Available online at www.sciencedirect.com

ScienceDirect

Lingua 262 (2021) 103155

Lingua

www.elsevier.com/locate/lingua

Lexical use and social class: A study on lexical richness, word length, and word class in spoken English

Yaqian Shi [a,1], Lei Lei [b,*]

[a] Huazhong University of Science and Technology, PR China
[b] Shanghai Jiao Tong University, PR China

Received 24 September 2020; revised 18 June 2021; accepted in revised form 20 June 2021; available online 27 August 2021

Abstract

Lexical use is an important indicator of an individual's social class, and previous studies found that speakers from different social classes show distinct features in their lexical use. However, those studies are more qualitative in nature and the findings are far from conclusive. To address these issues, this study examined the lexical features of the utterances produced by speakers from different social classes in terms of lexical richness, word length, and word class based on a large dataset of spontaneous utterances, i.e., Spoken British National Corpus 2014. The analysis yielded several interesting findings. First, speakers from the middle and upper-middle classes produce utterances of a higher lexical richness than those from the lower class. Second, individuals from all social classes tend to produce utterances of a lower lexical richness and with shorter words in spoken language than they do in writing, which indicates the spontaneous nature of spoken language. Third, speakers from the middle and upper middle classes have similar lexical features such as the more frequent use of derived *-ly* adverbs (particularly intensifiers), conjunctions, and prepositions. In contrast, those from the lower class use more negative words and first-person singular pronouns. Such differences are explained by factors closely related to the speakers' social class backgrounds.

Keywords: Lexical use; Social class; Lexical richness; Word length; Word class

1. INTRODUCTION

Social class refers to the social stratification in society where individuals are hierarchically classified into different social groups such as upper class, middle class, and lower class (Grant, 2001). It has been accepted as a multi-dimensional construct that may include a wide range of aspects, such as economic, social, and cultural ones (Block, 2015; Savage et al., 2013: 223). To be more specific, social class is comprised of a number of different dimensions such as occupation, place of residence, education, consumption patterns, and symbolic behaviours (Block, 2012, 2014). A person's social class is thus determined by a combination of these dimensions. Social class has long been a topic of interest in language variation research since it is one of the main constructs that organise our society

* Corresponding author at: 800 Dongchuan Road, Minhang, Shanghai, PR China.
E-mail addresses: leileicn@126.com, leileibama@outlook.com (L. Lei).
[1] Postal address: 1037 Luoyu Road, Wuhan, Hubei, PR China.

https://doi.org/10.1016/j.lingua.2021.103155
0016-7037/

(Block, 2014; Meyerhoff, 2006; Rampton, 2010; Snell, 2014). That is, social class is a fundamental part of our social structure, which is closely related with language use and evolution. A close investigation of social class differences in language use is hence crucial to our understanding of the nature of language variation, and even the nature of language (Guy, 2011).

Given the importance of social class in language variation, researchers have investigated how individuals from different social classes use language (e.g., Cheshire, 2005; Dickson and Hall-Lew, 2017; Eckert, 2000; Macaulay, 2002). For example, Labov (1966a, 1966b, 1972) investigated the relationship between social class and pronunciation. One interesting finding is that certain social classes were always linked to particular pronunciation features. For instance, the pronunciation of the post-vocalic [r] was more frequently used by individuals from higher social classes in the United States (Labov, 1966a). Another interesting finding is that speakers from the lower middle class pronounced more [r]-s than those from the upper middle class in formal contexts. The finding may be explained by the factor of upward social mobility (Labov, 1966b), which means the lower middle class aspire more to move upward on the social ladder and thus may linguistically imitate the upper middle class.

Following Labov, many other researchers have also explored how the speeches of different social classes vary at different linguistic levels, which provides more evidence to the social stratification of language use (e.g., Dickson and Hall-Lew, 2017; Mather, 2012; Trudgill, 1974). For example, Bernstein (1971, 1973, 1975) explored the relationship between social class and language use based on the language code theory. He found that students from the working class performed not as well as those from the middle or upper class in language-based subjects. Also, students from the working class tended to produce more informal and context-dependent utterances with shorter and less complex sentences (Bernstein, 1971). Such differences may be explained by the language code theory that was comprised of elaborated and restricted codes. Elaborated codes were characterised by longer, more complicated phrases and sentences while restricted codes were less formal with shorter phrases. Students from the working class may have access only to restricted codes while those from the middle or upper class may have access to both restricted and elaborated codes.

In addition, Bourdieu (1977, 1991) investigated the relationship between language and social class from the perspective of language habitus. Language habitus refers to the ability to use the possibilities provided by language in appropriate contexts, which plays an important role in language production (Bourdieu, 1977). Such a habitus is largely determined by individual experiences and living conditions, such as material conditions, upbringing, and social relation (Grenfell, 2014). Members of different social classes have different experiences and thus different language habitus, which in turn brings about language differences.

More recently, many researchers have reconsidered the social meaning of language variation and attempted to increase our understanding on how meaning is conveyed by language use and other semiotic resources (e.g., Blommaert, 2010; Canagarajah, 2013; Pennycook, 2012). In this line of research, language variation is considered as a social semiotic system that includes many social concerns in a given community (Eckert, 2012), which may include a wide range of sociocultural information such as stances, individual/-group styles, and identities (Eckert, 2008; Snell, 2018). Thus, the meanings of language variables are not specified and specific meanings are only gained in context (Eckert, 2012). For example, Snell (2018) investigated the personal pronouns used in right dislocation and local dialect *howay* in the speech of two groups of children, i.e., those from the working class and those from the middle class. *Howay* refers to a discourse-pragmatic feature that is unique to northeast England. It means “come on” and usually functions as a directive (Snell, 2017). However, its specific social and pragmatic meaning is determined by the context where it occurs. She found that children from the working class used personal pronouns and *howay* more frequently to build their stance of opposition and negative evaluation.

One issue in language variation research is that most previous studies examined language variation concerning phonological, morphological, or syntactic features, while less attention has been paid to lexical ones. One reason for this is that lexical use is closely related to the topics of spontaneous utterances. Speakers from different social classes may differ in lexical use because they engage in conversations of different topics. In contrast, phonetic, morphological, or syntactic features are less affected by the topics of utterances. Hence, the potential occurrence of lexical items in a conversation is different from that of phonetic, morphological, and syntactic ones. Another reason is that words are more open-ended in scope and hence researchers need to collect a large dataset for the quantitative analyses of lexical use (Upton et al., 1994). In addition, speakers in the dataset need to be labelled with various social tags such as social class. The unavailability of a large-sized corpus of spontaneously produced spoken data with social class annotations has been a possible challenge for lexical analyses. The aforementioned two reasons may explain why lexical variation is difficult to study and why previous studies mostly examined the social stratification of particular words such as personal pronouns and *howay* (e.g., Snell, 2018). Hence, what the lexical features of different social classes are and how they differ from one another have not been fully examined.

To fill the aforementioned research gaps, the present study aims to quantitatively explore the social class differences of British English at the lexical level. More specifically, the purpose of the study is three-fold. First, it intends to examine the lexical use by speakers from different social classes based on indices such as lexical richness, word length, and word class. Second, it aims to explore whether difference exists in the indices between speakers from different social classes. Third, if any differences are found, possible reasons for the differences will be explored.

2. LITERATURE REVIEW

In this section, we first review the studies on lexical use and social class. Then, we describe the lexical indices used in this study. Finally, we introduce the research questions that this study attempts to answer.

2.1. Lexical use and social class

Vocabulary is central to communication and a speaker's lexicon is an important marker of his/her social class (Hudson, 1996: 43). This is because speakers' lexical resource is influenced by their social backgrounds and life experience such as education, occupation, income, and interpersonal relationship (Hoff, 2006; Snow, 1999). Speakers may actively accumulate and use lexical alternatives to make more subtle social distinctions when they grow up (Jaspal, 2009; Schieffelin and Ochi, 1986).

Given its importance in social class, researchers have qualitatively and quantitatively examined how speakers use words that are particularly associated to their social class. For example, in the early 1950s, Mitford (1956) and Ross (1954) qualitatively analysed and identified pairs of words that were typically used by individuals of different social classes, i.e., the upper-class terms (U words) and the non-upper-class terms (non-U words). In a follow-up study, Fox (2004) found that some pairs of the U words and the non-U words, such as *serviette* vs. *napkin*, *sitting room* vs. *lounge*, and *pudding* vs. *sweet* were still applicable after half a century.

Researchers have also quantitatively explored the relationship between lexical use and social class. For example, Bernstein (1971) investigated the use of vocabulary across different social classes covering a wide range of word classes such as nouns, pronouns, adjectives, and adverbs. He found that speakers from higher social classes tended to use more adjectives while those from lower classes used more personal pronouns. In addition, he also found that speakers from lower classes tended to produce shorter utterances and use non-standard words compared with those from higher classes. However, Bernstein's (1971) study was limited because of the small size of the corpus used due to constraints of computer technology at that time. Specifically, he used a speech sample of less than 9000 words for his study. The small sample size might have prevented him from fully uncovering the differences in lexical use among speakers from different social classes.

Other researchers also explored how the speeches of different social classes varied in the use of some specific words or word classes (e.g., Aliakbari et al., 2012; Degaetano-Ortlieb, 2018; Kacewicz et al., 2014; Macaulay, 2002; Yusuf et al., 2019). For example, Macaulay (2002) investigated the use of adverbs across different social classes. He found that middle class speakers used more derived *-ly* adverbs than working class speakers did.

To summarise, most of the previous studies were based on small-sized corpus data or focused on some specific words and their findings were hence limited regarding the relationship between lexical use and social class. It is thus necessary to quantitatively explore this relationship and examine the lexical features of the speeches of different social classes using a large corpus.

2.2. Lexical indices

The lexical indices used in the study include lexical richness, word length, and word class. The reason for using these three indices is that they are distinctive features of words or texts, which may change as a function of various demographical factors such as gender (Koppen et al., 2019; Mikros, 2012; Singh, 2001), age, and profession (Zhang, 2014). Therefore, we hypothesise that they may also serve as indices of lexical use by speakers from different social classes (e.g., Kacewicz et al., 2014). Meaning is also an important property of words. However, it is difficult to quantitatively study the meaning of a word since it may have multiple meanings and the meaning of a word is largely determined by the context in which it is used. Thus, we do not consider the factor of meaning in this study.

Lexical richness refers to the number of unique words used in a text (Daller et al., 2003). It is a significant language feature since a speaker's lexicon is integral to every aspect of the person's language knowledge (Daller et al., 2007: 1). Lexical richness has hence been widely employed in studies regarding language variation (e.g., Singh, 2001; Zhang, 2014). For example, lexical richness was found to be significantly correlated with some demographical and socioeconomic variables such as age, level of education, and profession (Zhang, 2014).

Word length is defined as the number of the constituting elements of a word (Grzybek, 2007). As one property of a word, word length is helpful for the comparison of texts of different authors and genders (Lian and Li, 2019; Mikros, 2012). For example, Mikros (2012) conducted authorship attribution and gender identification in Greek blogs based on a variety of textual features, of which word length was a useful one for the identification of authorship and gender. One way to calculate word length is to count the number of letters. A letter is a written sign representing one of the sounds in a language and the number of letters has been widely used as a measurement of word length due to its simplicity (Fan, 2013; Lei and Liu, 2014; New et al., 2006). Another way to calculate word length is to count the number of syllables. A syllable is the smallest unit that individuals can produce in one pulse of breath, which usually contains a vowel or a vowel plus one or more consonants (Nweke, 2013). Controversy exists over whether the number of letters or syllables could serve as a better index of word length. For example, the number of syllables has recently been considered as a better proxy of word length since syllables are more immediate constituents of the word (Grzybek, 2007). Due to the controversy, we will use both indices as measures of word length in this study.

Word class is a set of words that display the same formal properties (Altenberg and Vago, 2010). Similar to word length, word class is also a property of a word that is useful for text characterisation and comparisons (Stamatatos, 2009; Tabata, 2002). Seven major word classes, i.e., nouns, verbs, adjectives, adverbs, pronouns, conjunctions, and prepositions, will be included in the analysis of this study. The reason is that many of them such as adjectives, pronouns, adverbs, and prepositions have been found to be important indicators of an individual's social class, gender, and personality (e.g., Bernstein, 1971; Hessner and Gawlitzek 2017; Macaulay, 1995, 2002; Kacewicz et al., 2014; Pennebaker, 2011). For example, Bernstein (1971) found that middle class speakers of English tended to use some prepositions and uncommon conjunctions (i.e., all conjunctions other than *and*, *so*, *or*, *because*, *also*, *then*, and *like*) more frequently than lower class speakers. Macaulay (1995) examined the use of adjectives and found middle class English speakers used a higher proportion of adjectives than lower class speakers. In a follow-up study, Macaulay (2002) investigated the use of adverbs in different social classes. He found that speakers from the middle class used more derived *-ly* adverbs than those from the working class

did. Similar class differences also existed in the use of words in other word classes such as pronouns. Speakers from the higher social class used more personal pronouns *we* and *you* while those from the lower class used more frequently *I* (Kacewicz et al., 2014). Based on such findings from the previous research, we will include adjectives, adverbs, prepositions, conjunctions, pronouns, nouns, and verbs in this study. Although nouns and verbs have been investigated less frequently in the previous relevant studies, they are included in the study since they are content words that play a major role in communication and they account for a large proportion of words in a language (Liang and Liu, 2013).

In addition, lemmas that are used more frequently by a certain social class will be extracted to examine the lexical features of each social class. We do this for two reasons. First, we follow the practice of previous studies (e.g., Macaulay, 1995, 2002; Kacewicz et al., 2014). For example, Macaulay (2002) identified that intensifiers such as *exactly*, *quite*, and *very* were used more frequently by speakers from higher social classes. Second, those more frequent lemmas are probably important indicators of the features of different social classes.

In summary, lexical indices such as lexical richness, word length, and word class are important properties of a word or a text that may be affected by demographical factors. In this study, we aim to identify and describe the lexical features of different social classes in terms of the three lexical indices. Based on previous research findings, we have two expectations. First, speakers from higher social classes such as the middle and upper middle classes may produce utterances with a higher lexical richness and longer words than those from the lower class do. Second, speakers from higher social classes may use some word classes, such as adverbs, conjunctions, and prepositions, more frequently.

2.3. Research questions

(1) Is there any difference in lexical use between social classes in terms of lexical indices such as lexical richness, word length, and word class? If yes, what are the differences?
(2) Is there any lemma that is more frequently used by a certain class?
(3) Why do speakers from various social classes differ in their lexical use?

3. METHODOLOGY

In this section, we first describe the corpus data used in the study and then introduce the procedures of data processing.

3.1. Spoken BNC2014

The Spoken British National Corpus 2014 (hereafter Spoken BNC2014), a sub-corpus of the British National Corpus 2014, is a publicly accessible corpus of present-day spoken British English (Love et al., 2017). It consists of 1,251 transcribed texts collected from 2012 to 2016, with a total of more than 11 million words.

The Spoken BNC2014 was used for the present study for two reasons. First, the corpus consisted of data that were spontaneously produced by English speakers in natural contexts. To be specific, the corpus contained spontaneous conversations produced and recorded by voluntary speakers on their smartphones. These utterances were daily-life conversations with friends, family members, or colleagues, which covered a wide range of topics such as food, film, sport, shopping, news, birthday, and neighbourhood. More importantly, speakers of the utterances were from different social classes. For example, of the five speakers in a conversation about their daily lives (Text ID: 52k7), one was from the upper middle class, one was from the middle class, and three were from the lower class. In other words, although the speakers were from different social classes, they participated and talked on the same or similar topics. The corpus is probably the largest publicly available set of such spontaneously spoken English data.

Second, the corpus contained indices of the speakers' social class based on their occupations. An important feature of the corpus was that it contained the demographic metadata of the speakers such as their age, occupation, and nationality. The speakers' socioeconomic status was hence estimated from their occupation based on the demographic classification system of National Readership Survey's Social Grade (see Table 1 for details). The system has been widely used for the classification of social classes over half a century (Collis, 2009).

Table 1
National Readership Survey Social Grade classifications.

Social grade code	Description
A	Higher managerial, administrative and professional
B	Intermediate managerial, administrative and professional
C1	Supervisory, clerical and junior managerial, administrative and professional
C2	Skilled manual workers
D	Semi-skilled and unskilled manual workers
E	State pensioners, casual and lowest grade workers, unemployed with state benefits only

Therefore, the spontaneous nature of the data in the corpus with its social class annotation made it particularly appropriate for its use in the present study.

For comparison purposes, we classified the six codes of social classes into three broad social classes, i.e., the upper middle class (A), the middle class (B, C1, C2), and the lower class (D, E). The upper middle class usually consisted of higher status members of the middle class who were well-educated professionals with comfortable incomes such as top executives, professors, and engineers (Thompson and Hickey, 2010). The middle class included individuals who had relatively secure jobs with a pay significantly above the poverty line such as semi-professionals and craftsmen (Thompson and Hickey, 2010). Last, the lower class was comprised of individuals who were employed in lower-paying jobs with little economic security. The statistics of the three social classes were presented in Table 2. A point worth noting here is that speakers who were not tagged with social class codes were not included in the analyses.

3.2. Data processing and analysis

We processed and analysed the data in the following steps. First, we calculated the lexical richness of the utterances produced by each social class. A traditional popular method to measure lexical richness is the Type-Token Ratio (TTR), which is the ratio of different words (types) to the total number of words (tokens) in a text (Malvern and Richards, 2013). For example, in the following sample sentence, the total number of words/tokens in the sentence was 10 and the number of different words/types was 8. Thus, its TTR was 80 (8/10*100 = 80). It meant that 80% of the running words were different words or different types of words.

"David likes playing football while his brother likes playing basketball."

However, the TTR is limited in its dependence on text length (McCarthy and Jarvis, 2007; Shi and Lei, 2020). That is, as a text becomes longer, the chance of new words entering into the text becomes increasingly lower (Daller et al., 2007), which results in longer texts with a lower lexical richness. As a result, when two texts are different in length, it is difficult and not valid/reliable to compare the lexical richness of a shorter text with that of a longer one by their raw counts. A more sophisticated and also more valid/reliable alternative to the TTR, i.e., the Standard Type-Token Ratio (STTR) that computes the TTR based on every 1,000 words, is regarded as a more valid indicator of lexical richness when texts of different lengths are compared. Given that the number of words in the Spoken BNC2014 varied substantially across the three social classes (see Table 2), we adopted the STTR as a proxy of lexical richness. Specifically, the data of each social class were first divided into subsamples of the same length, each with 1000 tokens. Then, the lexical richness of each subsample was calculated in terms of the STTR. Finally, the STTRs of the subsamples were averaged, which was used as the final value of lexical richness.

Second, we examined the word length across the three social classes. We calculated the length of each word in each social class in terms of the number of letters. For example, the word *class* had five letters, and thus its word length was 5. We included filled pauses such as *um* and *ah* in the analyses for the reason that they were produced in spoken language, and were characteristics of spontaneous utterances. In addition, the number of the filled pauses accounted for a very small proportion of the words of the corpus (approximately 2.7%). With such a small proportion, they may not affect the results of the statistical analyses. Then, we calculated the mean word length of each social class and calculated the raw frequency of each word length in each social class. For valid comparisons of word lengths across the three social classes, we normalised the frequency by dividing the raw frequency in a social class by the total number of words in that social class and then multiplying one million. The normalised frequency meant the frequency of occurrence per million words of a certain word length in each social class. For example, the raw frequency of words of two letters in the upper middle class was 449,968 and the total number of words in the upper middle class was 1,939,599. Thus, the normalised frequency of words of two letters in the upper middle class was 231,995 (449,968/1,939,599 * 1,000,000 ≈ 231,995). In addition, we also calculated the length of each word in each social class in terms of the number of syllables. The number of syllables in a word was usually measured by its number of vowels (Roach, 2009). For example, the word length of *young* was one since it had only one vowel and that of *architecture* was four since it had four vowels. We used *syllable_sum()*, a function in the package *qdap* of the R language, to count the number of syllables in a word, i.e., the syllable-based word length. Last, we conducted ANOVAs to examine the word length differences across the three social classes.

Third, we examined word classes across the three social classes. We investigated seven major word classes, i.e., nouns, verbs, adjectives, adverbs, pronouns, prepositions, and conjunctions. We counted the number of occurrences of a word class per speaker in each social class and made a normalised frequency list (see Table 3). The normalised frequency meant the frequency of occurrence per 1000 words of a word class by each speaker. Then, we performed ANOVAs to examine the normalised frequencies across the three social classes. One of the most common statistical tests used for comparing word frequencies was the Chi-squared test (Rayson et al., 2004). However, we conducted ANOVAs rather than Chi-squared tests since previous studies showed that it may be problematic to use the Chi-squared test to compare word frequencies in texts or corpora since the Chi-squared test was based

Table 2
Statistics of the three social classes.

Social class	Social grade codes	No. of words
upper middle class	A	1,939,599
middle class	B, C1, C2	4,606,231
lower class	D, E	4,489,888
Total		11,035,718

Table 3
The normalised frequency of VERBs.

Speakers	Upper middle class	Middle class	Lower class
speaker 1	233.09	251.95	229.86
speaker 2	228.32	241.47	236.53
. . .	. . .	. . .	. . .
speaker n	208.67	234.56	232.44

on the assumption that all the words in a corpus were statistically independent (Kilgarriff, 2005; Lijffijt et al., 2016; Paquot and Bestgen, 2009). A better alternative is to use tests that take individual texts/speakers as observations (Gilquin and Granger, 2014) such as the t-test, the ANOVA, the Wilcoxon rank-sum test, or the bootstrap test (Gablasova et al., 2017; Lijffijt et al., 2016). Given that the frequencies were compared across the three social classes in our study, we performed ANOVAs to examine whether the use of a word/word class was significantly different across the three social classes. For words/word classes with significant results, post hoc tests were performed to identify between which specific social classes the differences were significant.

Fourth, we examined lemmas across the three social classes. For the identification of lemma candidates, we set a minimum frequency of 28.57 times per million words in the corpus. The minimum frequency was high enough to ensure that the identified lemmas were of high frequency (Coxhead, 2000; Lei and Liu, 2016)). In addition, the minimum frequency was not too high to exclude any significant lemmas since it was found that 93% of the lemmas in the corpus were included in the analyses (10287976/11035718=0.93). Based on this frequency threshold, we obtained a total of 1597 unique lemmas. Then, with the same method used for the analysis of word class, we performed ANOVAs to identify the lemmas that significantly differed across the three social classes. For lemmas with a significant difference, we conducted pairwise post hoc tests to examine between which specific social classes the differences were significant and to identify lemmas that were used by a certain class more frequently.

Last, we manually checked the identified lemmas and qualitatively analysed the results. We manually identified derived *-ly* adverbs and negative words in these lemmas and examined their distributions across the three social classes. Concerning the determination of negative words, we used the following three steps. First, we focused on open-class words, i.e., nouns, verbs, adjectives, and adverbs since they were usually used to express one's feelings (Pérez-García and Sánchez, 2020). Second, we conducted a pilot study and developed three criteria to determine if a word was negative: (1) it has a negative connotation (e.g., *bad*, *die*, and *dead*); (2) it is related to undesirable things or violent behaviours (e.g., *problem, mistake*, *war, and shoot*); (3) it contains sense of negation (e.g., *fail* and *not*); (4) it is associated with difficulty, regret, or sadness (e.g., *struggle*, *sorry,* and *sad*). Then, two raters, who are doctoral students of applied linguistics, were invited to determine whether a word was negative. The result of the Cohen's kappa showed a good inter-rater agreement (Cohen's kappa = 0.855, p = 0.000). For words with different judgements, we discussed them until we reached a full agreement. For example, the word *shut* was judged as negative by one rater but neutral by the other rater. Based on a close read of the concordance lines in the corpus, we agreed that *shut* was negative in sense because it was mostly used in the corpus as part of the phrase *shut up*.

The automatic part of the work mentioned above was completed with homemade scripts with the programming language R.

4. RESULTS

In this section, we report the results concerning cross-class variations in terms of lexical richness, word length, word class, and lemmas used more frequently.

4.1. Lexical richness

The results of lexical richness of the three social classes were presented in Table 4. The lexical richness of all the three social classes was approximately 30, which meant approximately thirty different types or unique words occurred per 100 words. This result was in line with the findings of previous studies. For example, Baker (2006: 52) reported that the lexical richness of written texts (FLOB corpus of British English) was 45.53 while that of spoken texts (Spoken British National Corpus 1994) was 32.96. Our results, together with Baker's (2006), provided evidence that the lexical richness of spoken language was relatively low compared to the lexical richness of written language (Strömqvist et al., 2002).

Table 4
Lexical richness across the three social classes.

Social class	Lexical richness (STTR)
upper middle class	29.9
middle class	30.0
lower class	29.2

The results of ANOVA showed that lexical richness was significantly different across the three social classes ($F(2) = 122$, $p < 0.001$). In addition, the results of the post hoc tests showed that middle and upper middle classes produced utterances with a higher lexical richness than the lower class did.

4.2. Word length

The descriptive statistics of word length in the three social classes were presented in Table 5 and Fig. 1. The means of word length across the three social classes were all approximately 3.7 letters per word with no significant difference ($F(2) = 2.405$, $p = 0.0903$). Words with two, three, and four letters accounted for a large proportion in the speech data (see Fig. 1). The mean word length of our spoken utterance data was lower than that of written texts. For example, Fan (2011) reported that the word length of written texts such as those in the Lancaster-Oslo/Bergen (LOB) corpus was 4.4 letters per word. The LOB corpus was a one-million-word written corpus. It consisted of British English texts from 15 text categories such as fictions and essays. This result indicated that all speakers, regardless of their social class, used shorter words in daily communication.

Some may be concerned with the validity of using letters as a proxy of word length in spoken language. To address the concern, we also calculated the word length of the three social classes with syllables. The results were presented in Table 6 below. As shown in the table, the mean word length measured with syllables across the three social classes was approximately 1.2 syllables per word with no significant difference ($F(2) = 1.721$, $p = 0.18$). The result was similar to that of word length measured based on letters. It further confirmed that individuals, regardless of their social class, tended to use shorter words or words with fewer syllables in daily communication.

Table 5
Descriptive statistics of word length (measured with letters) across the three social classes.

Social class	Min	Max	S.D.	Mean word length
upper middle class	1	25	2.104	3.704
middle class	1	44	2.107	3.702
lower class	1	51	2.129	3.705

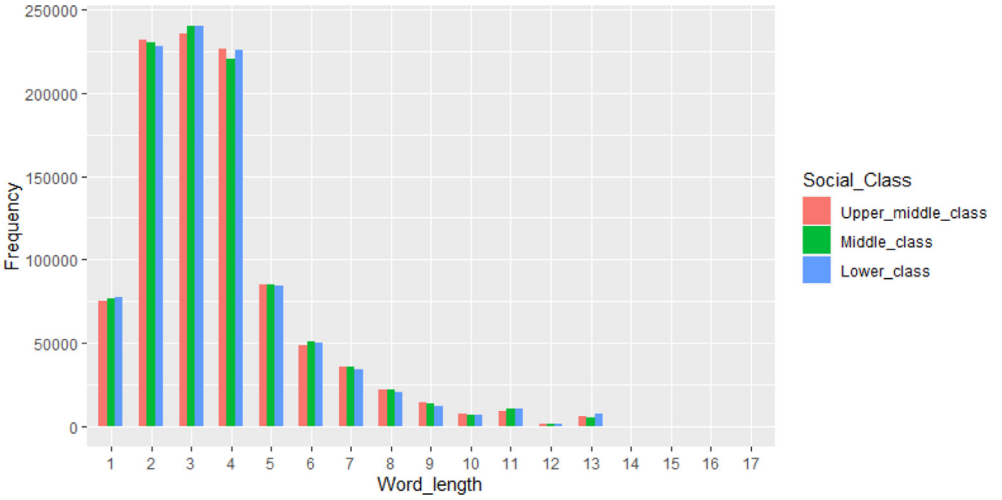

Fig. 1. Word lengths across the three social classes.

Table 6
Descriptive statistics of word length (measured with syllables) across the three social classes.

Social class	Min	Max	S.D.	Mean word length
upper middle class	1	10	0.625	1.255
middle class	1	11	0.630	1.259
lower class	1	17	0.620	1.252

4.3. Word class

The results of the use of word class across the three social classes were presented in Table 7. The results of ANOVAs showed that there was no significant difference in the use of words of all word classes except for pronouns, prepositions, and conjunctions across the three social classes (pronouns: F (2) = 3.59, $p = 0.028$; prepositions: F (2) = 12.42, $p < 0.001$; conjunctions: F (2) = 6.43, $p = 0.002$). One possible reason for the results was that nouns, verbs, adjectives, and adverbs were open-class words, which were large in scope and easy to expand (Kestemont, 2014). Speakers had many options to choose in these word classes in communication, and hence the occurrence or frequency of these words was not significantly different across the social classes.

The pairwise post hoc tests were conducted to identify the word classes that were used more frequently by a certain social class (see Table 7). Middle and upper middle classes used prepositions and conjunctions more frequently than the lower class, and the middle class used pronouns more frequently than the upper middle class. In contrast, the lower class used only pronouns more frequently than the upper middle class, a narrower range of more frequent word classes. Such results partly supported previous findings such as Bernstein (1971) and Aliakbari et al. (2012).

4.4. More frequently used lemmas

We also explored the more frequently used lemmas by each social class. The results summarised in Table 8 presented some interesting findings.

First, middle and upper middle classes used a larger number of derived *-ly* adverbs. To be specific, they used *-ly* adverbs more frequently, such as *actually*, *really*, *possibly*, and *slightly* (underlined in the adverb row of Table 8), and some of the derived *-ly* adverbs were intensifiers such as *really*, *actually*, and *particularly*. The finding was in line with those of the previous studies such as Macaulay (2002).

Second, middle and upper middle classes used more prepositions and conjunctions. Specifically, the upper middle class used four prepositions (i.e., *of, by, across*, and *within*) and three conjunctions (i.e., *that*, *or*, and *unless*) more frequently. The middle class used three prepositions (i.e., *to*, *for*, and *about*) and one conjunction (i.e., *even*) more frequently.

Third, the lower class frequently used a larger number of negative words such as *fuck*, *bad*, and *dead* (bolded in Table 8). For example, they used twelve verbs more frequently, half of which were negative verbs such as *shit*, *die*, and *hate*. In contrast, middle and upper middle classes only used three negative verbs more frequently (i.e., *fail*, *struggle*, and *lie*).

Last, the lower class used the first-person singular *I* and *my* more frequently, which was partly consistent with previous research findings such as those reported by Kacewicz et al. (2014), showing that individuals from a lower class preferred the first-person singular *I* over the plural *we*.

Table 7
Statistical results on word classes across the three social classes.

Word class	Raw frequency			F	*p*	Post hoc test	
	upper middle class (umc)	middle class (mc)	lower class (lc)			Paired social classes	*p*
verb	439,632	1,070,624	1,023,702	2.67	0.070	mc-lc	0.148
						umc-lc	0.776
						umc-mc	0.111
pronoun	323,932	786,554	772,725	3.59	0.028	mc-lc	0.999
						umc-lc	0.030
						umc-mc	0.0417
adverb	233,763	570,452	541,912	2.56	0.078	mc-lc	0.072
						umc-lc	0.977
						umc-mc	0.342
noun	228,989	552,814	535,699	1.23	0.293	mc-lc	0.342
						umc-lc	0.501
						umc-mc	0.998
adjective	161,259	383,708	367,277	2.11	0.122	mc-lc	0.825
						umc-lc	0.218
						umc-mc	0.105
preposition	144,525	344,128	317,685	12.42	<0.001	mc-lc	<0.001
						umc-lc	<0.001
						umc-mc	0.885
conjunction	119,627	275,818	267,291	6.43	0.002	mc-lc	0.012
						umc-lc	0.009
						umc-mc	0.725

Table 8
Lemmas used more frequently by the three social classes.

Word class	Upper middle class	Middle class	Lower class
Noun	*thing, time, so, sort, kind, work, job,* ***problem****, part, moment, language, sense, business, line, summer, difference, programme, fire, system, issue, board, sign, option, winter, project, government, education, meeting, feeling, screen, study, vegetable, example, rent, teaching, process, supermarket, roof, career, sentence, benefit*	*thing, kind, money, work, friend, job, load, food, holiday, weekend, Friday, class, cake, Saturday, Sunday, business, lunch, message, office, wedding, date, potato, issue, gym, restaurant, information, skill, talk, soup, girlfriend, decision, Germany, value,* ***mistake****, teaching, process*	*dad, girl, game, film, bus, music,* ***war****,* ***fuck****, club, horse, finger, hall, episode, tablet, queen*
Verb	*think, mean, work, need, find, cut, invite, burn, arrive, press,* ***fail****,* ***struggle****, argue*	*know, go, think, work, need, might, meet,* ***lie****, drive, spend, cook, travel, mortgage*	*like, watch, play, wait,* ***shit****,* ***die****,* ***hate****,* ***drop****,* ***shut****, kick,* ***shoot****, shout*
Adjective	*this, these, right, those, different, another,* ***sorry****, great, whole, interesting,* ***difficult****, certain, extra,* ***low****,* ***worried****, warm, exciting, used, thick, aware*	*another,* ***sorry****,* ***difficult****, sweet,* ***sad****, Chinese, common, driving*	***bad****,* ***dead****, asleep, gay*
Adverb	*so, actually, quite, maybe, more, together,* ***less****, slightly, particularly, clearly, therefore,*	*so, really, actually, quite, probably, also, later, possibly*	***not****, though*
Pronoun	*something*	*their, everything*	*I, she, my, one*
Preposition	*of, by, across, within*	*to, for, about*	/
Conjunction	*that, or, unless*	*even*	/

The lemmas underlined in the adverb row are the derived *-ly* adverbs; those bolded are negative words.

5. DISCUSSION

The study explored cross-social class differences in lexical use in terms of lexical richness, word length, word classes, and more frequently used lemmas. The results revealed lexical features that were closely associated with speakers' social classes. In other words, social class directly affected the social and life experience of an individual such as their education, career, income, and interpersonal relationship, which in turn affected their lexical resources (Hoff, 2006; Snow, 1999). Below, we discuss the major differences the results of our study have shown.

First, the three social classes showed significant difference in terms of lexical richness. Middle and upper middle classes tended to produce utterances with a higher lexical richness than the lower class. The finding partly met our first expectation that speakers from higher social classes might produce utterances with a higher lexical richness. The reason for the finding was probably that speakers from middle and upper middle classes generally received a better education than those from the lower class. As a result, the middle and upper middle classes might have a larger vocabulary and thus tended to use a wider variety of words.

Second, compared with those of writing, the lexical richness and the word length of the utterances produced by the three social classes were lower and shorter. That is, all speakers from the three social classes tended to produce utterances with a lower lexical richness and shorter words than those in writing, which probably disclosed the nature of spoken language and could be explained as follows. One possible reason was the time pressure in spontaneous speech. That is, speech was produced in real time and the speakers might be limited in time to plan or select words. The other possible reason was the constraint of working memory of the human beings (Cowan, 2016). That is, regardless of their social class, all speakers have a limited working memory, i.e., a limited ability to temporarily store information (Baddeley, 1983). Under both pressures, the speakers tended to choose shorter and simpler words from a restricted vocabulary to reduce the cognitive burden in communication (Lei and Wen, 2020). For example, they tended to use more two-letter words than longer words in spoken language such as *no*, *eh* etc., which hardly occurred in written language. This might also explain part of the differences between written and spoken languages.

Still another reason was attributed to the differences of stylistic properties between spoken and written corpora. Written corpora, such as FLOB, consisted of texts such as fictions, news articles, biographies, and stories composed by professionals (Hundt et al., 1998). However, spoken corpora, such as the Spoken BNC2014 used in this study, contained largely natural conversations in daily life produced by speakers from different social classes. The fact that the language users in the written corpus were less socially diverse and more linguistically professional may lead to texts with a higher lexical richness and longer words.

Third, the middle and upper middle classes shared similar lexical features. Both classes tended to frequently use a wider range of lemmas such as derived *-ly* adverbs (particularly intensifiers), prepositions, and conjunctions. The finding was in line with our second expectation that higher social classes might use certain classes of words more often. Intensifiers were linguistic terms that were used to modify words or phrases, which strengthened the speakers' positions and attitudes toward what they had said (Núñez Pertejo and Palacios Martínez, 2014). Hence, intensifiers were used to express intensity and emphasis, which enhanced the speakers' commitment

to the propositions (Labov, 1984). The higher frequency of intensifiers in our data suggested that speakers from the middle and upper middle classes were more likely than those from the lower class to make emphatic statements, clarify their opinions, and show their attitudes. Such linguistic tendencies may be associated with their social class in which they were brought up and socialised into society. The middle and upper middle classes generally possessed a better education and held more prestigious and secure occupational/professional positions such as top executives, managers, engineers, and doctors (Stephens et al., 2014). In other words, they held positions which gave them more power, prestige, and authority (Bullock and Lott, 2010; Stephens et al., 2014). As a result, they may be more confident and prefer to explicitly express and emphasise their opinions, ideas, and attitudes in communication. In example (1) below, two speakers were discussing what cakes they should make for their parents. One speaker from the upper middle class recommended a walnut coffee cake, and used intensifiers *particularly* and *really* (in bold font) to emphasise that it was delicious.

(1) "well we were gonna do for for my parents we were gonna do um walnut sort of coffee walnut which uh which I think is gonna be **particularly particularly** tasty. . .with the with the um a walnut cake with the with like this the coffee. . .I used to get those you know those little walnut things you used to get in a rectangular shape? They were **really** delicious too" (Text ID: S2AX, Utterance No. 813–817)

Prepositions were used to signal spatial or temporal relations between two linguistic elements, while conjunctions were used to connect words, phrases, or clauses (Huddleston and Pullum, 2002). With a preposition or conjunction between two linguistic elements, the internal logical relations of the phrases or sentences became more explicit (Aliakbari et al., 2012). Thus, the more frequent use of prepositions or conjunctions may indicate more explicitness in utterances. As discussed earlier, speakers from the middle and upper middle classes may possess a better education and hence have better language skills and were more likely to produce more grammatically correct utterances with prepositions and conjunctions. In addition, they were more likely to play the role of leaders and experts in society (Martin et al., 2017). As a result, they may be more confident and hence tended to more explicitly express their ideas and opinions (Richardson et al., 2012). In example (2) below, for example, the speaker, a person from the upper middle class used four prepositions and four conjunctions in one utterance (bolded for emphasis purposes). These prepositions and conjunctions not only offered information such as the location (e.g., *at the conference*), but also explicitly expressed the logical relations between them. The frequent use of prepositions and conjunctions hence made his statement seemingly more explicit.

(2) "it's like **if** you speak **at** the conference they give you **and** they it was like you've got to pick a gift **from** this bag **of** stuff **but** all **of** them were reindeer things **and** like" (Text ID: S23A; Utterance No. 3756).

Fourth, the lower class was characterised by the more frequent use of the first-person singular *I* and *my* as well as negative words. Pronouns were words used to refer to someone or something that had been mentioned earlier, which were important language features that indicated the speakers' focus of attention (Kacewicz et al., 2014). The use of the third-person singular/plural (e.g., *he*, *she*, *they*) indicated that the speaker's attention was on others while the use of first-person singular pronouns was on ourselves as distinct entities (Kacewicz et al., 2014; Zimmermann et al., 2013). When a person felt insecure and self-aware, he/she tended to focus more on himself/herself (e.g., their own feelings and behaviours) and thus used more first-person singular pronouns (Davis and Brock, 1975; Duval and Wicklund, 1972; Kacewicz et al., 2014). Speakers from the lower class may be less educated and usually engaged in physical work (Simpson et al., 2016). In such a working environment, they not only had less power or authority, but also had little economic security. As a result, they may feel insecure and paid more attention to themselves and used the first-person singular *I* and *my* more often in daily conversation. For instance, in example (3), the speaker, a person from the lower class frequently used *I* to express his feelings about his own accents.

(3) "**I** wish that's a talent **I** wish **I** had to be really good at accents. . .**I** wish **I** could be good at accents like. . .same **I** can do like a certain like line like. . ." (Text Id: S4C2; Utterance No. 253–257).

Last, the lower class probably used more negative words for two reasons. On the one hand, in a less favourable working and life environment, speakers from the lower class may have more negative emotions (Gallo and Matthews, 2003; Stellar et al., 2012). On the other hand, they may prefer to directly express their negative emotions since emotional communication was perceived as acceptable and appropriate in the lower class (Gist, 2017). Therefore, they may use more negative words to directly express their unpleasant feelings in daily conversations. In example (4) below, for example, when the speaker, who was from the lower class, complained of the tinny sound of the broken TV, she used four negative words (bolded) to express her unpleasant feelings.

(4) "it was so tiny it was **awful** absolutely **terrible** and this is **bad** enough sort of **cringes** when he comes in so you couldn't really watch anything on it" (Text Id: S3JF; Utterance No. 713).

6. CONCLUSION

Lexical use is an important indicator of an individual's social class. Previous studies found that speakers from different social classes showed distinct features in their lexical use. However, those studies are more qualitative in nature and the findings are far from

conclusive. To address these research inadequacies, this study used a more quantitative approach to examine the relationship between lexical use and social class with three research questions. The first question is what differences in lexical use exist between different social classes in terms of three lexical indices, i.e., lexical richness, word length, and word class. Results show that the middle and upper middle classes produce utterances with a higher lexical richness than the lower class, while no significant difference is found in word length. As for word classes, the middle and upper middle classes use more conjunctions and prepositions. In contrast, lower class speakers use more first-person singular pronouns. The second research question is what lemmas are more frequently used by a certain class. Results show that the middle and upper middle classes use more derived *-ly* adverbs (particularly intensifiers), while lower class speakers use more negative words. The third research question is why the three social classes show such differences in lexical use. The lexical differences are explained by factors closely related to the speakers' social backgrounds such as occupation, education, personal life, and social experience.

To the best of our knowledge, this is probably the first study that explored the lexical features of different social classes based on a large dataset of spontaneous utterances. It provides more evidence for language variation in lexical use due to social class differences. Similar to pronunciation and grammar, lexis is also significantly correlated with social class, and speakers from different social classes use words in a way specific to their social class.

The study may be limited in the corpus data used, which include utterances of only one language (i.e., British English) and one register (i.e., spoken data). It would be of interest to explore lexical use in other languages since it may reveal more patterns of lexical variation. Future studies may also examine the language variation in written texts. The reason is that writing is a more formal register for expressing one's ideas, and findings may provide more insights into social class differences at the lexical level across more registers. In addition, the study is also limited in that it has only one control variable, i.e., social class. Other variables such as gender and age may also affect lexical use. Thus, it may be of interest to examine the effect of gender and age on the use of lexical items.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aliakbari, M., Samaie, M., Sayehmiri, K., Qaracholloo, M., 2012. The grammatical correlates of social class factors: The case of Iranian fifth-graders. Linguistik Online 56 (6/12), 3–20.

Altenberg, E.P., Vago, R.M., 2010. English Grammar: Understanding the Basics. Cambridge University Press, London.

Baddeley, A., 1983. Working memory. Philosophical Transaction of the Royal Society of London 302, 311–324.

Baker, P., 2006. Using Corpora in Discourse Analysis. Continuum, London.

Bernstein, B., 1971. Class, Codes and Control, Volume I: Theoretical Studies Towards a Sociology of Language. Routledge & Kegan Paul, London.

Bernstein, B., 1973. Class, Codes and Control, Volume 2: Applied Studies Towards a Sociology of Language. Routledge & Kegan Paul, London.

Bernstein, B., 1975. Class, Codes and Control, Volume 3: Towards a Theory of Educational Transmissions. Routledge & Kegan Paul, London.

Block, D., 2012. Class and SLA: Making connections. Lang. Teach. Res. 16 (2), 188–205. https://doi.org/10.1177/1362168811428418.

Block, D., 2014. Social Class in Applied Linguistics. Routledge, New York.

Block, D., 2015. Social class in applied linguistics. Ann. Rev. Appl. Ling. 35, 1–19. https://doi.org/10.1017/S0267190514000221.

Blommaert, J., 2010. The Sociolinguistics of Globalization. Cambridge University Press, Cambridge.

Bourdieu, P., 1977. The economics of linguistic exchanges. Soc. Sci. Inform. 16 (6), 645–668.

Bourdieu, P., 1991. Language and Symbolic Power. Harvard University Press, Harvard.

Bullock, H.E., Lott, B., 2010. Social class and power. In: Guinote, A., Vescio, T.K. (Eds.), The Social Psychology of Power. Guilford Press, New York, pp. 408–427.

Canagarajah, S., 2013. Translingual Practice: Global Englishes and Cosmopolitan Relations: Lingua Franca English and Global Citizenship. Routledge, London.

Cheshire, J., 2005. Syntactic variation and beyond: Gender and social class variation in the use of discourse-new markers. J. Socioling. 9 (4), 479–508.

Collis, D., 2009. Social Grade: A Classification Tool. Retrieved from https://www.ipsos.com/sites/default/files/publication/680003/MediaCT_thoughtpiece_Social_Grade_July09_V3_WEB.pdf (last accessed September 2020).

Cowan, N., 2016. Working Memory Capacity. Psychology Press, New York.

Coxhead, A., 2000. A new academic word list. TESOL Quart. 34 (2), 213–238.

Daller, H., Milton, J., Treffers-Daller, J., 2007. Modelling and Assessing Vocabulary Knowledge. Cambridge University Press, London.

Daller, H., Van Hout, R., Treffers-Daller, J., 2003. Lexical richness in the spontaneous speech of bilinguals. Appl. Ling. 24 (2), 197–222.

Davis, D., Brock, T.C., 1975. Use of first person pronouns as a function of increased objective self-awareness and performance feedback. J. Exp. Soc. Psychol. 11, 381–388.
Duval, S., Wicklund, R.A., 1972. A Theory of Objective Self-awareness. Academic Press, New York.
Degaetano-Ortlieb, S., 2018. Stylistic variation over 200 years of court proceedings according to gender and social class. Paper presented at the Proceedings of the 2nd Workshop on Stylistic Variation.
Dickson, V., Hall-Lew, L., 2017. Class, gender, and rhoticity: The social stratification of non-prevocalic /r/ in Edinburgh speech. J. Engl. Ling. 45 (3), 229–259.
Eckert, P., 2000. Linguistic Variation as Social Practice. Blackwell, Oxford.
Eckert, P., 2008. Variation and the indexical field. J. Socioling. 12 (4), 453–476.
Eckert, P., 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. Ann. Rev. Anthropol. 41 (1), 87–100. https://doi.org/10.1146/annurev-anthro-092611-145828.
Fan, F., 2011. A corpus based quantitative study on the change of TTR, word length and sentence length of the English language. In: Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of His 75th Birthday, pp. 123–130. https://doi.org/10.1515/9783110894219.
Fan, F., 2013. Text length, vocabulary size and text coverage constancy. J. Quantit. Ling. 20 (4), 288–300.
Fox, K., 2004. Watching the English: The Hidden Rules of the English Behaviour. Hodder & Stoughton, London.
Gablasova, D., Brezina, V., McEnery, T., 2017. Exploring learner language through corpora: Comparing and interpreting corpus frequency information. Lang. Learn. 67, 130–154. https://doi.org/10.1111/lang.12226.
Gallo, L.C., Matthews, K.A., 2003. Understanding the association between socioeconomic status and physical health: Do negative emotions play a role?. Psychol. Bull. 129 (1) 10–51.
Gilquin, G., Granger, S., 2014. Learner language. In: Biber, D., Reppen, R. (Eds.), The Cambridge Handbook of English Corpus Linguistics. Cambridge University Press, Cambridge, pp. 418–435. https://doi.org/10.4324/9781315845890.
Gist, A.N., 2017. Class and organizing. In: Scott, C., Lewis, L. (Eds.), The International Encyclopaedia of Organizational Communication. John Wiley & Sons Inc, New Jersey, pp. 1–13.
Grant, J.A., 2001. Class, definition of. In: Jones, R.J.B. (Ed.), Routledge Encyclopedia of International Political Economy. Taylor & Francis, London, pp. 161–163.
Grenfell, M.J., 2014. Pierre Bourdieu: Key Concepts. Routledge, London.
Grzybek, P., 2007. History and methodology of word length studies. In: Grzybek, P. (Ed.), Contributions to the Science of Text and Language: Word Length Studies and Related Issues. Springer, Dordrecht, pp. 15–90.
Guy, G.R., 2011. Language, social class, and status. In: Mesthrie, R. (Ed.), The Cambridge Handbook of Sociolinguistics. Cambridge University Press, New York, pp. 159–185.
Hessner, T., Gawlitzek, I., 2017. Totally or slightly different? A Spoken BNC2014-based investigation of female and male usage of intensifiers. Int. J. Corpus Ling. 3 (1), 403–428.
Hoff, E., 2006. How social contexts support and shape language development. Dev. Rev. 26 (1), 55–88.
Huddleston, R., Pullum, K.G., 2002. The Cambridge Grammar of the English Language. Cambridge University Press, London.
Hudson, R.A., 1996. Sociolinguistics. Cambridge University Press, Cambridge.
Hundt, M., Sand, A., Siemund, R., 1998. Manual of information to accompany the Freiburg-LOB Corpus of British English (`FLOB'). Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau.
Jaspal, R., 2009. Language and social identity: A psychosocial approach. Psych-Talk 64, 17–20.
Kacewicz, E., Pennebaker, J.W., Davis, M., Jeon, M., Graesser, A.C., 2014. Pronoun use reflects standings in social hierarchies. J. Lang. Soc. Psychol. 33 (2), 125–143.
Kestemont, M., 2014. Function words in authorship attribution from black magic to theory?. In: Paper presented at Proceedings of the 3rd Workshop on Computational Linguistics for Literature pp. 59–66.
Koppen, K., Ernestus, M., Van Mulken, M., 2019. The influence of social distance on speech behavior: Formality variation in casual speech. Corpus Ling. Ling. Theory 15 (1), 139–165.
Kilgarriff, A., 2005. Language is never, ever, ever, random. Corpus Ling. Ling. Theory 1 (2), 263–276. https://doi.org/10.1515/cllt.2005.1.2.263.
Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., Mannila, H., 2016. Significance testing of word frequencies in corpora. Digit. Schol. Hum. 31 (2), 374–397. https://doi.org/10.1093/llc/fqu064.
Labov, W., 1966. The Social Stratification of English in New York City. Centre for Applied Linguistics, Washington D.C.
Labov, W., 1966b. The effect of social mobility on linguistic behaviour. Sociol. Inq. 36 (2), 186–203.
Labov, W., 1972. The social stratification of /r/ in New York City. In: Labov, W. (Ed.), Sociolinguistic Patterns. University of Pennsylvania Press, Philadelphia, pp. 168–178.
Labov, W., 1984. Intensity. In: Schiffrin, D. (Ed.), Meaning, Form, and Use in Context: Linguistic Applications. Georgetown University Press, Washinton D.C., pp. 43–70.
Lei, L., Liu, Z., 2014. A word type-based quantitative study on the lexical change of American and British English. Journal of Quantitative Linguistics 21 (1), 36–49. https://doi.org/10.1080/09296174.2013.856131.
Lei, L., Liu, D., 2016. A new medical academic word list: A corpus-based study with enhanced methodology. Journal of English for Academic Purposes 22, 42–53. https://doi.org/10.1016/j.jeap.2016.01.008, In press.
Lei, L., Wen, J., 2020. Is dependency distance experiencing a process of minimization? A diachronic study based on the State of the Union addresses. Lingua 239, 102762. https://doi.org/10.1016/j.lingua.2019.102762.

Lian, F., Li, Y., 2019. Word length distribution in German texts during the 17–19 Century. J. Quantit. Ling. 1–21. https://doi.org/10.1080/09296174.2019.1662536 (accessed 17 March 2020).
Liang, J., Liu, H., 2013. Noun distribution in 5 natural languages. Poznań Stud. Contemp. Ling. 49 (4), 487–507.
Love, R., Hawtin, A., Hardie, A., 2017. The British National Corpus 2014: User Manual and Reference Guide (version 1.0). ESRC Centre for Corpus Approaches to Social Science, Lancaster.
Macaulay, R., 1995. The adverbs of authority. Engl. Word-Wide 16, 37–60.
Macaulay, R., 2002. Extremely interesting, very interesting, or only quite interesting? Adverbs and social class. J. Socioling. 6 (3), 398–417.
Malvern, D., Richards, B., 2013. Measures of lexical richness. In: Chapelle, C. (Ed.), The Encyclopedia of Applied Linguistics. Blackwell Publishing Ltd, New Jersey, pp. 3968–3972.
Martin, S.R., Innis, B.D., Ward, R.G., 2017. Social class, leaders and leadership: A critical review and suggestions for development. Curr. Opin. Psychol. 18, 49–54.
Mather, P., 2012. The social stratification of /r/ in New York city: Labov's department store study revisited. J. Engl. Ling. 40 (4), 338–356.
McCarthy, P.M., Jarvis, S., 2007. vocd: A theoretical and empirical evaluation. Lang. Test. 24 (4), 459–488.
Meyerhoff, M., 2006. Introducing Sociolinguistics. Taylor & Francis e-Library, New York.
Mikros, G.K., 2012. Authorship attribution and gender identification in Greek blogs. In: Obradovic, I., Emmerich, K., Reinhard, K. (Eds.), Methods and Applications of Quantitative Linguistics. Academic Mind, Belgrade, pp. 21–32.
Mitford, N., 1956. Noblesse Oblige: An Enquiry into the Identifiable Characteristics of the English Aristocracy. Hamish Hamilton, London.
New, B., Ferrand, L., Pallier, C., Brysbaert, M., 2006. Re-examining the word length effect in visual word recognition: New evidence from the English Lexicon Project. Psychon. Bull. Rev. 13 (1), 45–52. https://doi.org/10.3758/BF03193811.
Núñez Pertejo, P., Palacios Martínez, I.M., 2014. That's absolutely crap, totally rubbish: the use of the intensifiers absolutely and totally in the spoken language of British adults and teenagers. Funct. Lang. 21 (2), 210–237.
Nweke, C.O., 2013. A Review of English syllable structure. Afrrev Laligens Int. J. Lang. Literat. Gender Stud. 2 (1), 142–157.
Paquot, M., Bestgen, Y., 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In: Jucker, A.H., Schreier, D., Hundt, M. (Eds.), Corpora: Pragmatics and Discourse. Rodopi, Amsterdam, pp. 247–269. https://doi.org/10.1163/9789042029101_014.
Pennebaker, J.W., 2011. The secret life of pronouns. New Sci. 211 (2828), 42–45.
Pennycook, A., 2012. Language and Mobility: Unexpected Places. Multilingual Matters, Clevedon.
Pérez-García, E., Sánchez, M.J., 2020. Emotions as a linguistic category: Perception and expression of emotions by Spanish EFL students. Lang. Cult. Curr. 33 (3), 274–289.
Rampton, B., 2010. Social class and sociolinguistics. Appl. Ling. Rev. 1 (1), 1–22.
Rayson, P., Berridge, D., Francis, B., 2004. Extending the Cochran rule for the comparison of word frequencies between corpora. JADT 2004: 7es Journées Internationales d'Analyse Statistique Des Données Textuelles, pp. 1–12. Retrieved from http://eprints.lancs.ac.uk/12424.
Richardson, A., Allen, J.A., Xiao, H., Vallone, D., 2012. Effects of race/ethnicity and socioeconomic status on health information-seeking, confidence, and trust. J. Health Care Poor Underserved 23 (4), 1477–1493.
Roach, P., 2009. English Phonetics and Phonology: A Practical Course. Cambridge University Press, London.
Ross, A.S.C., 1954. Linguistic class-Indicators in present-day English. Neuphilologische Mitteilungen 55 (1), 20–56.
Savage, M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J., Le Roux, B., Friedman, S., Miles, A., 2013. A new model of social class? Findings from the BBC's Great British Class Survey Experiment. Sociology 47 (2), 219–250. https://doi.org/10.1177/0038038513481128.
Schieffelin, B.B., Ochi, E., 1986. Language socialization. Ann. Rev. Anthropol. 15 (1), 163–191.
Shi, Y., Lei, L., 2020. Lexical richness and text length: An entropy-based perspective. Journal of Quantitative Linguistics, 1–18. https://doi.org/10.1080/09296174.2020.1766346.
Simpson, R., Hughes, J., Slutskaya, S., 2016. Gender, Class and Occupation: Working Class Men Doing Dirty Work. Palgrave Macmillan, London.
Singh, S., 2001. A pilot study on gender differences in conversational speech on lexical richness measures. Lit. Ling. Comput. 16 (3), 251–264.
Snell, J., 2014. Social class and language. Available at http://www.snell.me.uk/wp-content/uploads/HoP-Snell_-Social-class-and-language_updated_changes-accepted.pdf (accessed 30 December 2019).
Snell, J., 2017. Enregisterment, indexicality and the social meaning of `howay': Dialect and identity in north-east England. In: Moore, E., Montgomery, C. (Eds.), Language and a Sense of Place. Cambridge University Press, Cambridge, pp. 301–324.
Snell, J., 2018. Solidarity, stance, and class identities. Language in Society 47 (5), 665–691. https://doi.org/10.1017/S0047404518000970.
Snow, C.E., 1999. Social perspectives on the emergence of language. In: MacWhinney, B. (Ed.), The Emergence of Language. Taylor & Francis, London, pp. 257–276.
Stamatatos, E., 2009. A survey of modern authorship attribution methods. J. Am. Soc. Inform. Sci. Technol. 60 (3), 538–556.
Stellar, J.E., Manzo, V.M., Kraus, M.W., Keltner, D., 2012. Class and compassion: Socioeconomic factors predict responses to suffering. Emotion 12 (3), 449–459.

Stephens, N.M., Markus, H.R., Phillips, L.T., 2014. Social class culture cycles: How three gateway contexts shape selves and fuel inequality. Annu. Rev. Psychol. 65 (1), 611–634.
Strömqvist, S., Johansson, V., Kriz, S., Ragnarsdóttir, H., Aisenman, R., Ravid, D., 2002. Toward a cross-linguistic comparison of lexical quanta in speech and writing. Written Lang. Lit. 5 (1), 45–67.
Tabata, T., 2002. Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution. In: Saito, T., Nakamura, J., Yamazaki, S. (Eds.), English corpus linguistics in Japan. Brill Rodopi, Amsterdam, pp. 165–182.
Thompson, W.E., Hickey, J.V., 2010. Society in Focus: An Introduction to Sociology. Pearson Education Inc, Boston.
Trudgill, P., 1974. The Social Differentiation of English in Norwich. Cambridge University Press, Cambridge.
Upton, C., Parry, D., Widdowson, J.D.A., 1994. Survey of English dialects: The dictionary and grammar. Routledge, London & New York.
Yusuf, Y.Q., Nasir, C., Andib, N., 2019. Power and solidarity: The pronoun of address *ke* [ke] used in Indonesian by acehnese speakers. International Journal of Language Studies 13 (1), 77–98.
Zhang, Y., 2014. A corpus-based analysis of lexical richness of Beijing Mandarin speakers: Variable identification and model construction. Lang. Sci. 44, 60–69.
Zimmermann, J., Wolf, M., Bock, A., Peham, D., Benecke, C., 2013. The way we refer to ourselves reflects how we relate to others: Associations between first-person pronoun use and interpersonal problems. J. Res. Pers. 47 (3), 218–225. https://doi.org/10.1016/j.jrp.2013.01.008.

Yaqian Shi is a Ph.D. candidate of Applied Linguistics at Huazhong University of Science and Technology. Her research interests include corpus linguistics, quantitative linguistics, and Academic English. She has published articles in *English Today* and *Journal of Quantitative Linguistics*.

Lei Lei is Professor of Applied Linguistics at Shanghai Jiao Tong University. His research interests include corpus linguistics, quantitative linguistics, and Academic English. He has published extensively in journals such as *Applied Linguistics*, *International Journal of Corpus Linguistics*, *Journal of English for Academic Purposes, Lingua, System, Journal of Quantitative Linguistics*, and *English Today*.