Frequency distributions of punctuation marks in English

KUN SUN AND RONG WANG

Evidence from large-scale corpora

Introduction

The analysis of punctuation in philology is mainly carried out with a view to better understand the meaning of the literature concerned. Punctuation is generally believed to play the role of 'assisting the written language in indicating those elements of speech that cannot be conveniently set down on paper: chiefly the pause, pitch and stress in speech' (Markwardt, 1942: 156). Most of us often ignore the importance of punctuation in writing systems and tend to believe that punctuation only depends on tradition and the personal styles of writers. In fact, punctuation marks may contribute significantly to the clarity of expression. Many linguists associate punctuation with intonation, but the truth is more complex than that – punctuation marks may affect orthography, morphology, syntactic relations, semantic information, and can even influence textual structure.

Before the 1980s, most studies on punctuation were prescriptive, usually neglecting descriptive approaches, and much of the research focused on philology, in most cases seeking to explain the unclear meaning of punctuation in ancient literature. For example, style guides and grammar books (e.g. Partridge, 1953) gave prescriptive accounts of punctuation. The first descriptive attempt can be found in Meyer's PhD dissertation (Meyer, 1987) where he synthesized an account of punctuation through a small English corpus. Punctuation in grammatical relations has been explored in Quirk et al. (1985: 1610–1639), Nunberg (1990), Jones (1996), Huddleston & Pullum (2002: 1731) etc. For example, Nunberg (1990) first put forward the idea of the 'linguistics of punctuation', together with an attempt to situate punctuation as a linguistic subsystem which is closely associated with lexis and grammar. He did so through the construction of two concepts: text-grammar and lexical grammar.

The basic functions of punctuation marks can be classified into two types: grammatical and rhetorical. With regard to grammatical functions, punctuation marks are used to show the boundaries between segments and to indicate how the segments of text are supposed to relate to one another. In contrast, in rhetorical functions, they show the emphasis or prosody that readers want to give to a segment or a larger segment. However, the two functions have not been equally developed in the historical tradition. For one thing, few studies on punctuation up to now have directly engaged with the characteristics and development of language because most of these studies have focused on rhetorical, linguistic and orthographical

KUN SUN, Ph.D., is Associate Professor, Postdoctoral Fellow at Zhejiang University, China, as well as Postdoctoral Researcher at Tuebingen University, Germany. His academic interests cover issues in text grammar, corpus linguistics and punctuation studies. He has published more than 20 papers in journals, conference proceedings and edited books. Email: sharpksun@hotmail.com

RONG WANG, PhD, is a lecturer at Hangzhou Dianzi University, China. Her research interests include applied linguistics, corpus linguistics and literary translation studies. Email: rong4ivy@163.com

doi:10.1017/S0266078418000512

functions. Although the rhetorical, linguistic and orthographical functions of punctuation are closely related to language use, few studies on punctuation have probed into the characteristics and development of language through analyzing these functions. Further, the rapid development of corpora and the technology of natural language processing allow us to collect linguistic data and make more scientific investigations of linguistic phenomena. Specifically, with the help of new techniques, quantitative analysis will help us better understand patterns of punctuation although only few studies on punctuation hitherto have been built on data analysis. For example, the analysis of the frequency distributions of punctuation marks from synchronic and diachronic perspectives helps us discover patterns and regularities in language use.

The frequency distribution of punctuation based on large-sized corpora has rarely been investigated before. Anyone with a little awareness can see that commas and full stops are used heavily in written language. Frequency can be treated as a ranking of the occurrences of a given phenomenon. For instance, word frequency refers to words as ranked by their frequency in language (the words with highest frequency here are well known: *a, and, he* …) Similarly, various punctuation marks are used with differing frequency. Frequency plays an important role in language and in the physical world. Nature is strict in its rules and laws; hence we can often discover its patterns in the form of frequency distributions.

This paper will focus on punctuation in different registers, varieties of English and the development of English. There have been numerous corpus-based descriptions of linguistic characteristics of particular registers which are treated as different textual categories, such as the novel, spoken language, the lecture etc. Studies have also been made using comparisons across the registers. These studies are linguistic descriptions of lexical and grammatical features (e.g. Biber, 1988, 1995; Biber et al., 1999). However, although the frequency distribution of punctuation is helpful in showing differences between registers, the punctuation perspective differing across registers has seldom been taken. The frequency distribution of punctuation marks in different registers helps us to understand the many distinctions between these registers. Additionally, English has been widely used in many countries across the globe. These Englishes differ from each other to some extent, and they are treated as varieties of English. Considering the importance of frequency distributions, we can ask whether different punctuation marks differ in frequency across different registers and varieties of English as well as whether these patterns may have changed over time.

Addressing these questions can definitely help us to gain a better understanding of English and its characteristics. With the development of new technology, we have far greater access to frequency distributions of punctuation marks and can examine them from the perspective of big data.

A thorough study of punctuation using frequency data is therefore definitely helpful in understanding the role that punctuation marks have played in the English language from synchronic and diachronic perspectives. We will answer the following questions in this paper:

1) What statistical patterns do the frequency distributions for the different punctuation marks in English follow?
2) Are these differences in punctuation mark use across various English registers? Do punctuation marks show a similar frequency distribution in the global varieties of English?
3) What changes have these punctuation marks undergone in last five hundred years? Are there any regularities or patterns in these changes?

Data and method

Corpora and Google Books N-gram Viewer were employed heavily for obtaining the data for the present study. The main corpora used in the current study were the Brown corpus, COCA (Corpus of Contemporary American English) (2014), COHA (Corpus of Historical American English) (2014), BYU-BNC (British National Corpus) (2007), and GloWbE (Global Web-Based English) (2013). In addition, the Google N-gram viewer was used for further insights. In this study, the frequency data from COCA, COHA and GloWbE were given in normalized figures per one million words.

In what follows, some details are given about the electronic resources used for the present study. As the largest freely-available corpus of American English, COCA contains more than 520 million words of text and is comprised of five registers. Containing texts from the 1810s to the 2000s, COHA is the largest corpus of historical American English, consisting of texts from different registers. BNC contains 100 million words of British English text from a wide range of registers. The corpus of GloWbE is a large English corpus collecting international English from the internet, containing about 1.9 billion words of text from

twenty different countries. For further information on the corpora used, see https://corpus.byu.edu/.

Using a yearly count of N-grams found in the sources printed between 1500 and 2008 in Google's text corpora in English, Chinese, and other six languages, Google Books N-gram Viewer (Google Books, 2010) is an online search engine, which outputs a graph that depicts the historical changes of frequency for a particular phrase (or word). It is also currently the world's largest corpus and the only corpus that enables resolution at a fine temporal scale (yearly) over a long period of time (Michel et al., 2011). The developers of the viewer aimed to create a new approach to humanities research, which would make it possible to rigorously study the evolution of culture using distributional, quantitative data on a grand scale (Bohannon, 2010). Google N-gram viewer thus enables further understanding of the relationship between language and its culture.

The punctuation marks explored in the current study are the common ones: period, comma, colon, semicolon, hyphen, question mark, dash, exclamation mark, parenthesis, apostrophe and slash. The frequency distributions of these punctuation marks can be obtained using the corpora mentioned above.

Patterns in the frequency distributions of punctuations marks

Three corpora (Brown, COCA and BNC) were searched to collect the frequency data of each punctuation mark. The frequency of punctuation marks calculated in the current study is relative to word count in corpora. The frequencies of punctuation marks per million are shown in Figure 1.

In Figure 1, the x-axis shows the rank according to the punctuation frequency distribution; the y-axis shows the normalized frequency per million. Statistical analysis shows that the frequency patterns in the use of punctuation marks in COCA and BNC follow a growth regression model, while they follow a power law[1] for Brown.[2] On one hand, the growth regression model, also called growth curve model, captures how a particular quantity increases over time. Growth curves are used in statistics to determine the type of growth pattern of the quantity— be it linear, logarithmic, or exponential. Logarithmic regression in growth curves might present a heavy-tailed distributional pattern. On the other hand, the power law distribution captures a phenomenon whereby a small number of occurrences are common, while instances of larger occurrences are rare. The power law distribution can be found in a wide variety of physical, biological, cognitive, social and artificial phenomena (Kello et al., 2010; Clauset, Shalizi & Newman, 2009).

In theory, nature, animals (including people), and even well-designed machines will naturally choose the path of least effort so as to reach the best result. The two distributions shown in the frequency of punctuation marks are both heavy-tailed distributional patterns. The power law can explain the behavior that humans make the least effort (Kello et al., 2010), and the growth model also helps explain least-effort behavior. In such a

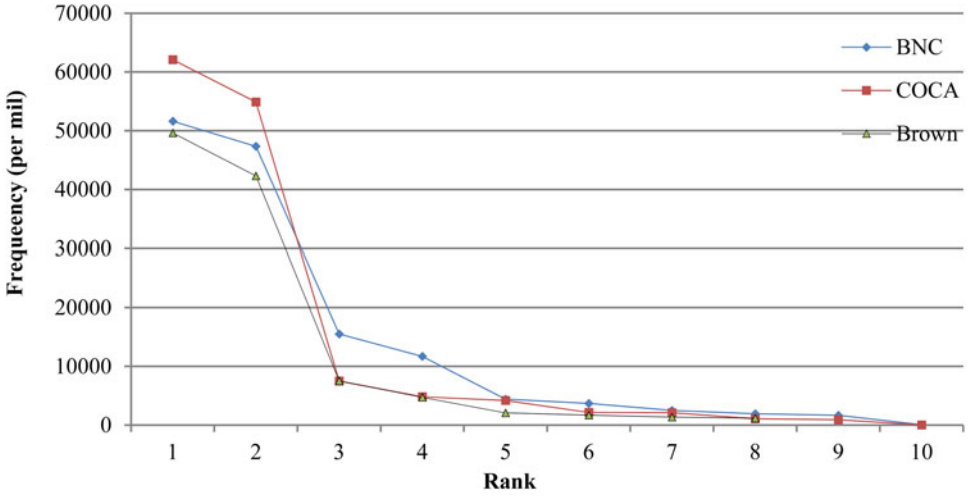

Figure 1. The distribution of frequency for all punctuation marks combined in three corpora

sense, there is very little difference between the two models to demonstrate the least-effort principle.

The frequency distribution of punctuation marks in different registers

After identifying the general statistical pattern of frequency distributions for punctuation marks, we can also ask whether the frequency distributions vary in the different registers. For example, academic English may use few exclamation marks so as to avoid displaying subjectivity and personal emotions in addressing serious topics. This section will examine how the frequencies of punctuation marks are distributed across various registers in English.

COCA provides five different registers for inspection. The data are depicted in Figure 2. Figure 2 shows the normalized frequency per a million (y-axis) for different punctuation marks (x-axis).

It is necessary for a statistical test to be taken to check whether there are significant differences among these data. AVONVA is a commonly effective method to check this kind of significance. One-way ANOVA is used to test the null hypothesis that the means of several data within a group are all equal. One-way ANOVA test in Excel shows that if $F > F\ crit$, we reject the null hypothesis, indicating that the difference among values is real. This is the case, 116.5 (F) > 2.124 ($F\ crit$). Therefore, we reject the null hypothesis, that is to say, in COCA, the five registers display different frequency distributions with regard to punctuation marks.

Firstly, academic writing displays the fewest periods[3], question marks and exclamation marks, but the highest figures for parentheses and semicolons. This demonstrates that sentences in the academic register are longer than in other registers. It is also clear that academic English seldom displays questions or exclamatory sentences. In addition, parentheses are used extensively in academic texts for citations, or occasionally for explanations, while they are seldom employed in other registers, probably so as not to interrupt the flow of text.

Secondly, fiction shows the highest frequencies of periods and exclamation marks. Sentences in fiction tend to be short. Exclamation marks are widely employed in fiction so as to show emotions or the atmosphere. Additionally, fiction uses the fewest hyphens.[4] In contrast, newspapers and fictions tend to combine words through hyphenation into new units to express up-to-date ideas and temporary concepts (see in the following paragraphs and sections for further discussions at this point).

Thirdly, question marks and colons are used the most frequently in spoken English (written transcriptions of recorded spoken language). In daily communication, people often use questions or exclamations, which is a sign of an informal genre for pragmatic communications. The use of the colon is a little complicated because its functions cover many aspects such as the annunciatory, explanatory, appositive, parallel, etc. The colon can therefore be widely used when spoken language is transcribed into written form so as to express complex situations. That is why the frequency of colon in transcribed spoken English is much higher than other registers.

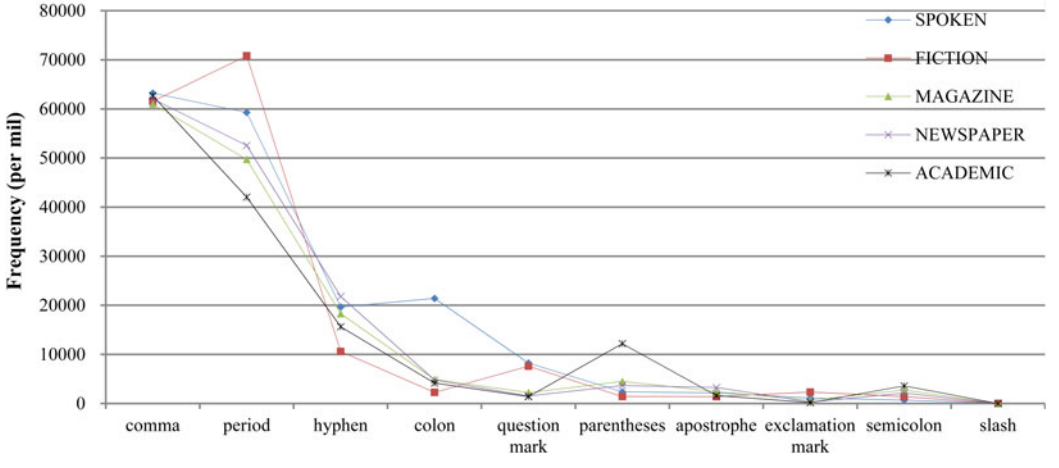

Figure 2. The distribution of punctuation marks for different registers in COCA

The newspaper genre has the highest frequency of apostrophes and hyphens. 'The 'kiss and tell' principle expresses the essence of good journalism: keep it short and simple and tell the story' (Busà, 2014: 96), and the principle indicates that news prefers simple and concise language; hence abbreviations (e.g. 'Strategic Health Authority' becomes 'SHA'), apostrophes and hyphens between words (e.g. some verb phrases need hyphens when they are used as nouns, like *check-up*, *break-in*, *turn-on*) tend to be extensively used. As for hyphens, Journalism BBC News style guide (2018) suggests, ' Hyphens are often essential, if the text is to make immediate sense', and the other reason is that hyphenation is easy to integrate several words into a unit, yielding a new meaning which allows journalists to express concepts in novel and appealing ways.

It is also of interest to look at the frequency of punctuation marks in British English (see Figure 3). BNC provides two more registers, NON-ACAD and MISC-miscellaneous, that are not included in COCA. A contrastive analysis between COCA and BNC is also useful for understanding the differences between American and British English.

One-way ANOVA shows that the difference is quite significant (57.67 (F) > 2.04 (F *crit*)) in BNC. The frequencies of the punctuation marks included in the analysis in these registers are quite similar to the results obtained for American English. However, British English newspapers display the most colons. Conversely, American newspapers use the most apostrophes and hyphens. British fiction uses the most periods, which indicates that the sentences are shorter than sentences in the American English material. However, apostrophes and quotation marks are mostly used in fiction and in this respect BNC data differ from American English.

Frequency distributions of punctuation marks in different varieties of english

The differences across varieties of English have been widely discussed from the perspectives of lexicon, syntax, semantics and discourse. However, few studies have investigated the differences in the uses of punctuation (The Punctuation Guide [2018] is an exception). Yet there may be differences in the frequency distributions for punctuation marks in different varieties of English. This section will examine the frequency of punctuation marks attested for 20 English-speaking countries and regions (for the varieties included in the study, see Table 2). The data were acquired through GloWbE.

SPSS was used to examine their discrete degrees, which are chiefly measured by the values of skewness. Skewness is a measure of symmetry. A distribution is symmetric if it looks the same to the left and right of the center point.

Shown in Table 1, skewness, as a statistical index, is a fitting measure of the discrete degrees within a group of numbers. A symmetric distribution such as a normal distribution has a skewness of 0. When data are skewed left, values for the skewness are negative; in contrast, positive values for the skewness indicate that data are skewed right. As positive skewness increases, the degree of asymmetry increases. For example, the skewness values for all twenty columns are 0.597 for comma and

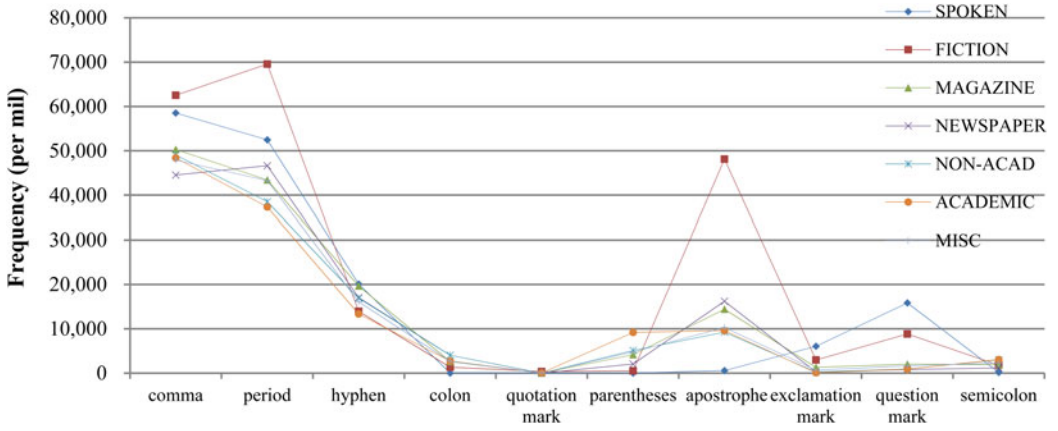

Figure 3. The distribution of punctuation marks for different registers in BNC

Table 1: Statistical values of punctuation frequency in 20 English-speaking countries

Name	Number	Minimum	Maximum	Mean	Skewness	Kurtosis
Comma	20	41440.91	48079.56	44189.96	0.6	−0.81
Period	20	40349.48	45144.21	42840.02	0.13	−0.9
Parentheses	20	3577.14	6448.4	4500.81	1.98	7.06
Question_mark	20	2799.68	5663.33	4154.78	0.19	0.8
Colon	20	2532.12	4115.78	3221.82	0.21	−0.44
Exclamation_mark	20	1162.83	3554.11	2057.22	1.01	1.68
Apostrophe	20	1958.05	5047.82	2980.35	1.11	0.87
Semicolon	20	944.02	1657.34	1355.22	−0.47	−0.01
Hyphen	20	7389.26	10565.52	9529.78	−0.97	2.49

0.125 for periods. This indicates that the difference of the frequencies of commas used in twenty countries is much larger than the difference of the frequencies of periods. The values of skewness in Table 1 show that the frequencies obtained for parentheses, exclamation marks, apostrophes, hyphens and commas vary greatly. However, the use of periods and question marks varies the least between varieties of English. It implies that the boundaries of statements and questions in English seem to be agreed more anonymously by English speakers with various backgrounds than the practices of other punctuation marks due to both obeying syntactic rules more rigidly and strictly than other punctuation marks.

As shown in Table 2, the length of sentences is very similar among the varieties of English. In strong contrast to the use of periods, the varieties of English exhibit great differences in the use of commas probably due to many and varied functions of the comma (Partridge, 1953: 14–41) distinguishes at least twelve different functions for commas).

The greatest difference concerns parentheses. Pakistani English shows the most parentheses (6448.4/mil), almost double in number compared to Nigerian English (3577.14/mil). UK English uses the most apostrophes (4585.83/mil), 2.34 times as many as the Canadian variety, which uses the least (1958.05/mil). Interestingly, Singapore English uses the most question marks (5214.09/mil) and exclamation marks (3554.11/mil). In Singapore English, there is an abundance of frequently used sentence-final particles (such as *har, hor, leh, lor, meh, siah*, *wat*) to express exclamations and questions (Leimgruber, 2013: 84–95), which might be one important factor responsible for the highest frequency of using question marks and exclamation marks in Singapore English.

A diachronic perspective

After the synchronic descriptions, a diachronic perspective will provide an interesting picture of the role played by punctuation marks in the development of English. This section will focus on changes in the frequencies of punctuation marks in the history of English from 1500 to 2008. The data are captured through Google N-gram viewer, as shown in Figure 4.[5]

In what follows, comments are provided on a number of the punctuation marks included in the analysis.

Period: By way of background to the use of the period in the history of English, Liberman (2011) once launched an American Presidency Project, showing that mean sentence lengths have been falling since the founding of the republic and have undergone a cumulative drop of roughly 50%. Haussamen (1994) found that the printed English sentence had become shorter by comparing the number of words in written sentences from 1600 to the 1980s. Haussamen (1994) accordingly suggested that the printed sentence will continue to develop into a similar direction over the next two centuries. More than 100 years ago, Lewis (1894: 34) concluded that the English sentence had decreased in average length by at least one half in 300 years (prior to the 1890s).

As the number of periods in English has continued to rise, the length of sentence is very likely

Table 2: Frequency Distribution of Punctuation Marks in Twenty Regional Varieties of English

Country	comma	period	parentheses	question mark	colon	exclamation mark	apostrophe	semicolon	hyphen
United_States	48,079.56	45,009.18	4,489.37	4,442.76	3,640.18	2,443.08	2,088.35	1,399.68	841.63
Canada	46,717.16	43,811.76	4,854.26	4,055.34	3,288.15	2,003.05	1,958.05	1,547.04	948.1
Great_Britain	42,895.05	40,986.32	4,025.45	4,040.89	3,276.34	2,104.74	5,047.82	1,199.63	1,783.70
Ireland	41,771.67	41,624.99	4,092.55	3,775.82	2,954.37	2,029.49	3,765.84	1,293.14	1,290.57
Australia	41,884.82	42,261.94	4,519.36	3,980.38	3,809.27	2,370.34	4,585.83	1,334.66	1,547.54
New_Zealand	41,440.91	42,747.16	4,746.67	3,869.13	3,050.04	1,985.76	3,717.01	1,211.00	1,526.85
India	43,684.18	44,509.51	4,246.41	4,101.95	3,860.88	1,658.92	2,947.36	1,192.38	1,066.46
Sri_Lanka	42,408.93	43,520.38	4,169.06	4,483.81	2,605.60	1,336.26	3,795.17	1,295.81	773.82
Pakistan	42,850.36	41,983.72	6,448.40	5,663.33	4,115.78	1,692.75	3,257.90	1,390.83	658.59
Bangladesh	43,784.92	44,079.39	4,651.85	3,332.31	3,302.73	1,162.83	2,792.06	1,251.47	816.93
Singapore	44,249.96	44,249.87	4,702.49	5,214.09	3,093.89	3,554.11	2,398.29	944.02	996.07
Malaysia	43,260.56	43,719.57	4,850.96	4,715.72	3,200.43	3,041.59	2,752.71	1,007.35	1,010.96
Philippines	47,240.42	45,144.21	4,735.97	4,209.22	3,479.30	2,557.57	1,995.80	1,603.57	905.07
Hong_Kong	47,254.74	43,074.82	4,602.59	4,428.92	3,361.98	1,661.10	2,492.52	1,489.70	891.09
South_Africa	42,418.74	41,981.82	4,388.48	3,915.18	3,405.77	2,095.98	2,588.30	1,425.12	1,198.16
Nigeria	44,639.61	41,758.10	3,577.14	4,703.88	2,532.12	2,253.41	2,469.09	1,657.34	667.7
Ghana	44,203.02	40,349.48	4,337.60	4,410.03	2,675.59	1,497.28	2,740.18	1,466.82	623.55
Kenya	43,422.31	42,095.22	3,885.32	3,178.30	2,688.38	1,731.22	2,368.15	1,470.15	893.06
Tanzania	44,871.09	42,149.22	4,486.23	2,799.68	3,293.77	1,882.56	2,789.82	1,596.98	1,271.24
Jamaica	46,721.21	41,743.64	4,206.07	3,774.89	2,801.74	2,082.30	3,056.68	1,327.66	1,177.37

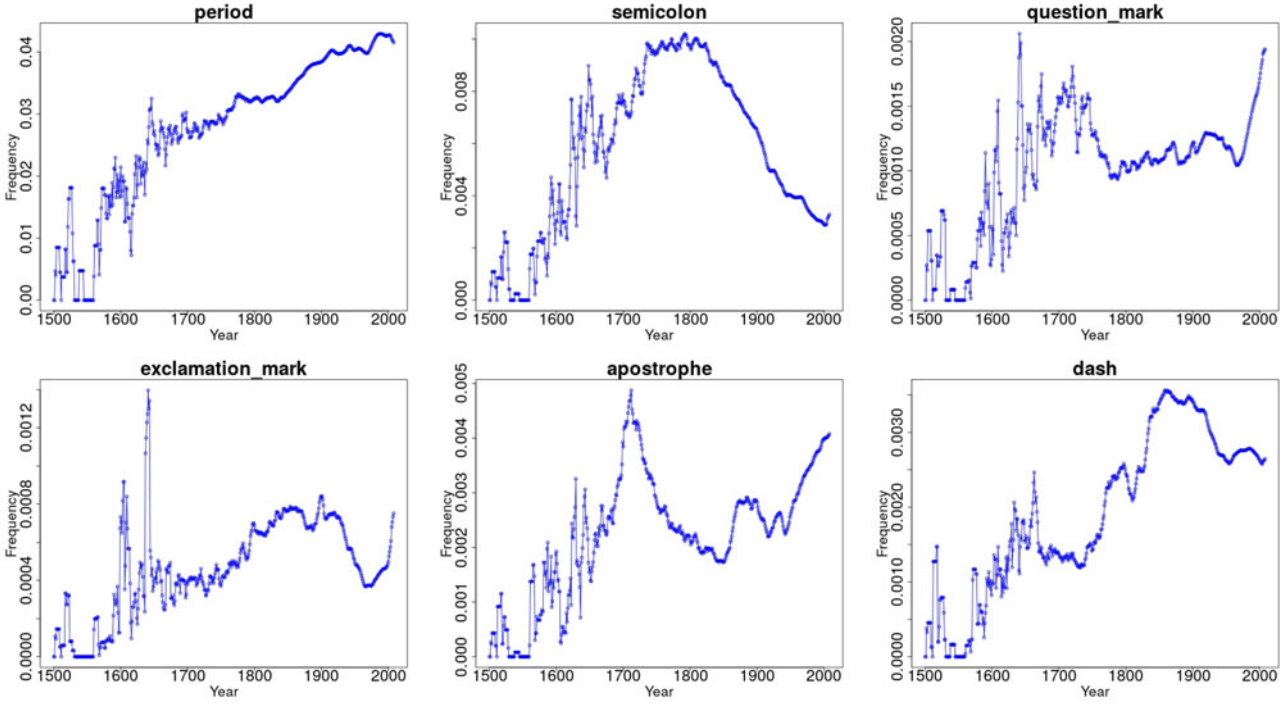

Figure 4. Changes in the frequencies of punctuation marks in Google N-gram viewer (1500–2008)

to have become shorter. These findings and predictions are confirmed by Google N-gram data. The graph on the top left of Figure 4 shows that the percentage of periods has continued to increase steadily over the last three hundred years, rising from 3% to 4.3%. Currently, social networking tools reinforce the tendency to use shorter sentences.

Semicolon: A jagged upward trend can be seen in the use of the semicolon in English, peaking around 1800, and afterwards the semicolon experienced a long, smooth decline. This tendency has also been identified in some earlier studies, such as Bruthlaux (1995).

Early 17th-century writers used colons, semicolons, and commas interchangeably.[6] The semicolon prospered before the 19th century; however, its frequency of use has fallen in the last two hundred years. It seems that currently the use of semicolons is associated with difficult or abstract topics; hence writers tend to decline to use semicolons, as previously described. Nunberg (1990), author of *The Linguistics of Punctuation*, holds the view that the semicolon seems to be reserved nowadays for certain kinds of highbrow and high-middlebrow writing (Kelly, 1999). The data in Figure 4 are consistent with the observation that the frequency of semicolon has declined.

Question mark: Since 1800, the frequency of question marks fluctuated mildly, but it drastically increased after the 1970s. This tendency is highly likely to continue in the future because of the rise of social networking media where question and exclamation marks are used frequently.

Exclamation mark: The frequency of the exclamation mark also kept fluctuating between 0.06% and 0.08% during the 19th century before continuing to increase between the 18th century and the mid-19th century with some fluctuations. However, its frequency decreased in the 20th century, reaching the lowest point in 1960s. This could indicate that there is a reduction in the number of exclamatory sentences used to express feelings and emotions. However, the last half-century has seen an upward trend in the use of the exclamation mark, most likely because of the wide application in social media.

Apostrophe: The data of the Google corpora show that the climax for the frequency of apostrophe was reached in the year 1712. However, the use made of apostrophes afterwards underwent a dramatic fall up until 1850. Strangely, it seems that the use of the apostrophe has become fashionable again in the last 60 years, especially in newspapers and magazines. The revival of the apostrophe is also likely to be the result of the rise of social networking, which will be supported by more evidence in the latter part of this section.

Dash: The dash referred to here is the so-called m-dash, '—', different from a spaced en dash (the m-dash is twice as long as the en dash), used as a break in a sentence or to set off parenthetical statements. The use of the dash increased after 1750, then reached its peak (about 0.35%) in 1860, but afterwards continued to drop up until the 1950s before starting to fluctuate between 0.25% and 0.275%.

As for other punctuation marks, in Google N-gram viewer, commas are missing as they are used as dividers; the colon is not available, either. Hence the COHA is used to supplement this missing data (see Figure 5).

The frequency of the comma has slowly dropped in the last two hundred years in American English, but the period evinces a reverse tendency to the comma. The decline of the use of commas might be caused by the fact that their rhetorical function has weakened in the history of English. Partridge (1953: 14) comments: 'In modern usage, the comma is used predominantly for the grammar, the construction or syntax, of a sentence; formerly the comma indicated primarily the rhetorical pauses, as quite often, it still does.' His point will be further discussed in the next section.

Hyphenated expressions are compounds with hyphenation, such as *short-term, would-be, decision-making*. Hyphenated expressions can express up-to-date ideas and temporary concepts (such as, *floor-to-ceiling* windows, a *back-to-back* connection, *1980s-style* dancing, *industrial-scale* organic producers), and the hyphenated use can make a phrase become a word, such as the premodifiers in '*state-of-the-art* article' and '*top-of-the-line* use'. A lexical-grammatical device of this kind seems to have become popular in contemporary English. For American English, the frequency of hyphenated two-word expressions has increased in the last 200 years, as shown in two graphs (having different perspectives: component numbers vs. which part of speech [POS] the expression as a whole belongs to) at the bottom of Figure 5. The same tendency can also be seen in British English.

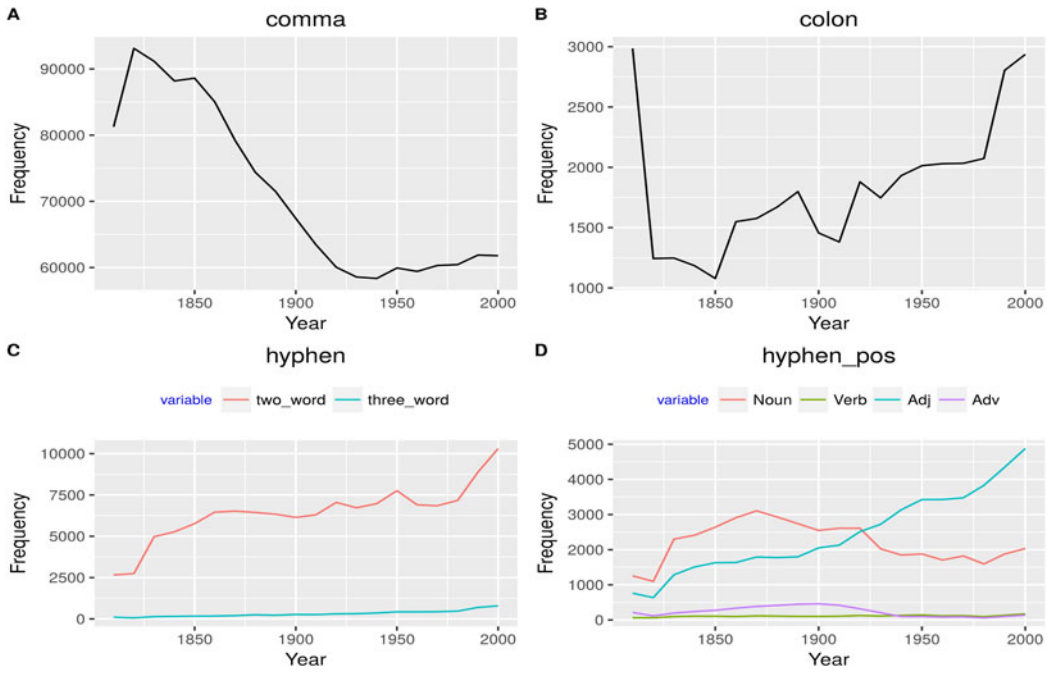

Figure 5. The change in the frequencies of the comma, colon, and hyphen in COHA (1800–2000)

Discussion

From the synchronic perspective, the frequency of punctuation marks follows a 'heavy-tailed' distribution, abiding by the pattern that humans make the least effort. In the frequencies of words in most languages, the phonological system follows the power law; a well-known example of this is Zipf's law (Zipf, 1949), which has also been widely attested for word frequencies and some syntactical phenomena. However, the frequency distribution for punctuation marks in English does not always seem to follow the power law. The question remains as to why this is so, unlike for the other subsystems in language.

The frequencies of punctuation marks in different registers are very helpful in distinguishing genres, styles and other linguistic features. The differences in the frequencies of punctuation marks attested in comparisons between COCA and BNC also reflect the differences in usage and styles between American and British English.

The varieties of English differ with regard to the frequencies of specific punctuation marks. While periods and question marks occur with quite similar frequencies across all the varieties of English, the frequency of parentheses, exclamation marks, apostrophes, hyphens and commas varies between varieties of English. The differences in the frequencies for these punctuation marks are likely to be caused by a wide of variety of complex social-cultural factors. Studies (e.g. Crystal, 1985; Kirkpatrick, 2010) have considered differences between varieties of English through a variety of approaches, among them pidgin and creole studies, lingua franca, linguistic futurology, and sociology. These frameworks might provide explanation of the soci-cultural factors potentially influencing differences on frequencies attested for punctuation in English varieties.

The present study has also shown that the frequencies of the individual punctuation marks in the history of English have undergone dramatic changes. As mentioned previously, the grammatical and rhetorical functions of punctuation have often been interwoven, but they have not been studied to an equal extent. The weakening rhetorical functions of punctuation have been explored in other studies. For example, Schou (2007: 213) points out that 'from 1800 to the present, the rhetorical aspect of punctuation has dwindled into the possibility of marking asyndetic coordination by a semicolon and the option of one specific rhetorical relation expressed by the colon; the rest is left to the full stop'. Schou's observation and Partridge's point on commas can be supported by our data. The other example is the semicolon and

Figure 6. Punctuation marks in Tweets

colon. The long-term decline in the use of the semicolon began in 1800 and has continued to the present day (middle top panel in Figure 4); in contrast, the frequency of the colon continued to drop before 1850, but began to rise after 1850 (left panel of Figure 5). This is most likely due to the colon's replacing the semicolon in certain rhetorical functions.

Therefore, as Schou (2007: 213) also proposes, 'the general experience is that syntax has been central at least since 1600, although prosody played and still plays a certain role. Punctuation and its theory have increasingly moved towards a syntactic orientation'. It also seems that English native speakers have been influenced unconsciously by the syntactic orientation of punctuation and started to use fewer commas owing to their redundant rhetorical function. In short, as Schou (2007: 214) puts it, 'this development can be therefore characterized as moving from the modern rhetorical-grammatical punctuation of 1800 to **the modernistic stylistic-grammatical punctuation**'. This can be observed and supported from the data presented in this study.

Stylistic functions feature in the use of some punctuation marks that have been used creatively for communicative purposes. We will now examine how these stylistic functions of punctuation have been influenced by modern communicative purposes. The use of punctuation has been greatly influenced by writing and communication technologies, particularly social networking tools in the internet era. For example, Twitter has a word limit, which means that its users have to use short sentences. Frequently they strengthen their emotional impact by using exclamation marks, question marks, en dashes for emphasis, and emoji, as demonstrated by President Trump and J. K. Rowling (see Figure 6).

Inversely, the use of punctuation marks in social networking influences contemporary English. Another example can be taken to illustrate this point. It was found that the use of apostrophes in contractions (e.g. *can't* rather than *cannot*) was higher in the Twitter messages (Denby, 2010) than in either the text messaging (MS) or instant messaging (IM) according to the data collected by Ling and Baron (2007). The use of apostrophes was identified in 97% of the occurrences with contractions in Twitter as compared to 94% for IM and 32% for TM. This study can at least partly explain why the frequency of apostrophe use has been increasing recently, as shown in Figure 4. The frequency of sentence-final punctuation in Tweets is much higher than in MS and IM, as shown in Table 3 (Denby, 2010). This indicates that sentences in tweets are shorter in comparison with MS and IM, which may have been influenced by the widespread use of social networking.

Written English in the internet era is likely to represent informal speech, just as Baron (2001: 56) suggests, 'The semi-stable grammatical model of the past century is being abandoned. Instead, punctuation increasingly marks the cadences of informal speech in the case of email and other contemporary language media, and helping the eye makes sense in messages that are intended to be viewed quickly'. Actually, the stylistic-grammatical function of punctuation facilitates informal speech represented in writing in the internet era. In particular, the wide use of social networking sites such as Facebook, Twitter and Blog has allowed people to read and write language in electronic media and to freely interact

Table 3: The Use of Punctuation in Social Networking

Feature	Twitter	MS (Ling & Baron, 2007)	IM (Ling & Baron, 2007)
Standard apostrophe use in contraction (as percentage of total)	97%	32%	94%
Sentence-final punctuation (as percentage of potential sentence-final punctuation)	81.1%	39%	45%
Transmission-final punctuation (as percentage of potential transmission-final punctuation)	70%	29%	35%

using language in public. That is why the internet can facilitate the use made of stylistic functions for some punctuation marks.

Conclusion

This study analyzes data on the frequency distribution for English punctuation marks from some large corpora. From both the diachronic and synchronic perspectives, we found that the frequency distribution for English punctuation follows the laws of least effort. The study showed that there were differences in how different punctuation marks were used in different registers. The varieties of English were also found to differ with regard to the frequencies of specific punctuation marks. In the last 300 years, the practices of punctuation marks have become more syntactical rather than rhetorical or prosodic in nature. These changes in the frequencies of punctuation marks are evident in, for instance, the shorter sentences and increase in the use of hyphenated compounds and apostrophes in contracted forms. These developments show that modern stylistic-grammatical punctuation is developing under the influence of modern writing and communication technologies.

Notes

1 The regression equations for BNC, COCA and Brown can be represented as respectively:

BNC: $\ln(y) = 11.729 - 0.603x$ ($R2 = 0.866$);
COCA: $\ln(y) = 11.878 - 0.717x$ ($R2 = 0.856$);
Brown: $\ln(y) = \ln(77777.7) - 2.064x$ ($R2 = 0.937$).

2 Brown corpus can be retrieved by online Sketch Engine (https://the.sketchengine.co.uk/bonito/corpus/first_form?corpname=preloaded/brown_1;align=); however, it is not possible to search for question marks (?) and there is no way to distinguish between the dash and the hyphen. The size of this corpus is small – roughly one million words. All might influence the fitting result.

3 Periods are also used following abbreviations, such as Mr., Dr., p.m.; however, the amount of such instances is not large. Due to the difficulty of identifying their proportion in the data, they are included in the counts with full-stops.

4 It is not possible to search '-' in COCA and BNC. Instead, '*-*' (*asterisk wildcard representing any word or morph) works in the search in two corpora. '*-*-*' represents a hyphenated compound with two hyphens, but many irrelevant symbols are calculated in COCA even if the amount of hyphenated compounds of this kind is not large (about over 100 per mil). Hence these are not included in the counts.

5 Although N-gram Viewer can output a graph that depicts the historical changes of frequency for a particular phrase (or word), the quality of graph is quite low. We therefore took use of 'ngramr' package (Carmody, 2015) in R programming language to extract the N-gram data and plotted new graphs based on these data. After that, all graphs were integrated into a page to yield a high-quality chart which is Figure 4 (the R script we wrote is provided to help those who would like to replicate it; please visit the link https://github.com/five-hills/punctuation).

6 Mulvey (2016) gave a detailed account of the relationship among punctuation marks and how they evolve in English and in other languages. Parkes (1993) is a good reference to the history of punctuation in western society.

Acknowledgements

This study is supported by the National Social Science Foundation of China (Grant No. 15CYY038). We appreciate the anonymous referee and editors whose valuable suggestions and questions helped improve this paper significantly. Our thanks go to Haitao Liu, Chunshan Xu, Jessie Nixon and Aengus Daly for their constructive suggestions and considerable help in revising the draft of this paper.

References

Baron, N.S. 2001. 'Comma and canaries: The role of punctuation in speech and writing.' *Language Sciences*, 23(1), 15–67.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
Biber, D. 1995. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.
Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Written and Spoken English*. London: Longman.
BNC (British National Corpus). 2007. The BNC Consortium. See http://www.natcorp.ox.ac.uk/. Online at < http://corpus.byu.edu/bnc/> (Accessed February 1, 2017).
Bohannon, J. 2010. 'Google opens books to new cultural studies.' *Science*, 330, 1600.
Bruthlaux, P. 1995. 'The rise and fall of the semicolon: English punctuation theory and English teaching practice.' *Applied Linguistics*, 16(1), 1–14.
Busà, G. M. 2014. *Introducing the Language of the News*. Routledge: London.
Carmody, S. 2015. 'Ngramr: Retrieve and plot Google N-gram data.' R package version 1.4.5.
Clauset, A., Shalizi, C. R. & Newman, M. E. 2009. 'Power-law distributions in empirical data.' *Annals of Applied Statistics*, 51(4), 661–703.
COCA (Corpus of Contemporary American English). 2014. Compiled by Mark Davies, Brigham Young University. Online at < http://corpus.byu.edu/COCA/> (Accessed February 1, 2017).
COHA (Corpus of Historical American English). 2014. Compiled by Mark Davies, Brigham Young University. Online at < http://corpus.byu.edu/COHA/> (Accessed February 1, 2017).
Crystal, D. 1985. 'How many millions? The statistics of English today.' *English Today*, 1, 7–9.
Denby, L. 2010. 'The language of Twitter: Linguistic innovation and character limitation in short messaging.' Online at <https://lewisdenby.files.wordpress.com/2010/06/the-language-of-twitter-linguistic-innovation-and-character-limitation-in-short-messaging.pdf> (Accessed February 1, 2017).
GloWbE (Corpus of Global Web-Based English). 2013. Compiled by Mark Davies, Brigham Young University. Online at <http://corpus.byu.edu/glowbe/> (Accessed February 1, 2017).
Google Books. 2010/2016. 'Google Books Ngram Viewer.' Online at <https://books.google.com/ngrams> (Accessed February 1, 2017).
Haussamen, B. 1994. 'The future of the English sentence.' *Visible Language*, 28(1), 4–25.
Huddleston, R. & Pullum, G. K. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Jones, B. 1996. *What's the Point? A (Computational) Theory of Punctuation?* PhD Diss. Edinburgh: University of Edinburgh.
Journalism BBC News style guide, 2018. 'Grammar, spelling and punctuation.' Online at <http://www.bbc.co.uk/academy/journalism/article/art20130702112133530> (Accessed July 22, 2018).
Kello, C. T., Brown, G, Ferrer–i–Cancho, R., Holden, J. G., Linkenkaer–Hansen, K. & Rhodes, T. 2010. 'Scaling laws in cognitive sciences.' *Trends in Cognitive Sciences*, 14 (5), 223–232.
Kelly, J. 1999. 'The secret of punctuation.' Online at <http://articles.chicagotribune.com/1999-05-12/features/9905120098_1_grammatical-task-semicolon-punctuation/2> (Accessed February 2, 2017).
Kirkpatrick, A. 2010. *The Routledge Handbook of World Englishes*. London: Routledege.
Ling, R. & Baron, N. 2007. 'Text messaging and IM: Linguistic comparison of American college data.' *Journal of Language and Social Psychology*, 26(3), 291–298.
Leimgruber, J. R. E. 2013. *Singapore English: Structure, Variation, and Usage*. Cambridge: Cambridge University Press.
Lewis, H. E. 1894. *The History of the English Paragraph*. PhD Diss. Chicago: University of Chicago.
Liberman, M. 2011 'Real trends in word and sentence length.' Online at <http://languagelog.ldc.upenn.edu/nll/?p=3534> (Accessed February 1, 2017).
Markwardt, A. H. 1942. *Introduction to the English Language*. New York: Oxford University Press.
Meyer, C. 1987. *A Linguistic Study of American Punctuation*. New York: Peter Lang.
Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team. 2011. 'Quantitative analysis of culture using millions of digitized books.' *Science*, 331, 176–182.
Mulvey, C. 2016. 'The English project's history of English punctuation.' *English Today*, 32(3), 45–51.
Nunberg, G. 1990. *The Linguistics of Punctuation*. Stanford, CA: CSLI.
Parkes, M. B. 1993. *Pause and Effect: An Introduction to the History of Punctuation in the West*. Berkeley: University of California Press
Partridge, E. 1953. *You Have a Point There: A Guide to Punctuation and its Allies*. London: Routledge.
Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
Schou, K. 2007. 'The syntactic status of English punctuation.' *English Studies*, 88(2), 195–216.
The Punctuation Guide. 2018. 'British versus American style.' Online at <http://www.thepunctuationguide.com/british-versus-american-style.html> (Accessed July 22, 2018).
Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.