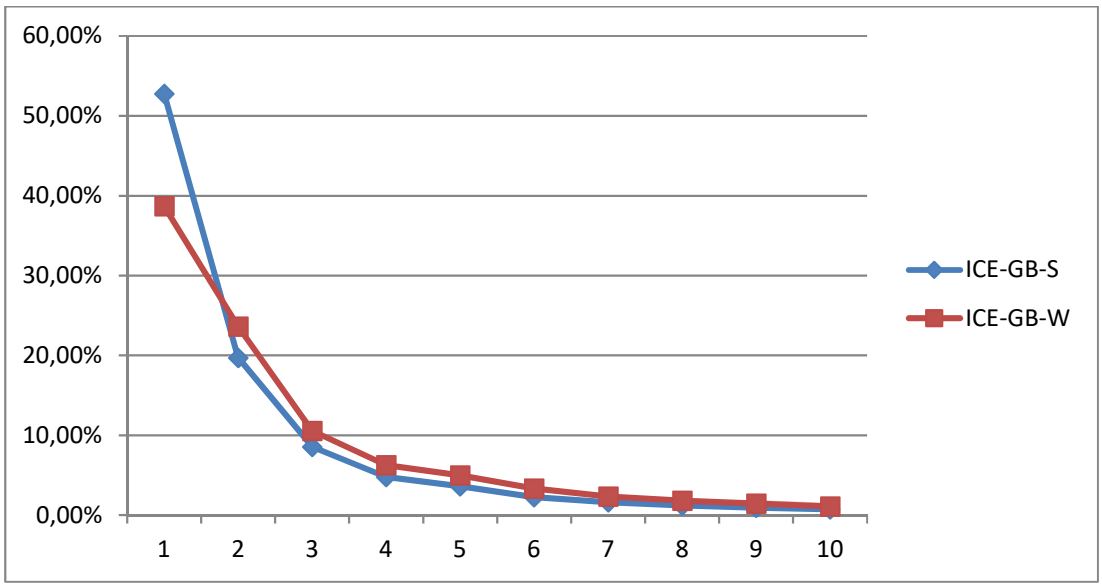

Figure 3. Distributions of NPs with the length smaller than or equal to 10 in both ICE-GB-S and ICE-GB-W

In Figure 3, it can be seen that both the frequencies of NPs in ICE-GB-S and in ICE-GB-W decrease as the NP length increases, but the frequency in ICE-GB-S decreases more rapidly than in ICE-GB-W.

Some possible function models were employed to capture the distribution of NP length in ICE-GB-S. The power model was one of them. It is one of the frequently used mathematical models for nonlinear regression. Its standard form is:

$$y = b_0 t^{b1} \quad (1)$$

The power model can be used to capture the relationship between the NP length (L) and the frequency of NP (F_l) for the corresponding NP length in ICE-GB-S. The power model can be changed into the following:

$$F_l = aL^b \quad (2)$$

R^2 is a number between 0 and 1. It is used to check how the formula fits the data. The larger R^2 is, the better the formula fits the data. R^2 for the power model was 0.956, the greatest value for all function models that have been used in this study. Therefore, the power model is the one that captures the data in ICE-GB-S the best. The two estimated parameters in the power model – a and b – are 833,300 and -2.93, respectively.

The power model was also used to capture the data of NP length in ICE-GB-W. Its R^2 is 0.932. Other possible mathematical models were fitted to the data as well. One of them is the inverse model.

The inverse model is also one of the frequently used mathematical models. Its standard form is as following:

$$y = b_0 + b_1/t \quad (3)$$

The inverse model can be used to capture the relationship between the NP length (L) and the frequency of NP (F_l) for the corresponding NP length in ICE-GB-W. The inverse model can be changed into:

$$F_l = a + b/L \quad (4)$$

R^2 of the inverse model for the distribution of NP length in ICE-GB-W is 0.965, larger than 0.932, which means that the inverse model captures the data in ICE-GB-W better. The two estimated parameters – a and b – are -1,720 and 52,530, respectively. In a word, the relation between NP length and frequency in ICE-GB-W can be captured by $F_l = a + b/L$, in which a = -1,720 and b = 52,530.

Two graphs of both of the predicted and observed values were drawn as following, to show the difference in the distributions of NP length between spoken and written English.

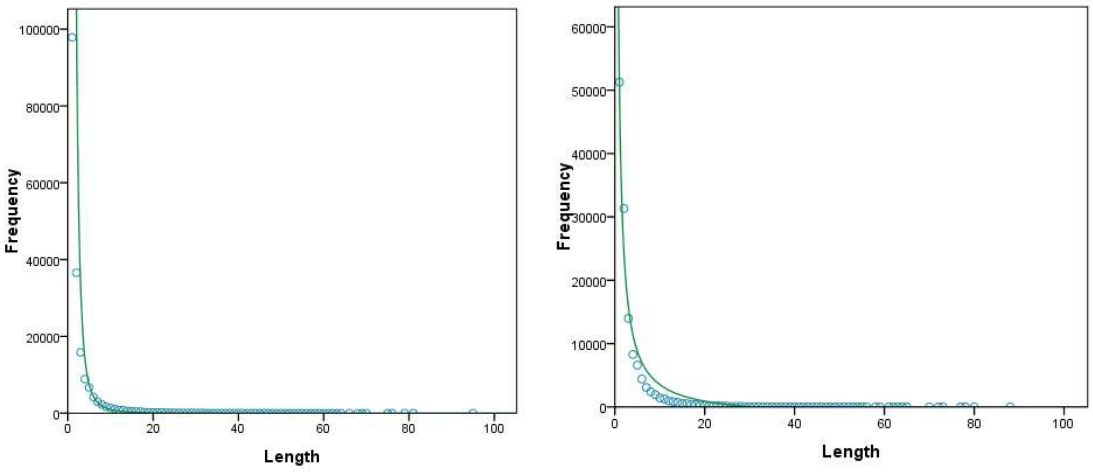

Figure 4. Model fit for the distributions of NP Length in ICE-GB-S (left) and ICE-GB-W (right). Solid line: model fit, small circles: the observed values.

It can be seen from Figure 4 that both of the two mathematical models fit the data well.

In a word, although the frequency of NP length in both ICE-GB-S and ICE-GB-W decreases as the NP length increases, they can be captured best by different mathematical models. The optimum mathematical model for the distribution of NP length in ICE-GB-S is $F_l = aL^b$, in which $a = 833{,}300$ and $b = -2.93$, and the one for ICE-GB-W is $F_l = a + b/L$, in which $a = -1{,}720$ and $b = 52{,}530$.

3.2 Comparison of NP Complexity between ICE-GB-S and ICE-GB-W

The way complexity is calculated in this study has been displayed in detail in 3.1. Its calculation mainly depends on the pattern of each NP. In other words, complexity in this study is a quantification of NP. NP complexity in both ICE-GB-S and ICE-GB-W is analysed in detail in this chapter.

3.2.1 Descriptive Statistics of NP Complexity

Before the descriptive statistics of NP complexities were obtained, a test was done in advance to check whether there was a significant difference between them.

Firstly, 1-Sample K-S test was done on NP complexity in both ICE-GB-S and ICE-GB-W to check the normality of the data. The result shows that the significance of the 1-Sample K-S test on the NP complexities of both ICE-GB-S and ICE-GB-W is 0.00, smaller than 0.05. It means that neither of the NP complexities is normally distributed. Since the two groups of data are not correlated at all, and neither of them is distributed normally, Mann-Whitney U test was used to check whether there is a significant difference between ICE-GB-S and ICE-GB-W in NP complexity.

The result of the test shows that the absolute value of Z is 80.017, larger than 1.96, and the *p*-value is 0.00, smaller than 0.05. Both of them confirm the existence of a significant difference in NP complexity between ICE-GB-S and ICE-GB-W.

Then, descriptive statistics of both ICE-S and ICE-GB-W were obtained with the help of SPSS.

There are altogether 185,469 NPs in ICE-GB-S and 132,563 NPs in ICE-GB-W. NP complexity in ICE-GB-S ranges from 1 to 23, while the one in ICE-GB-W ranges from 1 to 35.

The mean of NP complexity in ICE-GB-W, namely 2.05, is larger than that in ICE-GB-S, namely 1.77, which means that NPs in written English are more complex than those in spoken English.

Moreover, the standard deviation in ICE-GB-S, which is 1.01, is smaller than the one in ICE-GB-W, which is 1.15. It means that NPs in written English vary more in complexity than those in spoken English. It is similar to the situation of NP length analysed in 3.1.2. The deviations in both length and complexity of spoken English are larger than those of written English.

To sum up, there is a significant difference in NP complexity between ICE-GB-S and ICE-GB-W. The mean of NP complexity in ICE-GB-S, namely 1.77, is smaller than the one in ICE-GB-W, namely 2.05. The standard deviation in ICE-GB-S, which is 1.01, is smaller than the one in ICE-GB-W, which is 1.15. It means that NP complexity in spoken English is smaller and varies less than the one in written English.

3.2.2 Distribution of NP Complexity

The data of NP complexity were aggregated for further analysis of its distribution. The aggregated result is displayed in Table 1.

There are altogether 185,469 NPs in ICE-GB-S. The percentage of each complexity was obtained by dividing the frequency of NPs with the corresponding complexity and the total number of NPs. It can be calculated that NPs with the complexity of 1 account for approximately 52.81%, more than half of all NPs. NPs with the complexities of 1 and 2 together make up 78.55% of all NPs, with a total frequency of 145,680. The frequency of NPs with complexity smaller than or equal to 3 sum up to 175,081, which occupies 94.40% of all NPs. NPs with the complexity smaller than or equal to 5 take up 99.49% of all NPs. In other words, NPs with the complexity larger than 5 only account for less than 1% of all NPs. It means that most of the NPs in ICE-GB-S are simple. This is in accordance with native speakers' intuition. Native speakers prefer short and simple NPs while speaking because they are easy to comprehend.

A figure was drawn (Figure 5) to analyse the distribution of NP complexity further. Only those NPs with the complexity smaller than or equal to 10 were included in this graph, for NPs with the complexity larger than 5 take up a very small part of the whole, let alone NPs with the complexity larger than 10.

Table 1
Distribution of NP complexity in ICE-GB-S

Complexity	Frequency	Complexity	Frequency
1	97,955	10	18
2	47,725	11	10
3	29,401	12	8
4	7,609	13	5
5	1,841	15	1
6	534	16	1
7	231	17	1
8	92	23	1
9	36		

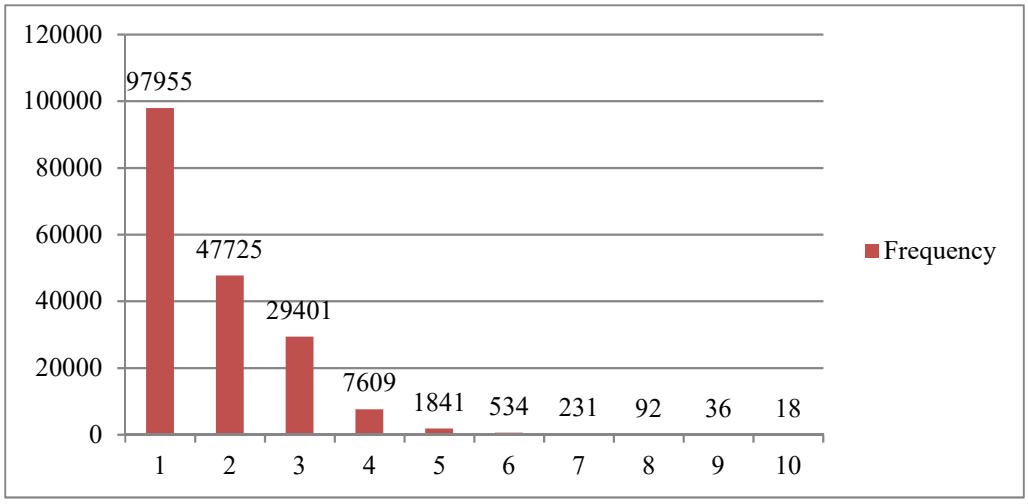

Figure 5. Distribution of NPs with the complexity smaller than or equal to 10 in ICE-GB-S

From the figure above, it is obvious that the frequency decreases continually as the complexity increases. It is similar to the situation of NP length in ICE-GB-S. However, NPs with the complexity larger than 5 are very rare in ICE-GB-S, with a percentage smaller than 1%. Differently from complexity, NPs with the length larger than 5 take up a larger percentage which is more than 10%.

Next, the distribution of NP complexity in ICE-GB-W was obtained and is displayed in Table 2.

There are totally 132,563 NPs in ICE-GB-W. The same calculation was done on the data above to get the percentage of each NP complexity. It can be calculated that NPs with the the complexity of 1 make up only 38.72%, much smaller than the one of ICE-GB-S, which cover 52.81%. It indicates that simple NPs with only one constituent are used much less frequently in written English than in spoken English.

Moreover, NPs with the complexities of 1 and 2 in ICE-GB-W take up 69.55%, which is by 9% smaller than the percentage in ICE-GB-S, namely 78.55%. The frequencies of NPs with the complexity smaller than or equal to 3 add up to 121,108,

which composes approximately 91.36% of all NPs in ICE-GB-W; this is only about 3% smaller than the one of ICE-GB-S, which is 94.40%. NPs with the complexity smaller than or equal to 5 in ICE-GB-W makes up 99.02% of all NPs, similar to the percentage of ICE-GB-S, which is 99.49%. It means that NPs with the complexity larger than 5 account for less than 1% in both ICE-GB-S and ICE-GB-W. The result indicates that very complex NPs (NPs with the complexity larger than 5) are rarely used in both spoken and written English.

Table 2
Distribution of NP complexity in ICE-GB-W

Complexity	Frequency	Complexity	Frequency
1	51,328	12	16
2	40,869	13	11
3	28,911	14	5
4	7,894	15	10
5	2,264	16	4
6	651	17	2
7	313	18	3
8	119	20	1
9	82	21	5
10	40	23	4
11	30	35	1

A figure about the distribution of NP complexity in ICE-GB-W was drawn, as follows. Since NPs with the complexity larger than 10 take up less than 1%, they were excluded from the figure.

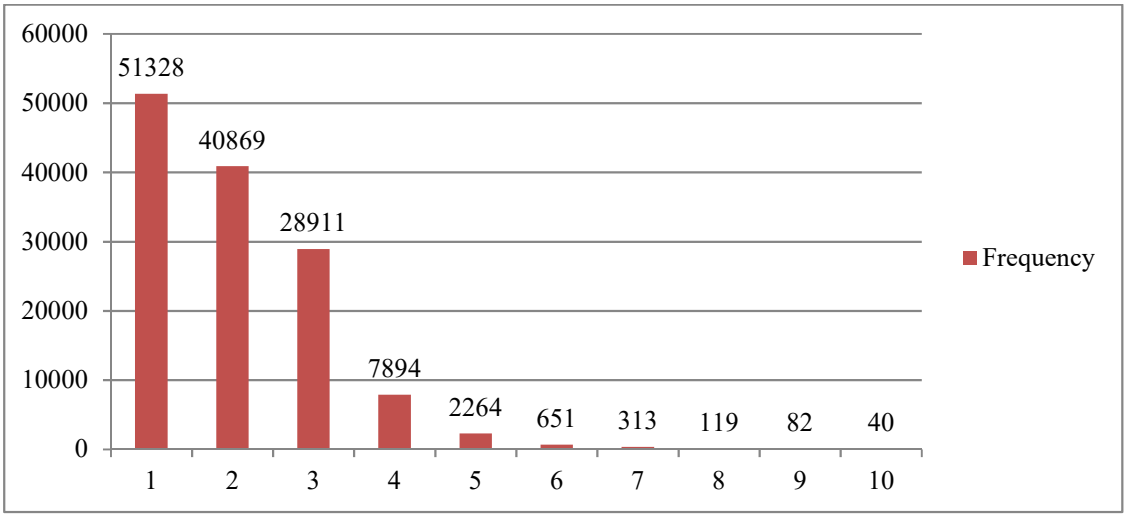

Figure 6. Distribution of NPs with the complexity smaller than or equal to 10 in ICE-GB-W

It shows that the frequency of NP decreases as the complexity increases in both

ICE-GB-S and ICE-GB-W. Then, a line chart of NP complexity in both ICE-GB-S and ICE-GB-W was drawn in Figure 7. Since the total numbers of NPs in ICE-GB-S and ICE-GB-W are different, the percentage was used in the graph.

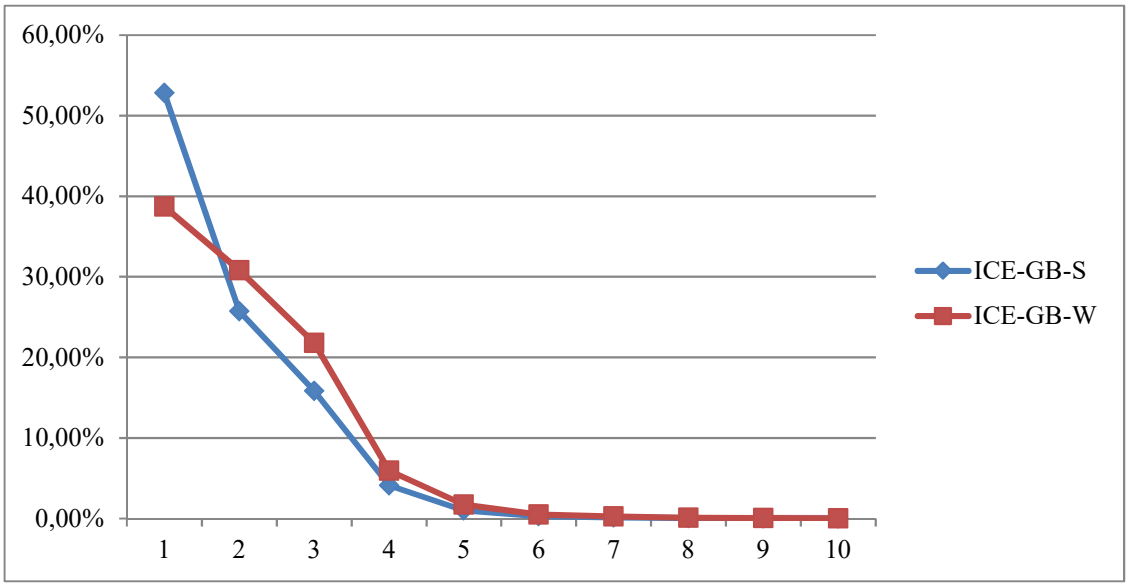

Figure 7. Distributions of NPs with the complexity smaller than or equal to 10 in both ICE-GB-S and ICE-GB-W

It is obvious in the figure above that the percentage of NP complexity in both ICE-GB-S and ICE-GB-W descends as the complexity increases, but at a different speed in each corpus. Generally speaking, NP complexity in ICE-GB-S decreases more rapidly than in ICE-GB-W. In addition, NP with the complexity of 1 plays a more dominant role in ICE-GB-S than in ICE-GB-W.

Then SPSS was used to obtain the mathematical model that fitted the data of NP complexity best in both ICE-GB-S and ICE-GB-W.

Firstly, the nonlinear regression was used on the data of NP complexity obtained from ICE-GB-S. All the possible function models were tried on the data. It turned out that the R^2 of the inverse model is 0.966, the largest in all R^2 of possible function models. It means that the inverse model captures the distribution of NP complexity in ICE-GB-S the best. The two estimated parameters in the inverse model – *a* and *b* – are 10,680 and 107,600, respectively. In a word, the optimum mathematical model for the distribution of NP complexity in ICE-GB-S is $F_c = a + b/C$, in which C denotes the NP complexity, and F_c denotes the frequency of NP for the corresponding NP complexity. In the model, $a = 10{,}680$, $b = 107{,}600$.

Then, the inverse model was also fitted to the NP complexity in ICE-GB-W. Its R^2 is 0.880. However, the R^2 of the power model is 0.932, which means that the power function captures the distribution of NP complexity in ICE-GB-W the best. The two estimated parameters in the power model – *a* and *b* – are 458,500 and -3.96 respectively. In a word, the optimum mathematical model for the relation between NP complexity and frequency in ICE-GB-W is $F_c = aC^b$, in which C denotes the NP complexity, and

F_c denotes the frequency of NP for the corresponding NP complexity. In the model, a = 458,500, b = -3.96.

Two graphs of both the predicted and observed values were drawn.

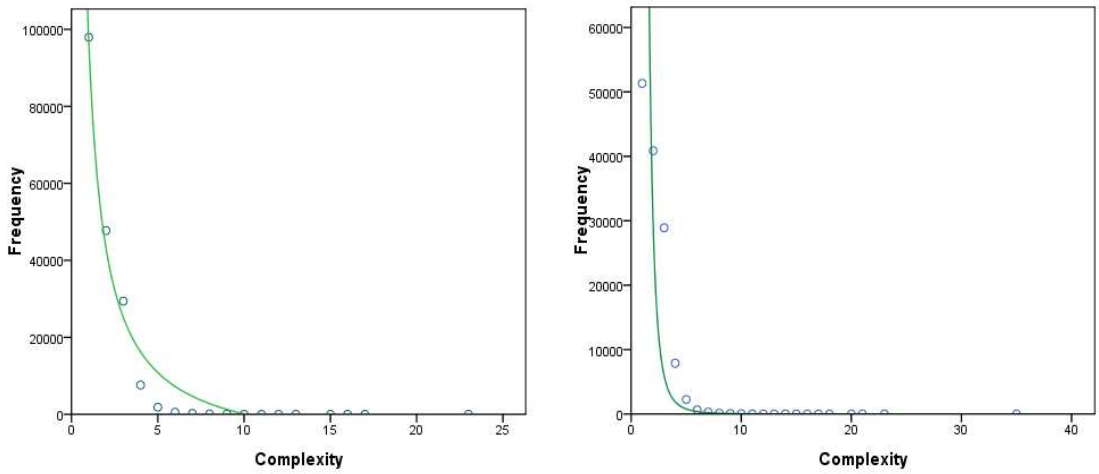

Figure 8. Model fit for the distributions of NP complexity in ICE-GB-S (left) and ICE-GB-W (right). Solid line: model fit, small circles: the observed values.

It can be seen from the two graphs that both the two mathematical models fit the data well.

To sum up, although the frequency of NP complexity in ICE-GB-S and ICE-GB-W share the same descending tendency, it does not necessarily mean that they behave according to the same optimum mathematical models. The optimum mathematical model for the distribution of NP complexity in ICE-GB-S is $F_c = a + b/C$, in which a = 10,680, b = 107,600, and the one for ICE-GB-W is $F_c = aC^b$, in which a = 458,500, b = -3.96.

3.3 Comparison of NP Pattern between ICE-GB-S and ICE-GB-W

Complexity is the quantification of a pattern. Therefore, there is a close relationship between them. In order to find what the relationship exactly is, further study was done on the pattern.

3.3.1 Distribution of NP Patterns in ICE-GB-S and ICE-GB-W

There are altogether 2,525 different NP patterns in ICE-GB-S, and 1,829 different NP patterns in ICE-GB-W. All NP patterns were sorted according to their frequencies from the highest to the lowest. The top 10 NP patterns with the highest frequencies are displayed in Table 3.

It can be seen in the table that among the top 10, three of them only have 1 linguistic constituent, namely NPHD_PRON, NPHD_N, and NPHD_NUM. NPHD_PRON refers to the pattern with only a pronoun as the head. It ranks first on the

list, which means that NPs such as *it*, *he*, etc., are used most frequently in spoken English. NPHD_N refers to the pattern with a noun as a head. It ranks third on the list. NPHD_NUM indicates the NP pattern with only a number as head. It ranks the tenth on the list.

Table 3
Top 10 NP patterns in ICE-GB-S

	Pattern	Frequency
1	NPHD_PRON	69,142
2	DT_DTP NPHD_N	30,414
3	NPHD_N	25,245
4	DT_DTP NPHD_N NPPO_PP	8,343
5	DT_DTP NPPR_AJP NPHD_N	6,452
6	DT_DTP NPHD_N NPPO_CL	3,564
7	NPPR_AJP NPHD_N	3,424
8	CJ_NP COOR_CONJUNC CJ_NP	3,350
9	NPHD_N NPPO_PP	2,405
10	NPHD_NUM	2,293

Moreover, there are three NP patterns with two linguistic components in the top 10 list, which are DT_DTP NPHD_N, NPPR_AJP NPHD_N, and NPHD_N NPPO_PP. DT_DTP NPHD_N refers to the NP pattern composed of a determiner and a noun, such as *the difficulty*. NPPR_AJP NPHD_N means the NP pattern with an adjective as the pre-modifier and a noun as the head, such as *good students*. NPHD_N NPPO_PP refers to the pattern with a noun as the head and a prepositional phrase as the post-modifier. It ranks the ninth on the list. Compared with the NP pattern that ranks fourth, namely DT_DTP NPHD_N NPPO_PP, it lacks a determiner. It can be inferred that NPs with prepositional phrases as post-modifier usually have determiners. The situation is the same in NP patterns with a clause as the post-modifier. Moreover, the NP pattern that ranks sixth, namely DT_DTP NPHD_N NPPO_CL, means the NP pattern with a determiner, a noun, and a clause as the post-modifier. The NP pattern with no determiner, but only a noun and a clause is not among the top 10. It means that NPs with a clause as the post-modifier usually have determiners.

In addition, DT_DTP NPHD_N NPPO_PP ranks fourth on the list, while DT_DTP NPHD_N NPPO_CL ranks sixth on the list. Moreover, NPHD_N NPPO_PP ranks ninth on the list, while NPHD_N NPPO_CL is not on the list of Top 10. It indicates that prepositional phrases are used more frequently as post-modifiers than clauses.

Next, all NP patterns in ICE-GB-W were obtained and aggregated.

As Table 4 shows, differently from NP patterns in ICE-GB-S, NPHD_N ranks first, instead of NPHD_PRON in ICE-GB-W. NPHD_PRON ranks third in the list of NP patterns in ICE-GB-W. It indicates that pronouns are used more frequently to refer to nouns in spoken English.

Table 4
Top 10 NP Patterns in ICE-GB-W

	Pattern	Frequency
1	NPHD_N	24,448
2	DT_DTP NPHD_N	23,987
3	NPHD_PRON	23,331
4	DT_DTP NPHD_N NPPO_PP	8,652
5	DT_DTP NPPR_AJP NPHD_N	6,139
6	NPPR_AJP NPHD_N	5,241
7	CJ_NP COOR_CONJUNC CJ_NP	4,193
8	NPHD_NUM	3,226
9	NPHD_N NPPO_PP	3,073
10	DT_DTP NPHD_N NPPO_CL	2,865

It can be seen in the table above that the top 10 NP patterns in ICE-GB-S – including three NP patterns with one component, three patterns with two components, and four patterns with three components – are the same as those in ICE-GB-W. However, they are in different orders.

It is shown that DT_DTP NPHD_N NPPO_PP ranks fourth and NPHD_N NPPO_PP ranks ninth on the list of ICE-GB-W. It means that NPs with a prepositional phrase as the post-modifier are usually with determiners. It is the same situation as in ICE-GB-S. Moreover, DT_DTP NPHD_N NPPO_CL ranks tenth, while NPHD_N NPPO_CL is not on the list. It means that NPs with a clause as the post-modifier usually have determiners before the nouns, which is the same as in ICE-GB-S. Therefore, NPs with a post-modifier are usually used with determiners in both spoken and written English.

Last but not least, it can be seen that DT_DTP NPHD_N NPPO_PP ranks higher than DT_DTP NPHD_N NPPO_CL on the list of both ICE-GB-S and ICE-GB-W. NPHD_N NPPO_PP ranks ninth on the list, while NPHD_N NPPO_CL is not among the top 10. It indicates that prepositional phrases are used more frequently as post-modifiers than clauses in both spoken and written English.

To sum up, NPHD_PRON ranks first on the list of top 10 NP patterns in ICE-GB-S, but NPHD_N ranks first on the list in ICE-GB-W. NPs with a post-modifier are usually with determiners in both spoken and written English. Prepositional phrases are used more frequently as post-modifiers than clauses in both of them as well.

3.3.2 Relationship between NP Complexity and Pattern

NPs with the same complexity can have different patterns. For instance, both DT_DTP NPHD_N NPPO_PP and DT_DTP NPPR_AJP NPHD_N have the complexity of three, but they are totally different patterns.

As complexity is the quantification of the pattern, firstly, it was calculated how many patterns there were for each complexity.

The data about NP complexity and the number of patterns of corresponding complexity in both ICE-GB-S and ICE-GB-W were amassed to see whether the relationship between NP complexity and pattern was different in spoken and written English. A figure (Figure 9) for both the data in ICE- GB-S and ICE-GB-W was drawn, as follows.

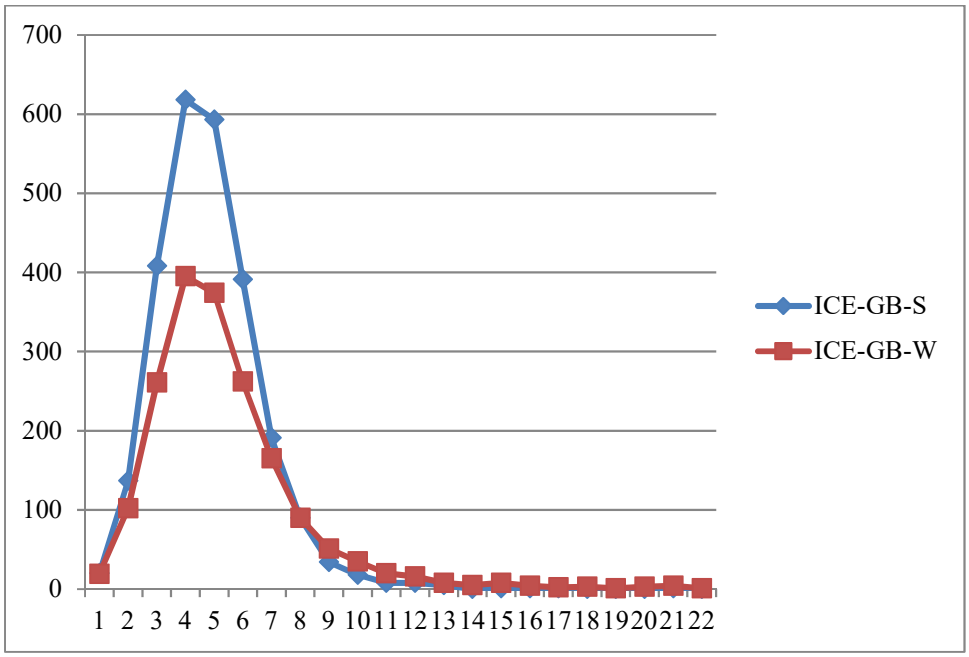

Figure 9. NP complexity and the number of patterns of the corresponding complexity in both ICE-GB-S and ICE-GB-W

In both ICE-GB-S and ICE-GB-W, the number of NP patterns increases continually as the NP complexity increases, reaching the top at the complexity of 4, and then decreases dramatically. It means that NPs with the complexity of 4 have the greatest number of different patterns in both spoken and written English.

In addition, the total number of NP patterns in ICE-GB-S, namely 2,525, is much larger than the one in ICE-GB-W, which is 1,829. Moreover, the maximum in ICE-GB-S, which is 618, is also much larger than the one of ICE-GB-W, which is 395. Generally speaking, the number of patterns in ICE-GB-S is larger than the one in ICE-GB-W, for almost each complexity. It means that English speakers prefer to use more various patterns in spoken English than in written English.

Then, SPSS was used to find the optimum mathematical model for the relation between complexity and pattern in ICE-GB-S and ICE-GB-W respectively. Firstly, the regression model was fitted to the data. However, none of the R^2 for those possible models are over 0.80, which means that none of the possible function models capture the data well. Therefore, some other function models should be found to capture the data.

There is a function that Nemcová and Serdelová (2005) used to describe the relationship between the number of synonyms (y) of a word and the length of the word in syllables (x):

$$y = a\,x^{b}\,e^{c\,x} + 1 \tag{5}$$

Wang (2012) used it to describe the relation between NP length and the number of NP patterns. This relationship may also hold for NP complexity (C) and the number of NP patterns (P) for the corresponding complexity. The function can thus be changed into the following:

$$P = a\,C^{b}e^{cC} + 1 \tag{6}$$

(6) was used as the function model in the regression for the relation between NP complexity and NP pattern. It was tried on both ICE-GB-S and ICE-GB-W.

Firstly, the function was tried to capture the relation between NP complexity and NP pattern in ICE-GB-S. Its R^2 value is 0.998, very close to 1, which means that the mathematical model captures the data well. The three estimated parameters – a, b, c – are 19.94, 7.70, and -1.80, respectively.

Then, the same function was tried on the data in ICE-GB-W. Its R^2 is 0.997, also very close to 1, which means that the mathematical model captures the data in ICE-GB-W well. The three estimated parameters – a, b, c – are 22.95, 6.14, and -1.42, respectively.

Therefore, the same mathematical model captures data in both ICE-GB-S and ICE-GB-W well, just with different parameters.

In order to show clearly how the mathematical model fits the data, all the observed values and the predicted values were drawn in a graph. The result is displayed in the following figures.

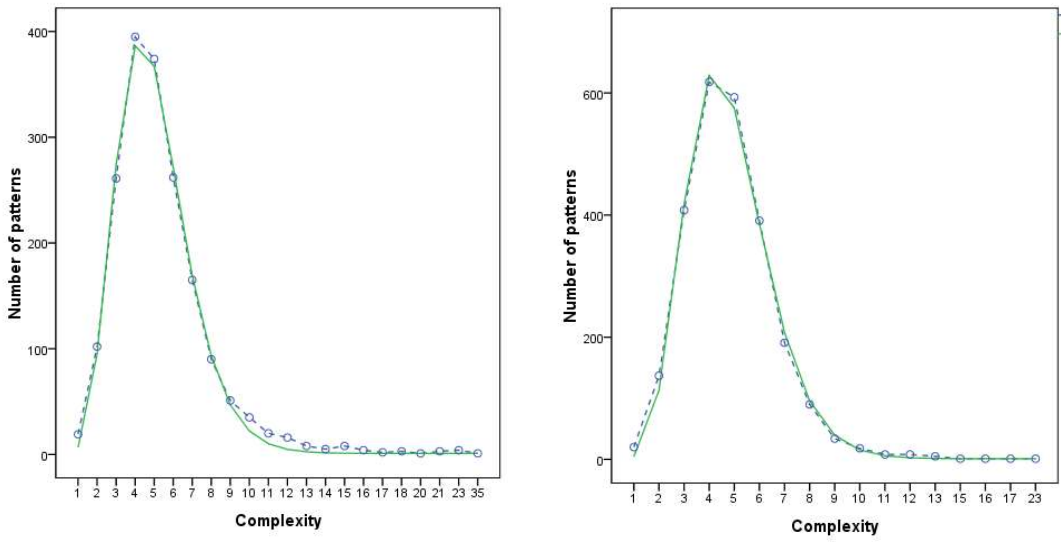

Figure 10. Model fit for the relation between NP complexity and the number of patterns of corresponding complexity in ICE-GB-S (left) and ICE-GB-W (right). Solid line: model fit, small circles: the observed values.

It can be seen that the mathematical model fits both of them well.

To sum up, the numbers of patterns in both ICE-GB-S and ICE-GB-W share the similar tendency. The number of NP patterns reaches its maximum at the complexity of 4 in both ICE-GB-S and ICE-GB-W. The relation between NP complexity and the number of NP patterns with the corresponding NP complexity in ICE-GB-S and ICE-GB-W can be captured with the same mathematical model, which is $P = aC^b e^{cC} + 1$. In ICE-GB-S, a = 19.94, b = 7.70, c = -1.80. In ICE-GB-W, a = 22.95, b = 6.14, c = -1.42.

4 Conclusion

This study aims to compare NP length, complexity, and pattern between spoken and written English. The major findings are summarized as following.

There is a significant difference in NP length between spoken and written English. The mean of the NP length in ICE-GB-S is only 2.74, much shorter than the one in ICE-GB-W, which is 3.44. The standard deviation of the NP length in ICE-GB-S is 3.90, also smaller than the one in ICE-GB-W, which is 4.37. The frequency of NPs decreases as the NP length increases in both ICE-GB-S and ICE-GB-W. However, the frequency of NP length in ICE-GB-S decreases more quickly than the one in ICE-GB-W. The optimum mathematical model for the distribution of NP length in ICE-GB-S is $F_l = aL^b$, in which a = 833,300 and b = -2.93, and the one for ICE-GB-W is $F_l = a + b/L$, in which a = -1,720 and b = 52,530.

There is a significant difference in NP complexity between ICE-GB-S and ICE-GB-W. The mean of NP complexity in ICE-GB-S, namely 1.77, is smaller than the one in ICE-GB-W, namely 2.05. The standard deviation in ICE-GB-S, which is 1.01, is smaller than the one in ICE-GB-W, which is 1.15. In ICE-GB-S, NPs with the complexity larger than 5 are very rare, with a percentage smaller than 1%. However, NPs with the length larger than 5 take up a larger percentage, which is more than 10%. NPs with the complexity of 1 take up only 52.81% in ICE-GB-S, much larger than the one of ICE-GB-W, which is 38.72%. Generally speaking, the frequency of NP in ICE-GB-S decreases more rapidly than in ICE-GB-W, as the complexity increases. The optimum mathematical model for the distribution of NP complexity in ICE-GB-S is $F_c = a + b/C$, in which a = 10,680, b = 107,600, and the one for ICE-GB-W is $F_c = aC^b$, in which a = 458,500, b = -3.96.

NPs with only one pronoun rank first on the NP pattern list of ICE-GB-S. However, NPs that rank first on the list of ICE-GB-W are the NPs with a noun only. It can be inferred that pronouns are used more frequently to refer to nouns in spoken English than in written English. DT_DTP NPHD_N NPPO_PP ranks fourth on the list of ICE-GB-S, while NPHD_N NPPO_PP ranks ninth. Moreover, the former also ranks higher than the latter on the list of ICE-GB-W. It means that NPs with a prepositional phrase as the post-modifier usually have determiners in both spoken and written English. DT_DTP NPHD_N NPPO_CL ranks sixth and tenth on the list of ICE-GB-S and ICE-GB-W respectively, but NPHD_N NPPO_CL is not among the top 10 of both lists. It indicates that, just like in case of the NPs with a prepositional phrase as the post-

modifier, NPs with a clause as the post-modifier usually have determiners in both spoken and written English. DT_DTP NPHD_N NPPO_PP ranks higher than DT_DTP NPHD_N NPPO_CL on the list of both ICE-GB-S and ICE-GB-W. Moreover, NPHD_N NPPO_PP is on the top 10 list, but NPHD_N NPPO_CL is not. Both of them indicate that prepositional phrases are used more frequently as post-modifiers than clauses in both spoken and written English.

Generally speaking, in the figure of NP complexity and the number of patterns of the corresponding complexity in both ICE-GB-S and ICE-GB-W, it can be seen that the number of patterns in ICE-GB-S is larger than that in ICE-GB-W for almost each complexity. It means that English speakers prefer to use more various patterns in spoken English than in written English. NPs with the complexity of 4 have the greatest number of different patterns in both ICE-GB-S and ICE-GB-W. It means that NPs with 4 components have most different patterns in both spoken and written English. The same mathematical model can be used to describe the relationship between complexity and pattern in both ICE-GB-S and ICE-GB-W, just with different parameters. The mathematical model is $P = a\ C^{b} e^{cC} + 1$. In ICE-GB-S, $a = 19.94$, $b = 7.70$, $c = -1.80$. In ICE-GB-W, $a = 22.95$, $b = 6.14$, $c = -1.42$.

References

Dan, M. & Adrian, N. (2004). Word sense disambiguation of WordNet glosses. *Computer Speech & Language*, 18(3), 301–317.

Li, B. (2009). *A Corpus-based Study on Patterns of Spoken English*. Unpublished master's thesis, Dalian Maritime University, Dalian, China.

Nemcová, E. & Serdelová, K. (2005). On synonymy in Slovak. In Altmann, G., Levickij, V., & Perebyinis, V. (eds.), *Problems of Quantitative Linguistics: A Collection of Papers*. Chernivtsi: Ruta, 194–209.

Quirk, R. (1985). *Comprehensive grammar of the English language*. London: Longman.

Wang, H. (2012). Length and complexity of NPs in written English. *Glottometrics* 24, 79–87.

Glottometrics 47, 2019, 123-137

Computational Stylistic Characteristics of American English

Fangfang Zhang[1]

Abstract. It is one of the most fashionable trends to learn American English in China, and American English is always the hot area for scholars to do research in. However, most of the studies are oriented to qualitative rather than quantitative aspects, which indicates the lack of quantitative research in this field. This study adopted the corpus-based approach to study the stylistic characteristics of American English. The Open American National Corpus (OANC) with the size of 18 million words was compared with a set of samples named BNCS, with the similar corpus size. The BNCS was drawn randomly from the 100-million-word British National Corpus (BNC). The comparison was made in order to reveal the stylistic characteristics of American English as to the aspects of word length, TTR, high frequency vocabulary, and sentence length. This study was carried out by the guidance of modern stylistics. With the aid of a computer programme, the tasks of data collection and calculation were carried out, while all the data on the four aspects were carefully studied and analysed with the help of statistical software SPSS. The results show that the word length of American English is longer than the one of British English, and the TTR is larger. Concerning the aspect of high frequency words, although the percentage of function words in the top 100 words is similar in the two corpora, it turns out that the sum frequency of the top 100 words in American English is smaller than the one in British English. In addition, American English displays shorter sentence length than British English.

Keywords: *word length*, *sentence length*, *computational stylistics, American English*

1 Introduction

In recent years, the "American English Craze" has accelerated in China. American English Craze takes root in the history of American English, and comes into shape within a certain time and in certain circumstances. In such a situation, it is necessary to have a better knowledge about the American English Craze and the history of American English before we do research on its characteristics.

The present study aims to give a detailed quantitative stylistic study of American English, adopting the lexical and grammatical criteria, and to identify and distinguish the stylistic features of American English with the corpus-based approach. This paper aims to study the following issues:

[1] Correspondence to Fangfang Zhang, Foreign Languages Department, Harbin University of Science and Technology Rongcheng Campus, Rongcheng, China. Email: fiona20079@foxmail.com

(1) the stylistic characteristics of American English as to the lexical criterion – i.e., word length, vocabulary richness, common word frequency, content and function words, and the differences from British English;

(2) the stylistic characteristics of American English as to the grammatical criterion – i.e., sentence length, and the distinctions from British English.

This study uses stylistic markers to study the characteristics of American English. It attempts not only to depict the whole picture of the stylistics characterizing American English, but also to pursue possible factors underlying these characteristics. This will help learners have a clear view of the linguistic characteristics of American English and be more at ease when they learn American English. Consequently, they can be purposeful in their learning and know what should be practiced more. They will know what they should pay attention to when they communicate with Americans, or when they write in American English.

2 Methods and Materials

2.1 Data source

The corpora employed in this study include Open American National Corpus (OANC) and British National Corpus (BNC). The following section gives a brief introduction into the two corpora.

2.1.1 Introduction to Open American National Corpus

The American National Corpus (ANC) consists of over 11 million words of American English. The ANC project has been creating a massive electronic collection of American English, including texts of all genres and transcripts of spoken data produced from 1990 onwards. The Second Release contains over 22 million words of written and spoken American English, annotated for lemmas, parts of speech, noun chunks, and verb chunks. The Open American National Corpus (OANC) project includes over 18 million words from the Second Release that can be freely downloaded from the internet.

The OANC contains 8,832 texts distributed into 6,424 written files and 2,410 spoken sources. The written component comes from a wide range of domains: government, technical issues, travel guides, fiction, letters, non-fiction, journals, and so on. The written part includes extracts from 911 reports, biomed, eggan, ICIC, OUP, plos, slate, berlitz, verbatim, and web data.

For the spoken part of the OANC, there are two major sources of texts. The first type of spoken texts is the face-to-face material, which consists of 93 files. The second type of spoken texts consists of 2,307 files, which are made through telephone conversations. The structure of OANC is illuminated in the following:

```
\---data
    +---spoken
    |   +---academic-discourse
    |   |   \---micase
    |   +---face-to-face
    |   |   \---charlotte
    |   \---telephone
    |       +---callhome
    |       \---switchboard
    +---written_1
    |   +---fiction
    |   |   +---eggan
    |   |   \---hargrave
    |   +---journal
    |   |   +---slate
    |   |   \---verbatim
    |   +---leisure
    |   |   \---blog
    |   \---letters
    |       \---icic
    \---written_2
        +---newspapers
        |   \---nytimes
        +---non-fiction
        |   \---OUP
        +---technical
        |   +---911report
        |   +---biomed
        |   +---government
        |   \---plos
        \---travel_guides
            +---berlitz1
            \---berlitz2
```

The OANC data include annotations for word and sentence boundaries, part of speech (4 tagsets), and noun and verb chunks. Parts of the corpus are annotated for additional linguistic features. The written portion of the ANC has been tagged for parts of speech using the C5 tagset and the C7 tagset by the University of Lancaster.

2.1.2 Introduction to the British National Corpus

The British National Corpus (BNC) is a 100 million-word collection of samples of

written and spoken English from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century.

The 4,124 texts in the BNC come from written (90 percent) and spoken (10 percent) sources. The BNC written component consists of about 75% informative prose, all after 1975, and about 25% imaginative fiction, all after 1960. The proportion of imaginative text samples was set to 25% in consideration of the continuing influential cultural role of literature and creative writing. The informative prose section comes from a wide range of fields: natural and pure science, applied science, social and community, world affairs, commerce and finance, arts, beliefs and thoughts, and leisure.

For the 10 percent spoken part of the BNC, there are two major sources of texts (Crowdy, 1993). The first type of the spoken texts is the context-governed material, consisting of educational and informative events in lectures, tutorials, and classrooms, news reports, business events, official and public events, leisure events, and so on. The second type of the spoken texts consists of 2,000 hours of transcribed recordings made by 124 volunteers from all walks of life, recruited systematically from 38 different parts of the UK from four different socio-economic groupings and with a balanced coverage of male and female speakers of a wide range of ages between 15 and 60+.

The codes in the BNCS provides a wide range of important information, such as the boundary and part of speech of each word, speech turns, pausing, paragraphs, sections and headings, meta-textual information, and so on.

2.2 Instruments Adopted in This Study

The quantitative analysis is based on the theories of statistics, and according to these theories, a proper type of test should be chosen with enough scrutiny. Just like as Bulter claimed, “Whenever we wish to collect quantitative data on language, we need to pay careful attention to the design of our study, and to the selection of appropriate statistical methods for summarizing the data, and for testing the hypothesis concerning differences between sets of data” (Butler, 1985, ix). In this section, we introduce the instruments used in this study and the research procedure.

2.2.1 Tools for Data Processing

In the study, the programming language FoxPro and statistical software SPSS were applied to obtain the necessary data.

Moreover, Visual FoxPro 6.0 facilitates the analysis of the huge amount of data contained in the two corpora. First, the original data were separated into sentences by a FoxPro programme. Other programmes were used to calculate word length, sentence length, and vocabulary growth.

SPSS has a number of ways to summarize and display data in the forms of tables and graphs. Here, it was used to carry out t-tests and draw the graphs manifesting the differences between American English and British English in sentence length,

vocabulary growth, word length, and so on.

2.2.2 Statistical Test for This Study

T-test is used for testing significance of differences between means of two independent samples. It makes two assumptions about the distributions of the populations from which the samples are drawn – that they are approximately normal, and that they have approximately equal variances. However, it has in fact been shown that the t-test is fairly robust, in the way it is tolerant of all but rather large deviations from normality and equality of variance. The test assumes that the populations have the same variance, so the best estimate of this common variance can be obtained by pooling the variability of the two samples. If the calculated value of *t* is greater than or equal to the critical value as determined by the table given in *Statistics in Linguistics* by Christopher Butler (1985), the null hypothesis can be rejected. If the calculated value of *t* is smaller than the critical value, the null hypothesis cannot be rejected, and the conclusion can be made that there is a significant difference between the means.

2.3 Process of Data Collection

Data collection is the crucial part of this study. With the aid of SPSS and FoxPro, the procedures of data processing were carried out by the steps described in the following section.

As presented above, OANC includes over 18 million words, while BNC contains over 100 million words. To make the data obtained from the two corpora comparable, the first step is to select a set of sample texts from the BNC randomly and rebuild a corpus with a size similar to OANC. By running a FoxPro programme, the sampled BNC, which is named as BNCS hereafter, was built up.

The second step is to remove all the coding of the sample texts in the two corpora, leaving only the clean texts to make lexical analysis. By running a FoxPro programme, the sample texts of OANC were decoded.

The third step is to access individual words of each corpus. The distribution of vocabulary of texts is explored for each corpus. FoxPro programme was run in OANC and BNCS to perform tokenization, lemmatization to get word frequencies, and then displayed the TTR, sentence length, and word length of the two corpora. The data outputs by the programme were further analysed and processed with SPSS, during which the histogram and t-test were performed. The data and figures were presented in Section 3 for a further discussion.

The fourth step is intended to discover the difference between OANC and BNCS with the aid of FoxPro programme. Specifically speaking, we make comparisons of TTR, word length, sentence length, and the 100 most frequent words in the two corpora by using t-test and histogram, which will be presented with details in Section 3.

3 Results and Discussion

This part is intended to present the findings of this research and to provide corresponding interpretations of the results. It mainly comprises the following sections: the analysis of word length, vocabulary growth and common words frequency on the basis of the lexical criterion, and the analysis of sentence length on the syntactic level within the grammatical criterion. The data obtained by FoxPro will be analysed in SPSS, and then the research results will be discussed and summed up.

3.1 Comparison of word length

The word is a central element for any language. The analysis of word length is an important factor to determine the style of American English. The word length in this study was measured in syllables. By running a FoxPro programme, the data of the word lengths in syllables are obtained and listed in Figure 1. Superficially, there seem to be no significant differences between the two corpora.

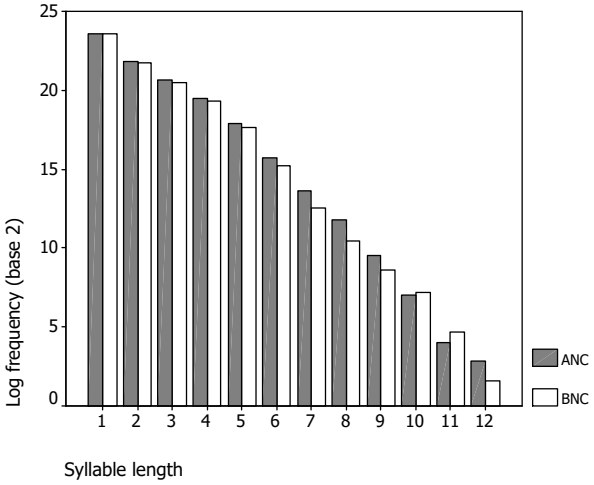

Figure 1. Syllable length bar chart (logarithmized frequencies; base 2)

The mean word lengths of OANC and BNCS are listed in Table 1. As shown in Table 1, the mean word lengths in each corpus are 2.76 and 2.71, respectively, which means the words in OANC are longer than those in BNCS.

After the statistic description of the word length in the two corpora, we need to perform a statistical test to check whether there exists a significant difference between OANC and BNCS, as the superficial statistics are not very reliable in revealing significant differences. In this study, t-test was adopted to solve this problem.

As shown in Table 2, the significance level (2-tailed) is 0.000, which is by far smaller than 0.05. In that case, the conclusion can be drawn that there is a significant difference in the aspect of word length between OANC and BNCS. In other words, the word length of OANC is longer than the one of BNCS.

Table 1
Average word syllabic lengths of OANC and BNCS

	CORP	N	Mean	Std. Deviation	Std. Error Mean
SY_LENGTH	1	275,357	2.76	1.548	.003
	2	205,573	2.71	1.443	.003

Note: Group 1 represents OANC; Group 2 represents BNCS

Table 2
Independent Sample Test on Word Syllabic Length

	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
SY_LENGTH	assumed	1300.921	.000	11.902	480,928	.000
	not assumed			12.024	458,096.309	.000

3.2 Comparison of vocabulary growth

People differ systematically in the richness of their vocabularies. Some have large vocabulary and use many relatively infrequent words, and others have smaller vocabulary and use many more frequent words. This has led to the reasonable assumption, often unstated, that vocabulary richness provides a kind of fingerprint that can distinguish one from another. Whatever the methods of calculation on vocabulary richness are, the appropriate one can capture something distinctive about some people. In the same way people are influenced by vocabulary richness, language is also greatly affected by it. In this research, we will discuss vocabulary richness through one important marker — vocabulary growth rate.

According to Fan (2006), in the inter-textual type-token relationship, the number of types is a function of the number of inter-textual cumulative tokens, with an ever decreasing TTR as the number of tokens increase. The number of types increases rapidly at the beginning; then, the rate of increase slows down. However, the type growth curve is still on the rise as it reaches the end.

The growth rates in OANC and BNCS are shown in Figure 2. In Figure 2, the *x*-axis shows the numbers of corpus size, while the *y*-axis shows the size of vocabulary. The solid line represents the growth rate in OANC, while the dashed one represents the growth in BNCS.

In Figure 2, it can be seen that the two curves have some superposition initially. Generally speaking, both the curves still rise as they reach the end. The number of types increases rapidly at the beginning and then the rate of increase slows down, which corresponds to Fan's theory. As Figure 2 shows, the number of types of OANC is larger

than that number in BNCS. Besides, the growth curve of OANC is steeper than the one of BNCS. Based on all the information given above, it can be said that American English has a larger TTR than British English. According to Williams (1970), the higher TTR indicates the tendency of a wider vocabulary diversity and higher information load. Thus, the result indicates American English has the tendency to a wider vocabulary diversity and a higher information load.

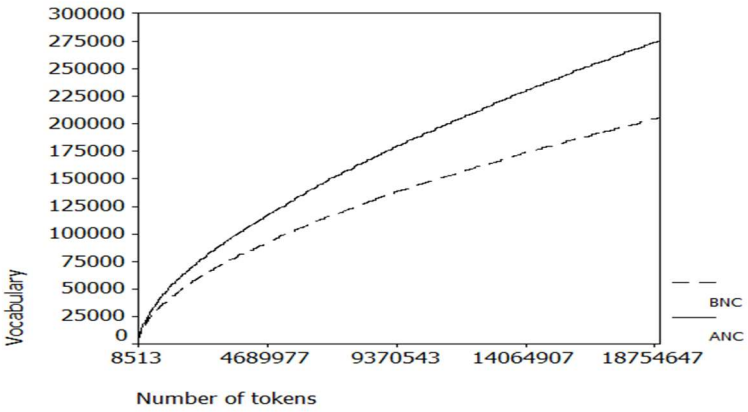

Figure 2. Vocabulary Growth: OANC and BNCS

3.3 Comparison of top 100 words

Word frequency distribution is one of the most important markers in computational stylistic analysis. In this research, we study the top 100 words of OANC and BNCS, respectively.

By running a FoxPro programme, the basic data about the top 100 words and their frequencies in OANC and BNCS were obtained. It can be seen in the Appendix that the first three most frequent words in OANC and BNCS are the same, and they are "the", "be", and "of". However, the respective frequencies of the three words in OANC are smaller than the ones in BNCS. In addition, the sum frequency of the top 100 words is 8,822,607 in OANC, while it is 9,380,376 in BNCS, which means the sum frequency of the top 100 words is much smaller than the one in BNCS.

A well-known index of lexical concentration was proposed by Guiraud (1954):

$$C = (\sum_{50}^{1} F)/N \quad (1)$$

In formula (1), this index represents the portion of cumulative frequency of the most frequent 50 meaningful words in the given text (*F* represents the frequency of a certain meaningful word and *N* represents the text size). A high value of the index indicates that the author concentrates his/her attention on a relatively small range of meaningful words used in the given text. This can testify to the thematic compactness and to the concentration on the main theme. For example, a comparison of 7 fiction texts has

demonstrated great differences, and the values of the index vary from 0.10 to 0.057 – i.e., the most frequent meaningful words from the vocabulary of the text cover 10% of the text, whereas in another text, the 50 most frequent meaningful words cover only 5.7% of the text; that is, the degree of concentration of autosemantic lexis is almost half as small (Tuldava, 1995).

Then, we will have a discussion about the index of lexical concentration in the two corpora. The results are shown in Table 3.

Table 3

Index of lexical concentration of the top 100 words

	OANC	BNCS
Corpus size	18,878,716	18,914,498
Sum Freq. of T_{100}	8,822,607	9,380,376
Index	0.4673	0.4959

It can be seen from Table 3 that the index of lexical concentration of the top 100 words in OANC is smaller than the one of BNCS. The data shown in Appendix indicate that a relatively narrow range of words with full meanings are used in the given text very often in British English. Oppositely, a relatively wide range of words is used in American English. This result may indicate that in comparison with British English, in American English, there are more rare or peculiar words, and less repetition of words.

3.4 Comparison of sentence length

To figure out the stylistic characteristics on the syntactic level, the sentence lengths of OANC and BNCS will be discussed in the following section.

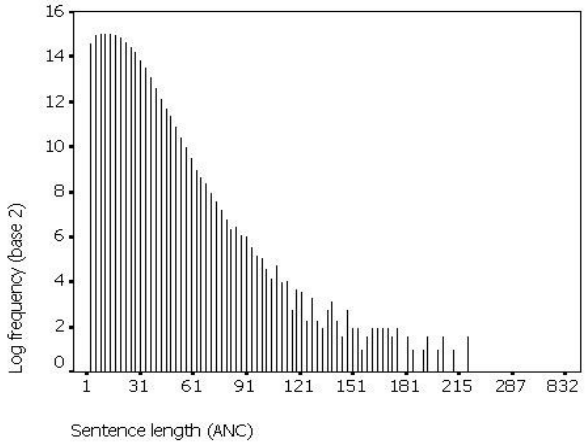

Figure 3. Sentence length bar chart of OANC

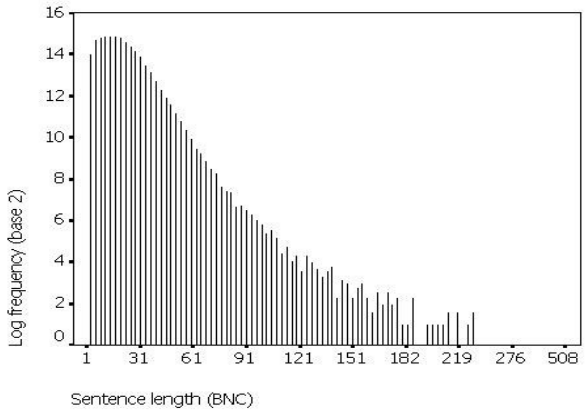

Figure 4. Sentence length bar chart of BNCS

According to the two figures, it can be seen that on the whole, the number of sentences decreases with the growth of sentence length both in OANC and BNCS. In addition, the sentences with shorter lengths occupy larger percentages than those with longer lengths.

The mean sentence length of OANC and BNCS is listed in Table 4. As shown in Table 4, the mean sentence length of OANC is 20.27732, while the mean sentence length of BNCS is 22.26215. It can be seen that the sentences in OANC are, on average, shorter than the ones in BNCS.

Table 4

Average sentence lengths of OANC and BNCS

	CORP	N	Mean	Std. Deviation	Std. Error Mean
SY_LENGTH	1	4,480	20.27732	5.2730691	.0787815
	2	4,480	22.26215	6.2190747	.0929152

Note: Group 1 represents OANC; Group 2 represents BNCS

After the statistic description of the sentence length in the two corpora, we need to perform a statistic test to check whether there exists a significant difference between OANC and BNCS. The t-test was performed to solve this problem.

As shown in Table 5, the significance level (2-tailed) is 0.000, which is by far smaller than 0.05. Based on test result, the conclusion can be drawn that there is a significant difference between the sentence length of OANC and BNCS. In other words, OANC has shorter sentence length than BNCS. According to Leech and Short (1981), the longer the sentence is, the more difficult its sentence structures are, and the more formal the variety is. Therefore, the result can indicate that the sentence structure of American English is less complex than the one of British English.

Table 5
Independent sample test on sentence length

	Equal variances	F	Sig.	t	df	Sig. (2-tailed)
SY_LENGTH	assumed	75.426	.000	-16.293	8958	.000
	not assumed			-16.293	8724.692	.000

4 Conclusion

In this corpus-based research, through comparing the data collected respectively from the two corpora – OANC and BNCS –, we discovered that there are significant differences on the basis of lexical and grammatical criteria between American English and British English.

Firstly, the word syllabic length in OANC and BNCS has some significant differences according to the analysis made before. American English has a longer word syllabic length than British English.

Secondly, although both BNCS and OANC show rising TTRs, American English has a larger TTR than British English. The result indicates American English has the tendency to a wider vocabulary diversity and a higher information load.

Thirdly, American English has a higher index of lexical concentration than British English, which probably suggests American English uses a relatively wider range of words than British English. In addition, the research shows function words are used most frequently in both American English and British English.

Fourthly, although sentences with short lengths form the major part in both OANC and BNCS, the analysis shows that American English has a shorter sentence length than British English. The result can indicate that the sentence structure of American English is less complex than the one of British English.

According to this study, the stylistic characteristics of American English can be obtained. This paper may provide the reader with an interesting and coherent account of lexical and syntactical differences between American English and British English. It may shed light on teaching and learning the standard American English in the new decade.

A better understanding of the stylistic characteristics of American English can help learners in the process of language education. Besides, the high frequency words are important in vocabulary learning. The lists of the top 100 words in this paper can offer learners a clear picture about what is used most frequently by native speakers.

The paper can be helpful for teachers, too. Sinclair and Renouf (1998) propose that students should be taught the words that are mostly used by native speakers in the contexts in which they are most likely to occur. Thus, English teachers may make full use of the top high frequency words in this study, and try to increase the students' exposure to them.

References

Butler, C. (1985). *Statistics in Linguistics.* New York: Basil Blackwell Ltd.

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing* 8(4), 259–265.

Fan, F. (2006). A corpus-based study on inter-textual vocabulary growth. *Journal of Quatitative Linguistics* 13, 111–127.

Guiraud, P. (1954*). Les Caractères statistiques du vocabulaire: essai de méthodologie.* Paris: Presses Universitaires de France.

Kennedy, G. (2000). *An Introduction to Corpus Linguistics*. Beijing: Foreign Language Teaching and Research Press.

Leech, G. N., & Short, M. H. (1981). *Style in Fiction: A linguistic Introduction to English Fictional Prose*. London: Longman Froup Ltd.

Sinclair, J. & Renouf, A. (1998). *Vocabulary and Language Teaching*. London: Longman.

Tuldava, J. (1995). *Methods in Quantitative Linguistics*. Trier: WVT.

Williams, C. B. (1970). *Style and Vocabulary: Numerical Studies*. London: Griffin.

Appendix

The Top 100 words in OANC and BNCS

	OANC	Frequency	BNCS	Frequency
1	The	1,110,794	The	1,238,794
2	Be	718,572	Be	746,162
3	Of	549,309	Of	640,419
4	A	474,085	To	523,042
5	And	473,528	And	511,675
6	To	459,382	A	497,103
7	In	367,594	In	397,534
8	That	242,879	Have	237,727
9	I	206,715	That	200,003
10	Have	198,790	For	177,486
11	For	184,314	It	174,078
12	It	157,840	On	139,741
13	Not	147,454	As	137,083
14	With	136,735	With	133,102
15	On	125,281	Not	127,489
16	As	114,991	He	120,399
17	He	104,122	I	117,131
18	Do	97,787	Will	110,159
19	By	96,854	By	109,642
20	At	87,951	At	101,792
21	But	83,712	This	87,203
22	They	81,882	From	86,844
23	From	81,630	His	85,763
24	This	81,363	But	83,153
25	You	73,224	You	80,159
26	Or	73,207	Which	77,973
27	Say	68,860	Do	77,815
28	We	64,329	Or	73,862
29	His	58,091	They	69,516
30	One	51,297	She	64,132
31	All	47,942	Her	61063
32	Will	47,406	There	54,555
33	Would	46,895	One	54,068
34	Can	45,584	Their	53,603
35	About	45,526	We	52,423
36	More	43,630	Say	52,423

37	Who	42,775	All	50,426
38	Which	41,658	Can	46,319
39	She	41,512	If	44,882
40	If	41,367	More	43,034
41	Time	41,265	Make	42,747
42	Use	41,001	Who	40,072
43	There	39,494	When	39649
44	Their	39,308	So	39,538
45	Make	37,774	Other	36,508
46	So	37,362	Up	36,458
47	Get	36,787	Out	36,341
48	Other	36,550	No	36,073
49	When	36,533	What	35,479
50	My	35,922	Its	35,390
51	New	34,860	Time	35,350
52	What	34,344	About	34,145
53	Than	34,245	Take	33,427
54	Like	34,021	See	32,741
55	Go	33,649	Some	32502
56	Out	33,600	Year	32,070
57	Its	31,961	Into	31,861
58	Think	31,680	Go	31,149
59	No	31,661	Could	30,638
60	Also	30,799	Than	30,543
61	Up	30,468	Use	30,472
62	Year	30,194	Only	29,967
63	See	29,860	Him	29,786
64	Just	29,839	Them	29,422
65	Some	27,954	Work	28,482
66	Know	27,300	May	26,761
67	These	26,841	Two	26754
68	Only	26,102	Also	26,313
69	Good	26,084	My	25,841
70	Into	24,587	Then	25,711
71	Because	24,436	New	25,471
72	Two	24,380	Give	25,290
73	Could	23,852	Know	24,925
74	Take	23,671	Come	24,866
75	People	23,239	Get	24,807
76	After	22,622	These	24,657
77	Most	22,555	Like	24,553

78	Show	22,047	Over	24,181
79	Work	21,793	Such	24117
80	Find	21,775	First	24,017
81	Now	21,604	Any	23,602
82	May	21,569	Now	23,245
83	Cell	21,253	After	22,910
84	First	21,076	People	22,301
85	Her	20,567	Think	22,097
86	Even	19,649	Me	21,401
87	How	19,552	Way	21,157
88	State	19,230	Most	21,137
89	Over	19,000	Your	20,932
90	I	18,801	Should	20,837
91	Report	18,692	Very	20,621
92	Come	18,471	Look	20,228
93	Then	18,346	Man	19,750
94	Your	18,259	Where	19,628
95	Way	18,226	Find	19,560
96	Our	18,023	Between	19,491
97	2	17,939	Many	18,397
98	Give	17,930	Even	18,328
99	Such	17,621	Just	18,025
100	Any	17,517	Back	17,856
Sum Freq.		8,822,607		9,380,376

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

1. U. Strauss, F. Fan, G. Altmann, *Problems in Quantitative Linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G. Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro.* 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in Quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4.* 2014, VI + 148 pp.
15. K.-H. Best, E. Kelih (Hrsg.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014, IV + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, *Unified modeling of length in language.* 2014. III + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical approaches to text and language analysis.* 2014, IV + 230 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA. Quantitative Index Text Analyzer*. 2014, IV + 106 pp.
19. K.-H. Best (Hrsg.), *Studies zur Geschichte der Quantitativen Linguistik. Band 1.* 2015, III + 159 pp.
20. P. Zörnig et al., *Descriptiveness, activity and nominality in formalized text sequences*. 2015, IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5.* 2015, III+146 pp.
22. P. Zörnig et al. *Positional occurrences in texts: Weighted Consensus Strings.* 2016. II+179 pp.

23. E. Kelih, E. Knight, J. Mačutek, A. Wilson (eds.), *Issues in Quantitative Linguistics Vol 4.* 2016, 287 pp.
24. J. Léon, S. Loiseau (eds). *History of Quantitative Linguistics in France*. 2016, 232 pp.
25. K.-H. Best, O. Rottmann, *Quantitative Linguistics, an Invitation*. 2017, V+171 pp.
26. M. Lupea, M. Rukk, I.-I. Popescu, G. Altmann, *Some Properties of Rhyme.* 2017, VI+125 pp.
27. G. Altmann, *Unified Modeling of Diversification in Language*. 2018, VIII+119 pp.
28. E. Kelih, G. Altmann, *Problems in Quantitative Linguistics, Vol. 6.* 2018, IX+118 pp.
29. S. Andreev, M. Místecký, G. Altmann, *Sonnets: Quantitative Inquiries.* 2018, 129 pp.