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

(Open Access)

Indexed in ESCI by Clarivate Analytics and SCOPUS by Elsevier

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift (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 **Druckversion** 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

Editorial and Peer Review Process

Glottometrics is a peer-reviewed scientific journal with a rigorous editorial screening and assessment process made up of several stages:

Orders for CD-ROM or printed copies to: RAM-Verlag@t-online.de
Free PDF-Download: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 47 (2020), Lüdenscheid: RAM-Verlag, 2020. 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 (2020)
online/ e-version ISSN 2625-8226 (print version ISSN 1617-8351)

Glottometrics 47, 2019

Abstracts

Yingying Xu

The Distribution of Word Families in Chinese College English Textbooks

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.

Zhao Gao

A Quantitative Lexical Study on Commercial English

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.

Jingjie Li

Inter-textual Vocabulary Growth Patterns for Marine Engineering English

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.

Hong Su

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

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.

Yaobin Yan

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

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.

Pianpian Zhou

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

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.

Yujia Zhu

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

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

Fangfang Zhang

Computational Stylistic Characteristics of American English

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.