Language Testing in Asia

RESEARCH Open Access

How effective are lexical richness measures for differentiations of vocabulary proficiency? A comprehensive examination with clustering analysis

Yanhui Zhang[1*] and Weiping Wu[2]

* Correspondence: Yanhui.Zhang@nottingham.edu.cn
[1]University of Nottingham Ningbo China, Ningbo, China
Full list of author information is available at the end of the article

Abstract

This study proposed an innovative automated approach to differentiation of the vocabulary proficiency of Chinese speakers. A robust K-means algorithm was designed to compare the oral proficiency between L1 and L2 Chinese speakers regarding lexical richness and how relatively effective the various lexical measures were in performing the differentiation task. Eighteen lexical richness measures were surveyed and compared using the clustering analysis. The effectiveness of each selected measure as well as an overall evaluation of all the measures for the concerned differentiation tasks were comprehensively calibrated. The results demonstrate that, while the L1 versus L2 group difference in lexical richness was observed with statistical significance for each of the chosen measures, the clustering and membership prediction accuracy of individual speakers varied greatly from one measure to another. The implication is that a more fully defined metric of lexical richness is still a worthwhile endeavor for language proficiency assessment, with optimal directions for such endeavors discussed in the concluding remarks.

Keywords: Language proficiency, Lexical richness measures, Chinese as a foreign language, Language testing, L1 and L2 comparison, Clustering analysis

Introduction

Lexical proficiency is one of the most critical components of linguistic competence to second language (L2) learning. Past studies have documented abundant and consistent evidence that lexical proficiency is the predominant element directly affecting the learners' performance at all major fronts such as L2 reading and writing literacy, oral fluency, and academic achievements (Anderson & Freebody, 1981; Daller et al., 2003; Huckin & Coady, 1999; In'nami et al., 2016; Koda, 1988; Li, 2018). At a practical level, effective communication in L2 is improbable without a sound mastery of the vocabulary of the target language (Akiyama & Saito, 2016; Alqahtani, 2015; Ellis, 1995; Gu, 2019; Newman et al., 2016; Wright & Cervetti, 2017). As such, appropriate measurement of lexical proficiency has been a meaningful quest to enhance the effectiveness and efficiency of L2 education in general.

Traditionally, researchers rely more on experimental approaches for discerning the nativeness or non-nativeness of lexical productions, whereas corpus-based numerical analysis is rarely seen on a large scale. Apart from the notion that human judgment is, by default, most accurate, an apparent reason causing the rare use of numerical analysis is the lack of a sufficiently calibrated corpus of L2 speakers (for the case of spoken language). A well-contrasted spoken corpus of Mandarin Chinese is exceptionally scarce, due to—among other reasons—the smaller number of Chinese as foreign language (CFL) learners relative to that of English as the second language (ESL) learners (despite growing popularity in Chinese in recent years) as well as the high technical (e.g., segmentation for non-alphabetic languages) and financial cost for constructing such a corpus. This is in spite of the existence of several influential online resources such as Beijing Mandarin Spoken Corpora (BJKY, developed by Beijing Language and Culture University), for instance, because to be applicable to specific aspects of second language acquisition (SLA), such as the lexical knowledge differentiation concerned in the current study, it is best for the corpus to have related latent variables embedded in the design in the first place. Further discussions on challenges surrounding the construction of a spoken corpus of Mandarin Chinese can be found in Xiao et al. (2004), for instance.

One of the more theoretical reasons leading to a heavy reliance on experimental approaches and human judgment instead of systematic and quantitative computations in lexical-related research is the ambiguity surrounding the measurement of lexical proficiency. A clear-cut definition is still missing: the term ‘lexical proficiency’ refers to vocabulary size, the depth of word knowledge, and the degree of sophistication of word use (Crossley et al., 2011). A survey of existing studies, where lexical proficiency is measured quantitatively, shows that lexical richness (LR) is most widely used, both conceptually and practically, for research along the lines of the current study, although the measurement of LR has proven quite an open problem in its own right (Jarvis, 2013; Malvern et al., 2004; Tweedie & Baayen, 1998). Although some earlier studies (Connor, 1984; Reynolds, 1995) suggest that quantitative LR parameters are not effective enough to predict L2 proficiency in general, many recent studies have proved otherwise. For instance, Ellis (2009) demonstrated that LR is a valuable tool to help SLA researchers to analyze the effectiveness of different task designs and learning strategies. Déogratias’ (2011) conclusion that lexical competence is embodied mainly in LR and hence constitutes a reliable predictor of L2 proficiency implies that L2 instruction and proficiency tests should place great emphasis on lexical knowledge. Using the speech data of ESL learners with diversified L1 backgrounds, Crossley et al. (2011) has shown that LR measures (particularly D) are highly predictive of the speaker’s proficiencies, rated by native human raters; altogether, LR accounts for about 45% of the total variance in human ratings. As a further example, similar LR measures were used by Bosker et al. (2014) to differentiate the native and non-native German speakers. In addition to serving as an immediate indicator of various dimensions of lexical proficiency, LR often provides information upon which other advanced discourse constructs, such as word sense, word networks, and cohesion, are founded.

The main purpose of the current study is to quantify the LR differences between CFL learners and native Chinese speakers based on a full range of LR measures that have been proven relevant in one way or another. Analytically, the research addresses the question of whether or not any of the quantitative measures such as RootTTR (Guiraud, 1960) or D (Malvern & Richards, 1997), on its own or collectively, can be reliably used to predict the nativeness versus non-nativeness of a given individual based on their transcribed spoken texts. Lastly, the relative effectiveness of the selected LR indices is evaluated in terms of the goodness of clustering and classification accuracy for the L1 versus L2 differentiation task. The data for the study is a digitized spoken corpus of CFL learners and native Chinese speakers. A detailed introduction of the corpus will be provided as part of this paper's "Research method" section. The Computerized Language Analysis Program (CLAN) generates the raw frequency information of each text based on which the subsequent LR analysis can be carried out. CLAN is an analytical tool, freely downloadable via the link at the Child Language Data Exchange System (CHILDES), which is increasingly recognized and used by researchers in many domains, including SLA and corpus linguistics (MacWhinney, 2007). The classification technique applied in the current study is K-means clustering, whose application in applied linguistics can be found in Kaufman and Rousseeuw (1990), Sultana (2019), and Golshaie (2016), for instance.

Although research using LR measures to fathom and foster lexical proficiency has been increasing, most of them involve EFL. To the best of our knowledge, this is the first systematic study of LR-based lexical proficiency differentiation concerning L1 and L2 Chinese speakers at a large corpus level (average tokens of 2000+ characters per text). Given the vast difference between English and Chinese (e.g., morphologically, syntactically, word networks), it is a reasonable and effort-worthy quest to test against a Chinese corpus the empirical findings involving the LR measuring of lexical proficiencies. The significance of the current study is not only that it validates various LR measurements using a corpus-based approach, but also that—by pinpointing the lexical weakness of CFL learners with a statistically-tested global picture—it could lead to the potential improvement of the overall CFL learning effectiveness with more focused instructional strategies.

The rest of the paper unfolds as follows: the "Measures for quantifying lexical richness" section overviews the range of LR measures to be investigated in the current study grouped according to how the frequency information is used to define such measures. The "Research method" section outlines the methodological steps of the current study, where the description of the corpus is focused on highlighting the distinctive features of Chinese text processing. The "Results" section provides the clustering analysis results together with the goodness of fit comparisons for each group of LR measures; subsequently, the conclusion of the overall efficacy of all the measures is presented. The paper concludes with a further comment on possible ways to improve the differentiation task performance and potential future directions.

Measures for quantifying lexical richness

Lexical richness is a multidimensional concept focusing on the quality of vocabulary in a language sample (Jarvis, 2013; Malvern & Richards, 2012; Siskova, 2012; Zhang, 2014). The notion of lexical richness, as currently perceived by many researchers, has undergone a gradual expansion of meaning over time from its original specific reference of the number of words in a person's mental lexicon (Yule, 1944) to a superordinate category encompassing various lexical aspects of a spoken or written discourse (Jarvis, 2013). This contemporary notion of lexical richness includes lexical density, which basically refers to the percentage of the words with specified lexical properties (e.g., adjectives as a set of content words) out of the total running words of a text; lexical diversity, which involves how many different words are deployed in a text of a given length (Hoover, 2003; Malvern et al., 2004; Shin, 2019); and lexical sophistication, which mainly looks at to what extent rare and advanced words (or "difficult" words according to Vermeer (2004) are being used (Tweedie & Baayen, 1998).

Intuitively, the higher the lexical richness of a discourse, the more verbally diversified and complex it is perceived to be (although an optimal degree of sophistication should not go beyond the reader or listener's comprehension for the purpose of effective perception). Because of this indicative function, LR is routinely applied in quite a number of areas of language studies, such as lexical complexity, as well as in vocabulary knowledge, language fluency, and lexical proficiency measurement, which is close to the aim of the current study (Crossley et al., 2011; Farahani et al., 2019; Johansson, 2008; Laufer & Nation, 1995; Skehan, 2009). The applications of LR are not restricted to the mainstream SLA topics. Still, they have spread to neighboring disciplines such as authorship detection (Smith & Kelly, 2002), language attrition (Schmid, 2010), language disorder and therapy (Silverman & Ratner, 2002), or even more distant fields such as social-economic appraisal and healthcare assessment (McCarthy & Jarvis, 2007).

Given the multifaceted nature of quantitative measurement, it is probably not surprising to take many different forms in literature. On the other hand, no matter what aspect of LR one chooses to measure, the fundamental leverage is always—from a computing and formulation perspective—the frequency distribution of individual words among the whole text. Thus, the likelihood of each word appearing in the text, which computationally corresponds to the ratio of the number of the occurrence of a particular word divided by the number of tokens, constitutes the building blocks of all the LR measures seen in the literature. Based on this criterion, the LR measures used for the current study were divided into the following three groups.

The first group of LR measures (group I) is those involving only the type and/or token information of a text, namely, Types, Tokens, TTR, RootTTR (Guiraud, 1960), LogTTR (Herdan, 1960), Uber (Dugast, 1979), and D. Here, Types is defined as the total number of different words or characters of a text. Tokens refer to the total number of running words of a text, defining the length of the text. TTR is simply the ratio between Types and Tokens. RootTTR, LongTTR, Uber, and D are all variations or transformations of TTR (termed by Tweedie & Baayen, 1998), all aiming to overcome or at least mitigate the length dependence of TTR.

The group II measures rely on a partial spectrum of frequency types. They are called partial because not all types of frequency spectrum information are included in the formula of these indices. They are the number of *hapax legomena*, denoted as V_1, which is the number of words appearing only once; V_2, or the number of *dis legomena*, which is the number of words appearing exactly twice; and so on. Certain measures based on algebraic manipulations of partial frequency spectrum information can also be found in the literature on the subject. For instance, Honored (1979) proposed a new measure (referred to as 'Honore' in this paper) by claiming that the ratio between V_1 and Types is linearly dependent on the logarithm of Tokens. Sichel (1975), on the other hand, noticed that V_2 divided by Tokens is roughly invariant to the text size, and therefore suggested this ratio to measure the LR of different texts or authors.

The group III measures include those incorporating full-frequency spectrum information of all types, such as entropy and relative entropy, where the latter is simply the entropy divided by the maximum possible entropy of a given text, achieved when each type occurs one but only one time (Shannon, 1951). Measures falling into this category also include Yule K (Yule, 1944), defined as a function of all relative frequencies of all types, where the relative frequency of a type is simply the ratio between the number of occurrences of this type and Tokens; Yule I (Yule, 1944), which is essentially the reciprocal of Yule K, and V_m (Herdan, 1960), which is again a transformation of Yule K.

Table 1 lists all the LR measures studied in the current paper, each accompanied by its notation, specification, calculation formula, notable references, and commonly seen variations in names or notations.

Table 1 LR measures studied in the current paper

Group	Label (notation)	Notes and explanation	Specification or formula	Variant labels in literature
	Types (T)	Number of different words	T	NDW
	Tokens (N)	Total running words	N	
I	TTR	Types divided by Tokens	T/N	
	RootTTR	square root adjustment of TTR	$T/\sqrt{N}$	Guiraud Index
	LogTTR	Logarithm adjustment of TTR	$\log(T)/\log(N)$	Herdan C
	D	A lexical diversity measure based on iterated calculation of TTR (Malvern & Richards, 1997)	$TTR=\frac{D}{N}(\sqrt{1+2\frac{N}{D}}\ -1)$	
	Uber	A quantity based on Types and Tokens (Dugast, 1979)	$\frac{\log^2 T}{\log(N)-\log(T)}$	Maas
	V_1	Number of *hapax legomena*	V_1	V(1,N)
	V_2	Number of *dis legomena*	V_2	V(2,N)
II	V_1TR	V_1 token ratio	V_1/N	V_1 Ratio
	V_2TR	V_2 token ratio	V_2/N	V_2 Ratio
	Honore	A quantity involving V_1, Types, and Tokens (Honored, 1979)	$100\log(N)/(1-V_1/N)$	R
	Sichel	a characteristic constant proposed by Sichel (1975)	V_2/T	S
	Entropy(E)	A quantity measuring the complexity of a text (Shannon, 1951)	$\mathrm{E}=-\sum_{i=1}^{T}p_i\ \log p_i$	
III	Relative entropy (RE)	Entropy scaled by the maximum entropy of a text	$\mathrm{RE}=\mathrm{E}/\log(N)$	
	Yule K	A characteristic constant proposed by Yule (1944)	$10^4(-\frac{1}{N}+\sum_i V_i(\frac{i}{N})^2)$	
	Yulk I	An algebraic transformation of Yule K	$\frac{10^4}{\text{Yule K}}$	
	V_m	A modification of Yule K proposed by Herdan (1960)	${V_m}^2=\text{Yule K}+(1/N-1/T)$	

Research method

Context and corpus description

The data used for the current study is the speech corpus of both CFL learners, and native speakers of Mandarin Chinese generated from the oral assessment of Putonghua (Mandarin Chinese), administered at The Chinese University of Hong Kong. The purpose of the test is to provide CFL learners a way to assess their Chinese oral proficiency. Insofar as it focuses exclusively on oral Chinese, the assessment contrasts to other language tests, namely, the HSK (abbr. of *Hanyu Shuiping Kaoshi* in Pinyin, or Chinese Proficiency Test), an official Chinese proficiency test that approximates TOEFL for English, which does not have oral tasks for level 1 and 2 tests, or BCT (Business Chinese Test), a test developed by Peking University for assessing Chinese skills in business occasions, the emphasis of which is formal and professional communication.

Participants and test design

The test asks the CFL participant to identify their mother tongue before formal test questions are prompted. If the mother tongue is English, Japanese, Korea, or Cantonese, the participant will hear the test rules and directions in their mother tongue. Otherwise, the directions are given in English. All the 28 CFL learners who participated in the current study are native English speakers. In terms of the education level, all participants were at least holding, or studying towards, a college degree in various fields. The ages of the participants range from 20 to 43 (mean 27, standard deviation 4.5); none of them reported having a hearing problem. The self-reported Chinese learning in a formal class setting ranges from 0 to 4 years (mean 2.4 years, standard deviation 0.9 years), which does not include their prior experience of Chinese learning in various informal environments such as self-teaching and personal tutoring. Table 2 presents the summary of the background information of the CFL learners for the current study. Participants in the control group who were recruited to take the same test are 25 native Chinese speakers (average age 24 with standard deviation 1.4). All of them are attending or have completed a university-level education in China. All the L1 participants in the current corpus speak standard Mandarin Chinese at home, after the deletion of two who reported speaking non-Mandarin Chinese dialects.

Lasting for about 30 min, the test consists of four tasks covering a variety of topics with sociocultural themes commonly encountered in daily, professional, or social settings. The first task comprises five short-answer questions. Each asks the participant to provide a brief response in 10 s to a conversational inquiry in daily life settings, such as

Table 2 Background information of the CFL learners

Number of participants		**28**
Education		College and above
	Range	20–43
Age (years)	Mean	27
	Standard deviation	4.5
	Range	0–4
Chinese learning experience (years)	Mean	2.4
	Standard deviation	0.9

"When does your friend get up every day?". The second task asks the participant to propose one or more questions in 15 s to each of the four pictures, where the scene of the picture can be an apartment building, for instance, where a sensible response in Chinese can be "这间公寓有多大面积?每个月的租金是多少?从这里乘公交车去学校需要多长时间?" (How large is this apartment? How much is the monthly rent of it? How long does it take to go from here to school by bus?). The third task asks the participant to narrate a coherent story based on pictures in 90 s. The fourth task includes 12 free-response questions in the simulation of linguistic functions needed to describe or explain something or some event, command or instruction, debate, assertion, argument, persuasion, defending, or other daily and social interactions. Nine of these questions ask the participant to respond in 60 s and three in 90 s.

Data analysis

The Computerized Language Analysis (CLAN) is the tool deployed to segment the transcribed texts and generate the spectrum of raw Types and Tokens frequency statistics. For a complete introduction of CLAN, including the installation, tutorial, examples of LR analysis, etc., one can refer to MacWhinney (2007). K-means clustering is the principal tool for the L1 versus L2 classification task concerned in the current paper. Compared with traditional statistical methods such as linear or nonlinear regressions and other structural equation models, clustering analysis is more robust since it is essentially non-parametric from a machine learning perspective. The unsupervised nature of the algorithmic design makes it particularly applicable to the classification problem faced with the current study since the corpus did not foretell how many native or non-native speakers were included. The clustering procedure and the quality assessment of it can be found in Kaufman and Rousseeuw (1990) and Hennig et al. (2015). Once the optimized classification of the L1/L2 mixed dataset is done, one can compare the programmed classification results with the true classification, i.e., classification of L1 versus L2 based on the mother tongue of the speakers. From this, mis-clustered points can be detected accordingly and the overall classification accuracy for each LR measure can be calculated as one minus the percentage of the mis-clustered.

Results

Summary of ANOVA

Tables 3 and 4 provide the key statistics of the ANOVA procedure for all the 18 LR measures concerned, using the raw frequency distribution of each text contained in the corpus of characters and words. Table 3 is based on the raw frequency distribution of characters contained in the texts of the corpus, while Table 4 is based on the frequency distribution of words. Seven columns of statistics are tabulated in both tables. Columns 1–3 tabulate the mean, standard deviation (SD), and normalized standard deviation (NSD) of each measure generated from the L1 data. Columns 4–6 repeat the statistics of columns 1–3 with L2 data. Columns 7–9 are the results of the two-sample comparison of means, tested with the assumption that the variances of the LR measure for L1 and L2 data are not necessarily equal. More specifically, column 7 records the *p* value of the comparison of means. Columns 8–9 are the 95% confidence interval (CI) of the difference of an LR measure. All tests are carried out at a default level of significance of

Table 3 ANOVA statistics for comparison of LR difference by characters in L1 and L2

LR measure	L1			L2			t test		
	Mean	**SD**	**NSD**	**Mean**	**SD**	**NSD**	***p***	**CI**	
Type	517.53	48.24	0.09	314.27	85.36	0.28	0.00	(165.15,	240.88)
Token	2483.39	378.38	0.14	1772.87	675.14	0.39	0.00	(411.04,	1008.66)
TTR	0.21	0.03	0.12	0.18	0.04	0.24	0.02	(0.00,	0.04)
RootTTR	10.52	0.83	0.08	7.49	1.07	0.15	0.00	(2.37,	3.42)
LogTTR	0.84	0.01	0.02	0.76	0.02	0.03	0.00	(0.02,	0.04)
D	69.70	13.16	0.19	35.80	10.49	0.29	0.00	(27.27,	40.54)
Uber	25.12	2.16	0.09	19.51	2.43	0.12	0.00	(4.34,	6.87)
V_1	194.60	28.24	0.15	117.29	33.66	0.29	0.00	(60.23,	94.39)
V_2	95.84	14.07	0.15	56.29	18.56	0.33	0.00	(30.52,	48.59)
V_1TR	0.08	0.02	0.20	0.07	0.02	0.31	0.11	(0.00,	0.02)
V_2TR	0.04	0.01	0.18	0.03	0.01	0.35	0.06	(0.00,	0.01)
Honore	1254.49	69.96	0.06	1186.49	90.18	0.08	0.00	(23.71,	112.28)
Sichel	0.19	0.02	0.13	0.18	0.03	0.17	0.35	(− 0.01,	0.02)
E	5.48	0.14	0.02	4.87	0.27	0.06	0.00	(0.49,	0.72)
RE	0.70	0.02	0.03	0.66	0.03	0.05	0.00	(0.03,	0.06)
YuleK	92.34	17.77	0.19	173.88	72.09	0.41	0.00	(− 110.27,	− 52.81)
YuleI	111.92	20.10	0.18	66.41	23.87	0.36	0.00	(33.38,	57.65)
V_m	0.01	0.00	0.22	0.01	0.01	0.47	0.00	(− 0.01,	0.00)

Table 4 ANOVA statistics for comparison of LR difference by words in L1 and L2

LR measure	L1			L2			t test		
	Mean	**SD**	**NSD**	**Mean**	**SD**	**NSD**	***p***	**CI**	
Type	499.39	53.53	0.10	291.37	88.63	0.31	0.00	(167.31,	247.06)
Token	1639.21	232.66	0.13	1284.20	494.78	0.39	0.00	(142.78,	564.88)
TTR	0.32	0.03	0.11	0.24	0.05	0.21	0.00	(0.05,	0.09)
RootTTR	12.78	1.08	0.09	8.27	1.40	0.17	0.00	(3.52,	4.88)
LogTTR	0.85	0.01	0.02	0.80	0.02	0.03	0.00	(0.04,	0.06)
D	114.21	21.27	0.19	45.27	16.53	0.37	0.00	(58.31,	79.57)
Uber	32.80	3.55	0.11	22.10	3.56	0.16	0.00	(8.74,	12.67)
V_1	278.84	36.78	0.13	140.14	43.59	0.31	0.00	(116.52,	160.87)
V_2	89.28	12.12	0.14	52.32	17.46	0.33	0.00	(28.73,	45.19)
V_1TR	0.17	0.02	0.14	0.12	0.06	0.53	0.00	(0.02,	0.08)
V_2TR	0.06	0.01	0.15	0.05	0.02	0.53	0.05	(0.00,	0.02)
Honore	1677.78	99.82	0.06	1368.76	137.18	0.10	0.00	(243.27,	374.77)
Sichel	0.18	0.02	0.11	0.18	0.02	0.14	0.87	(− 0.01,	0.01)
E	5.00	0.58	0.12	4.71	0.33	0.07	0.03	(0.03,	0.56)
RE	0.68	0.07	0.10	0.67	0.04	0.06	0.32	(− 0.02,	0.05)
YuleK	115.53	46.66	0.40	243.69	123.98	0.51	0.00	(− 179.34,	− 76.97)
YuleI	101.51	42.17	0.42	51.13	22.55	0.44	0.00	(31.21,	69.55)
V_m	0.01	0.00	0.46	0.02	0.01	0.56	0.00	(− 0.02,	− 0.01)

5%. A couple of overall patterns can be drawn from these statistics. First, and not surprisingly, it is evident that L1 speakers outperformed L2 speakers on average by all the measures. The conclusion is, in general, consistent with those found in previous studies based on alphabetical languages, especially English (Siskova, 2012; Crossley & McNamara, 2009; Skehan, 2009; Daller & Xue, 2007). However, a closer look at the p values and the mean difference CIs finds that not all the LR measures differentiate L1 and L2 speakers with enough statistical significance. For instance, the *p* value for V_1TR is 0.1138, which is substantially higher than the default level of significance of 5%. Similar results have been observed for relative entropy and Sichel. In particular, the *p* value for Sichel is so large (0.8690) that the alternative hypothesis has to be rejected. Another observation is that SDs and within-group diversity are much more apparent for L2 speakers than L1. For instance, the SDs for Type, TTR, V_1, and YuleK are 48.2368, 0.0250, 28,2415, and 17.7680, respectively, for L1 data, but are 85.3587, 0.0413, 33.6616, and 72. 0909 respectively for L2 data. Also, the average NSD for all the 18 measures for L1 speakers is only 0.1243, but 0.2433 for L2.

On the other hand, a few exceptions exist, such as NSD = 0.1164 versus 0.0707 for entropy calculated for L1 and L2. Lastly, considerable differences can be seen for LR measures computed based on characters versus by words. For instance, D values calculated by characters and words differ substantially in L1 and L2 speakers (69.7009 for L2 speakers versus 114.2130 for L1 speakers). Other measures profiling notable differences between character-based and word-based computation include TTR, RootTTR, V1, Honore, as in mean, V1TR, V2TR, entropy, relative entropy, and YuleK, YuleI, and V_m in deviation. These observed patterns underscore the importance of a systematic comparison of all the LR measures in differentiation tasks undertaken by the current research and the clustering analysis using both dimensions in character-based and word-based computations.

Group-wise comparison

This subsection presents the group-wise comparative analysis regarding those measures categorized in group I, II, and III in "Measures for quantifying lexical richness" section of the current paper. For each group of measures, the ANOVA results aimed to compare the lexical proficiencies of L1and L2 speakers at overall scale are firstly presented, followed by the clustering results attempting to classify the given individual speakers into distinct groups based on the chosen LR measures, the benchmark for which is based upon whether they are native Mandarin Chinese speakers or not. Meanwhile, an assessment of clustering quality is provided to demonstrate and compare how effective these LR measures are for the purpose of the intended classification task.

For group I, it is evident that L1 speakers outperformed L2 speakers, on average, in all seven measures. For instance, the average values of the measure D, calculated by characters, are 69.7003 for L1 versus 35.7961 for L2, and a two-sample comparison of means gives t = 10.2869 and p < 10E−4. In addition, it is noted that the standard deviations of the LR values for L1 speakers are all smaller than those for L2 speakers in terms of lexical usage in the oral discourse of the common sociocultural topics covered by the test. Here, standard deviations are normalized in order to make the figures more comparable. The following Fig. 1 plots the results of the iterated K-means clustering using the measure D in this group, with the maximum number of replicates set as 10. The replication of clustering is the simulation process to reduce the random error induced from the initiation of the clustering program. Plotted along with Fig. 1 are the 90% confidence ellipses for the original L1 and L2 data. Red triangular markers mark the mis-clustered points (where clustering error occurred), while the centroids of clusters are marked with magenta squares.

The results show that the clustering based on this group of LR measures is statistically significant, with an average Silhouette of 0.6151 (specifically, in decreasing order, 0.6611 for D, 0.6469 for RootTTR, 0.6431 for Uber, 0.6318 for LogTTR, 0.6289 for Types, 0.5532 for TTR, and 0.5408 for Tokens). And the average standard deviation for the recorded Silhouette values is 0.1744 (specifically, in increasing order, 0.1593 for RootTTR, 0.1577 for Uber, 0.1664 for LogTTR, 0.1703 for TTR, 0.1732 for D, 0.1804 for Tokens, and 0.2187 for Types). In addition, the average accuracy of classification (calculated as the percentage of the correct classification points in relation to the total number points to be clustered) is 89% (specifically, in decreasing order, 96% for D, 94% for RootTTR and Types, 91% for Uber, 89% for LogTTR, 81% for TTR, and 79% for Tokens). D is the best among this group of LR measures for the L1–L2 differentiation task at hand. It beats all the rest measures in various aspects of the quality of classification, especially the accuracy rate. In addition to the quantitative comparisons, one can also graphically infer the effectiveness of D from Fig. 1, where the 90% confidence ellipses based on D for the data plotted are located farthest apart from each other.

The next comparison is regarding the LR of L1 and L2 speakers calculated in terms of the second group of LR measures, namely those covering the partial spectrum of the frequencies of all types in a text. As demonstrated by Table 2, the performances of L1 speakers are better than those of L2 speakers overall. For example, the average values

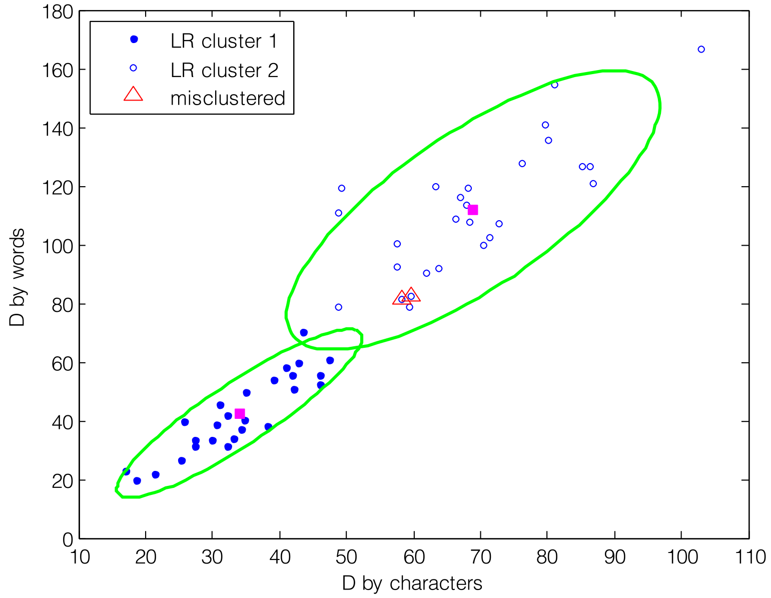

Fig. 1 Differentiation of L1 and L2 speakers using LR measure D

of V2, calculated by characters, are 95.8400 for L1 versus 56.2857 for L2 (t = 8.7976, p < 10E−4 for two-sample comparison of means, equal variance not assumed). However, all the speakers' normalized standard deviation did not show as apparent a disparity between L1 and L2 speakers as those of the first group of measures, where type and token information are the focused concern. In other words, according to the second group of LR measures, L1 speakers exhibit less homogeneous lexical patterns in oral discourse of the sociocultural topics selected by the test. The iterated clustering results using V_2 (as an example of this group of LR measures) and the quality of classification of the clustering procedure are provided in Fig. 2. These results, intuitively explained by the higher overlapping between the 90% confidence ellipses of the original L1 and L2 datasets, tend to suggest that the classification quality of this group of LR measures is lower than that of the first group. The average Silhouette values for the second group of measures is only 0.5304, with a 14% decline from that of the first group (0.6124 for V_1, 0.5876 for V_2, 0.5277 for Honore, 0.5221 for V_2TR, 0.4845 for V_1TR, 0.4482 for Sichel, in decreasing order, to make a complete comparison). The average standard deviation for the recorded Silhouette values is 0.1874 (specifically, in increasing order, 0.1680 for Sichel, 0.1705 for V_2TR, 0.1837 for V_2, 0.1933 for V1, 0.2022 for Honore, 0.2065 for V_1TR). The accuracy rates for classification, in decreasing order, are 94%, 89%, 89%, 85%, 70%, 51%, respectively, for V_1, V_2, Honore, V_1TR, V_2TR, and Sichel. For within-group ranking, it is reasonable to conclude that V_1 is the best classifier, as it reports the highest accuracy of classification and Silhouette values, followed perhaps by V_m and Honore. Cleary, Sichel reports the worst classification results (51% for accuracy and 0.4482 for Silhouette).

For group III measures used to differentiate LR between L1 and L2, namely, entropy, relative entropy, Yule K, Yule I, and V_m, the means for L1 are again evidently higher than L2. As an example of this group, the average Yule I value is 111.9223 for L1 versus 20.0994 for L2, calculated by characters. The two-sample comparison of means (with

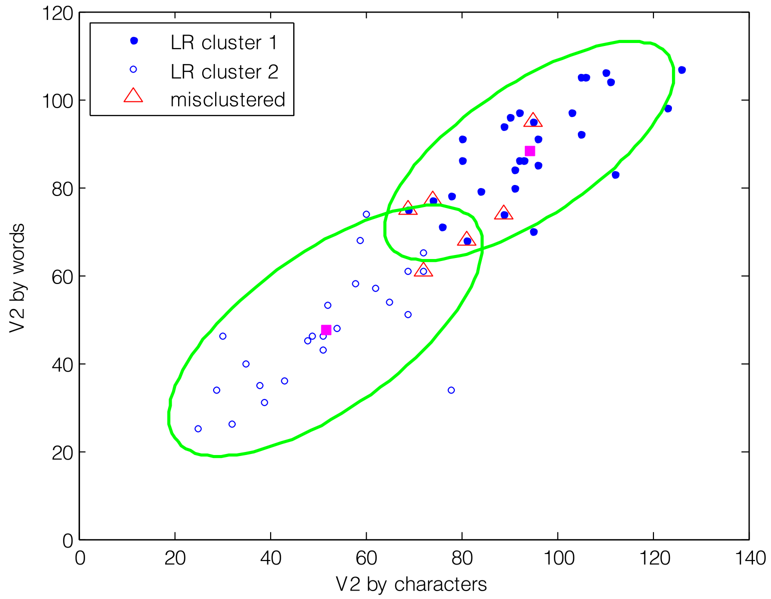

Fig. 2 Differentiation of L1 and L2 speakers using LR measure V1

equal means not assumed) yields t = 7.5320 and p < 10E−4. However, the normalized standard deviations of the results exhibit more mixed patterns compared to the first and second measure groups. The NSDs of this group are generally more extensive than the other two, particularly Yule K, Yule I, and V_m (the NSDs for them are 0.4039, 0.4154, and 0.4587, respectively, calculated by words). This implies that while L1 speakers outperform L2 speakers in general, the within-group difference from one speaker to another is substantial, which further stresses the necessity of clustering analysis. The individual-specific information is more of concern.

Taking Yule I to represent the group III measures, Fig. 3 plots the iterated clustering results and the corresponding classification quality, where again each clustering program is replicated 10 times to minimize the error induced from the randomness of initial guesses when the program is run. Within this group, the best classification measures, judged by the accuracy rate alone, are Yule I, with an accuracy rate of 77%. For the other four measures, namely, relative entropy, V_m, Yule K, and entropy, the accuracy rates range from 70 to 66%, which are roughly at the same level, disregarding the statistical errors applicable to the relatively small sample size of the current study. Although the average Silhouette values for V_m and Yule K are higher than those for the other three (0.7126 and 0.7028 versus 0.5121, 04754, and 0.4295), it is likely that Yule I is the best performer in this group when the two criteria are combined.

Classification performance of the LR measures

In summary of the ANOVA and clustering results based on the three different LR measures, one fundamental conclusion is that the lexical proficiency of L1 speakers is significantly higher than that of L2 speakers, whichever usual LR measure is applied in the analysis. This is indeed consistent with the results reported in similar studies such as

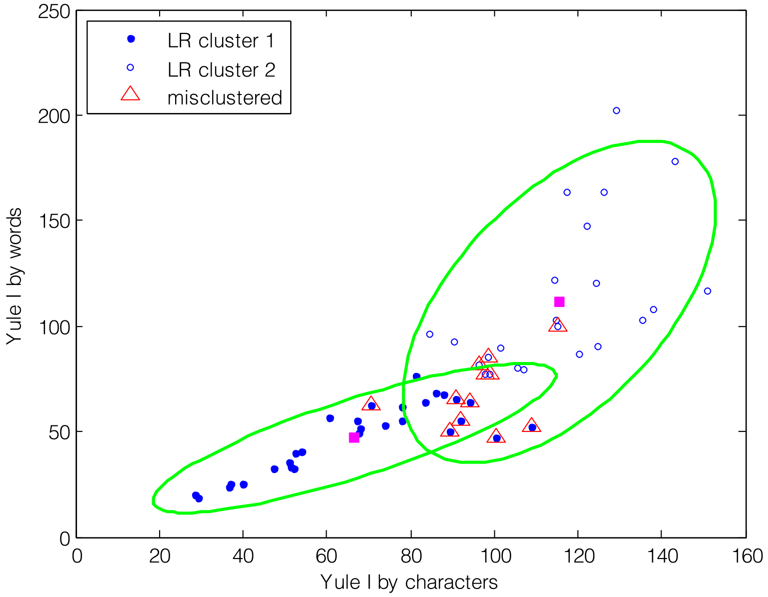

Fig. 3 Differentiation of L1 and L2 speakers using LR measure Yule I

Siskova (2012), Crossley and McNamara (2009), and Daller and Xue (2007) for comparison of L1 and L2 of English. But the disparity is valid for sure only in the average sense. Less predictable or even less accurate classifications exist when the clustering procedure is applied to decide whether a speaker's spoken text is native or non-native. This is even though some measures, namely, D or RootTTR, for instance, can yield relatively highly accurate classifications. On the other hand, the effectiveness of the many existing LR measures for L1–L2 lexical proficiency differentiation is highly varied, as large variances are observed in both the mean Silhouette values and the accuracy classification rates (0.561 for mean Silhouette values and 0.432 for correct classification rates). Considering aspects of LR that are not addressed in the current study, these results underscore the necessity of the quest for more comprehensive LR measures (Jarvis, 2013; Tweedie & Baayen, 1998).

Figure 4 presents the 2-dimensional scatterplot of the performance of all the 18 LR measures used for the current study. The tabulated numeric results of the clustering quality and classification performance of all the LR measures are provided in Appendix.

It is demonstrated that the first group of measures performed overwhelmingly better than the other two groups in terms of both classification accuracy and the mean Silhouette values. The other two groups of measures, namely those attempting to accommodate partial- or full-spectrum frequency information of the types appearing in a text, produced mixed performances. For instance, the performance of V_1 alone is close to that of D or RootTTR, while the V_1 to Token Ratio of the same group performed much less satisfactorily. Sichel, also a member of the second group of measures, performed even worse in terms of classification accuracy and Silhouette values. For another example, Yule K and entropy consider a full spectrum of types and produced very different Silhouette values, one very large and the other relatively small. On the other hand,

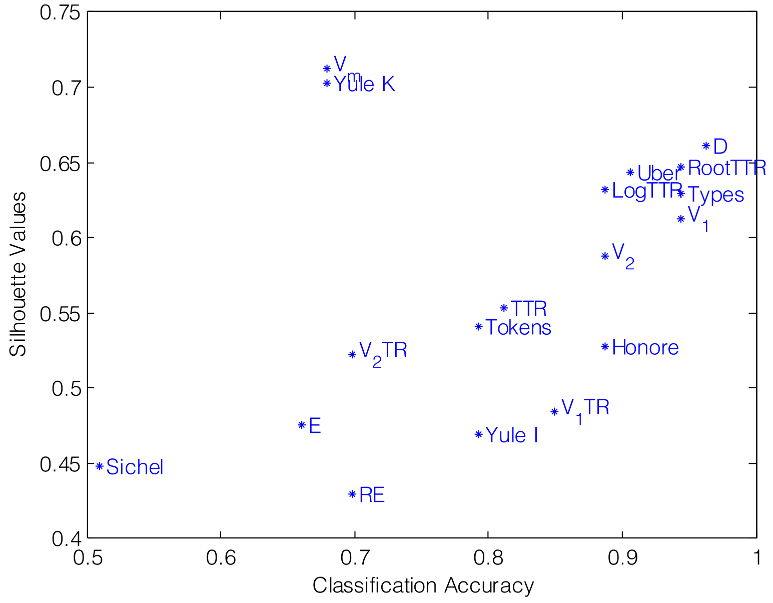

Fig. 4 Performance comparison of all the selected measures in the study

V_m, which is only a simple algebraic transformation of Yule K, yields a substantially more accurate classification.

Discussion

The current study categorized the LR measures according to how much and how substantially the spectrum information of all types in a text is used in constructing the measure. The more profound hope of such a categorization, instead of being a matter of convenience, is to validate whether more spectrum information correlates to more comprehensive accounts of LR. The more spectrum information is probably included in the construct of measure, the more effective it will be to the concerned differentiation task. The classification of the LR measures based on the spectrum inclusion is an effort to validate such a notion. However, the results provided by the study are mixed in addressing this notion. Nominally, D does not belong to the full-spectrum group, whereas it records the best performance. However, D has at least some tinge of full spectrum according to the iterative and interpolative nature of how it is calculated (i.e., the recursive procedure to truncate the whole text into consecutive short subtexts with smaller lengths and then average the D's obtained for each subtext). On the other hand, entropy and relative entropy, which have been successfully and routinely used in many different fields such as ecology and biology for quantifying the level of diversity and complexity of a system, have not generated as sound performance for L1 versus L2 differentiation as other primary measures, such as RootTTR.

It is worthy to note that, although the findings were based on Chinese, a generalization of the approach could be reasonably extended to other languages, including English. Such extension is logical despite the difference between English and Chinese. For instance, the basic unit of Chinese writing is the character, each standing for a morpheme rather than a phoneme. Such differences entail more computing challenges for Chinese text parsing instead of a fundamental difference in the applicableness of the clustering approaches herein discussed. This is because, as suggested by Halliday (2016), for instance, the semantic differences between any two languages should be bounded by common semantic space. In short, the difference between English and Chinese does not pose a substantial hurdle for the LR measures and their classification powers to be applied for English or other alphabetic languages.

Given its main focus is to compare the L1 and L2 production of Mandarin Chinese, the current study did not restrict the CFL learners to be homogeneous. Suppose one lexical richness measure is effective enough in predicting the nativeness versus non-nativeness of a speaker. In that case, its predicting power should be only more effective when homogeneity of the CFL learners is imposed. Accordingly, however, one limitation of the current study is that it remains unclear how such homogeneity requirement may impact the relative rankings of all the interested lexical richness measures regarding the accuracy in L1 and L2 differentiation. For instance, will D be consistently demonstrated as the most effective lexical richness measure for the same prediction task when CFL learners are restricted to a particular specific level of proficiency? Or will the relative performances of all the concerned lexical richness indices be affected by the demographic parameters such as age and level of education? Studies in this regard constitute a natural extension of the current study. The CFL learners in the current study had varied self-reported prior experience of Mandarin Chinese. It is not probable that the intermediate CFL learners will have the same production proficiency as advanced learners. Thus, it is an even further extension to investigate whether these lexical measures remain equally effective in differentiating the different proficiency levels among L2 learners and discerning the nativeness versus non-nativeness of a speaker. Ideally, the classification performance of the LR measures could be improved if a numeric value denoting the level of sophistication or difficulty pertaining to each Chinese character or word used in a text. Then, all the measures should be reevaluated with a weight for each lexical component incorporated. This could be a plausible direction, echoing, in a sense, the lexical frequency distribution approach for English (Beglar & Nation, 2013; Laufer & Nation, 1995); or the incorporation of information beyond frequency alone, e.g., collocation and semantic association (Pace-Sigge, 2018, for instance). Through comparing the notions of diversity and complexity in neighborhood fields, particularly ecology and biology, where compository properties are more naturally attached to the definition of diversity and richness, Jarvis (2013) suggests including several more dimensions atop the current concept of LR in linguistics. This is a theoretically promising yet practically challenging journey since the concepts such as dispersion and evenness can be challenging to quantify in their own right. Nevertheless, how this multivariate idea facilitates the relevant linguistic realities in the Chinese language is a worthwhile future direction.

Conclusion

Lexical proficiency differentiation between native and non-native English speakers based primarily on quantifiable LR analysis has been scarce. The application to Chinese speakers is virtually non-existent. The current paper is the first study to systematically investigate how L1 and L2 speakers of Chinese differ in LR using a reasonably large and authentic spoken corpus (average tokens of 2000+ characters per text). Altogether, 18 LR measures, grouped in 3 categories according to how the Chinese character and word spectrum information are used, have been selected and tested against each other using clustering algorithms. The relative efficacy and efficiency of the chosen measures are thoroughly calibrated. D records the best performance at an individual measure level in terms of clustering quality and correctness of group prediction (L1 versus L2 Chinese speakers). The conclusion is consistent more or less with those suggested by Jarvis (2002), Malvern and Richards (2012), Crossley et al. (2011), Silverman and Ratner (2002), and Duran et al. (2004), for instance. Other measures recording performances close to D include RootTTR, Type, V_1, LogTTR, and Uber. At categorical levels, it appears that improved variants of TTR measures have performed best (counting D into this category). For full-frequency spectrum accounted measures, the performance varies but is below that of the first group in general. For instance, the Silhouette values of Yule I and V_m are by far the highest among them all, with an acceptable level of classification accuracy. On the other hand, entropy and relative entropy did not profile impressive enough performances. Measures falling in the partial spectrum category have shown the most varied performances, where V_1 and V_2, for instance, work well enough both in terms of classification rate and Silhouette values while Sichel performs the worst.

Theoretically, the current study's findings demonstrated that the CFL learners' lexical proficiency, compared to that of Chinese native speakers, can be effectively fathomed, jointly if not single-handedly, by LR indices profiled by the learner's oral discourse. A complete ranking of the effectiveness of all the lexical richness measures for such differentiation tasks was generated, shedding important insights to the ongoing as well emerging lexicon-based researches in a very broad context, including, but not limited to artificial intelligence and natural language processing, for instance. Essentially non-parametric and model-free in nature, the clustering analysis and the algorithm proposed in the current study do not require a prior specification of the percentages of the participants with different language backgrounds. Thus, it may serve as a very robust benchmark scheme to help the researchers in language testing, for instance, to develop and calibrate new LR measures or conduct corpus-based linguistic analysis in a broad sense.

Practically and pedagogically, the current study's findings showed that the LR measures could effectively facilitate, if not completely replace, human assessment of the lexical progress of L2 learners. Some may argue that such a differentiation task might not be very challenging to human raters. But such argument overlooked the fact that human raters may make use of information beyond lexical richness itself such as cohesion patterns, tones and accents, prosodic features, and other sociolinguistic references contained in the oral production to distinguish a non-native speaker from a native one. Whereas when resources to determine the nativeness of a speaker is limited or restricted to the lexical components of their linguistic production, the benefits of the automated LR measures recommended in the current study will prevail, and many times can avoid the inconstancies seen in the ratings across different human raters or those

Appendix

Table 5 Classification performance of the LR measures selected in the study

	Mean silhouette	SD (silhouette)	Number of misspecifications	Classification accuracy
D	0.6611	0.1732	2	96.23%
Types	0.6289	0.2187	3	94.34%
RootTTR	0.6469	0.1539	3	94.34%
V_1	0.6124	0.1933	3	94.34%
Uber	0.6431	0.1577	5	90.57%
LogTTR	0.6318	0.1664	6	88.68%
V_2	0.5876	0.1837	6	88.68%
Honore H	0.5277	0.2022	6	88.68%
V_1TR	0.4845	0.2065	8	84.91%
TTR	0.5532	0.1703	10	81.13%
Tokens	0.5408	0.1804	11	79.25%
Yule I	0.5121	0.1773	12	77.36%
RE	0.4295	0.202	16	69.81%
V_2TR	0.5221	0.1705	16	69.81%
Yule K	0.7028	0.189	17	67.92%
V_m	0.7126	0.1797	17	67.92%
E	0.4754	0.1698	18	66.04%
Sichel S	0.4482	0.168	26	50.94%

of the same rater across different times. Overall, given the fallibility and surging cost of human ratings, the results reported by the current study should be highly beneficial for L2 educators who strive for a reliable automated rating tool based on lexical information.

Abbreviations

L2: Second language; CFL: Chinese as a foreign language; ESL: English as the second language; BJKY: Beijing Mandarin Spoken Corpora; SLA: Second language acquisition; LR: Lexical richness; CLAN: Computerized Language Analysis Program; CHILDES: Child Language Data Exchange System; BCT: Business Chinese Test; SD: Standard deviation; NSD: Normalized standard deviation; CI: Confidence interval

Acknowledgements

Not applicable

Authors' contributions

WW handled the experiment, collected the raw data, and provided preliminary analysis. YZ proposed the approach to differentiate lexical proficiencies with LR measures, surveyed the literature, analyzed the data, carried out the empirical test, and completed the manuscript writing. Both authors read and approved the final manuscript.

Funding

Not applicable

Availability of data and materials

Interested researchers may communicate with the corresponding author about the transcribed data for academic research only. Other conditions and limitations may apply.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]University of Nottingham Ningbo China, Ningbo, China. [2]The Chinese University of Hong Kong, Shatin, Hong Kong.

Received: 16 February 2021 Accepted: 25 June 2021
Published online: 01 September 2021

References

Akiyama, Y., & Saito, K. (2016). Development of comprehensibility and its linguistic correlates: a longitudinal study of video-mediated telecollaboration. *The Modern Language Journal*, *100*(3), 585–609. https://doi.org/10.1111/modl.12338.

Alqahtani, M. (2015). The importance of vocabulary in language learning and how to be taught. *International Journal of Teaching and Education*, *3*(3), 21–34.

Anderson, R. C., & Freebody, P. (1981). Vocabulary and knowledge. In J. T. Gutrie (Ed.), *Comprehension and teaching: Research review*, (pp. 77–117). Newark, DE: International Reading Association.

Beglar, D., & Nation, P. (2013). Assessing vocabulary. *The Companion to Language Assessment*, *2*(10), 72–184.

Bosker, H. R., Quene, H., Sanders, T., & de Jong, N. H. (2014). The perception of fluency in native and nonnative speech. *Language Learning*, *64*(3), 579–614. https://doi.org/10.1111/lang.12067.

Connor, U. (1984). A study of cohesion and coherence in ESL students' writing. *Papers in Linguistics: International Journal of Human Communication*, *17*(3), 301–316. https://doi.org/10.1080/08351818409389208.

Crossley, S. A., & McNamara, D. S. (2009). Computationally assessing lexical differences in second language writing. *Journal of Second Language Writing*, *17*(2), 119–135.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243–263.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in spontaneous speech of bilinguals. *Applied Linguistics*, *24*(2), 197–222. https://doi.org/10.1093/applin/24.2.197.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge*, (pp. 150–164). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.011.

Déogratias, N. (2011). The relationship between lexical competence, collocational competence, and second language proficiency. *English Text Construction*, *4*(1), 113–145.

Dugast, D. (1979). Vocabulaire et stylistique. I Théâtre et dialogue. In *Travaux delinguistique quantitative*. Geneva: Slatkine-Champion.

Duran, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*(2), 220–242. https://doi.org/10.1093/applin/25.2.220.

Ellis, R. (1995). Modified oral input and the acquisition of word meanings. *Applied Linguistics*, *16*(4), 409–435. https://doi.org/10.1093/applin/16.4.409.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *30*(4), 474–509. https://doi.org/10.1093/applin/amp042.

Farahani, A. A. K., Nemati, M., & Montazer, M. N. (2019). Assessing peer review pattern and the effect of face-to-face and mobile-mediated modes on students' academic writing development. *Language Testing in Asia*, *9*(1), 1–24.

Golshaie, R. (2016). A corpus study on identification and semantic classification of light verb constructions in Persian: the case of the light verb xordan 'to eat/collide'. *Language Sciences*, *57*, 21–33. https://doi.org/10.1016/j.langsci.2016.05.002.

Gu, Y. (2019). Vocabulary Learning Strategies. *The Encyclopedia of Applied Linguistics*, 1–7.

Guiraud, P., 1960. Proble'mes et me´thodes de la statistique linguistique. D. Reidel.

Halliday, M. A. K. (2016). English and Chinese: similarities and differences. *Aspects of Language and Learning*. https://doi.org/10.1007/978-3-662-47821-9_6.

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2015). *Handbook of cluster analysis*. CRC Press. https://doi.org/10.1201/b19706.

Herdan, G. (1960). *Quantitative linguistics*. London: Butterworth.

Honored, A. (1979). Some simple measures of richness of vocabulary. *Association of Literary and Linguistic Computing Bulletin*, *7*, 172–177.

Hoover, D. L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, *37*(2), 151–178. https://doi.org/10.1023/A:1022673822140.

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, *21*(2), 181–193. https://doi.org/10.1017/S0272263199002028.

In'nami, Y., Koizumi, R., & Nakamura, K. (2016). Factor structure of the Test of English for Academic Purposes (TEAP®) test in relation to the TOEFL iBT® test. *Language Testing in Asia*, *6*(1), 1–23.

Jarvis, S. (2002). Short texts, best fitting curves, and new measure of lexical diversity. *Language Testing*, *19*(1), 57–84. https://doi.org/10.1191/0265532202lt220oa.

Jarvis, S. (2013). Capturing diversity in lexical diversity. *Language Learning*, *63*, 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Lund Working Papers in Linguistics*, *53*, 61–79.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316801.

Koda (1988). Cognitive process in second language reading: transfer of L1 reading skills and strategies. *Second Language Research*, *4*, 133–156.

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307.

Li, L. (2018). L1 influence on coherence-building skills in L2 Chinese reading. In X. Wen, & X. Jiang (Eds.), *Studies on learning and teaching Chinese as a second language*, (vol. 1, pp. 86–104). Routledge. https://doi.org/10.4324/9781351208673-5.

MacWhinney, B. (2007). The TalkBank Project. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora: Synchronic databases*, (vol. 1, pp. 163–180). Houndmills, UK: Palgrave-Macmillan. https://doi.org/10.1057/9780230223936_7.

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (Eds.), *Evolving models of language*, (pp. 58–71). Clevedon, UK: Multilingual Matters.

Malvern, D., Richards, B., Chipere, N., & Duran, P. (2004). Lexical Diversity and Language. In *Development: Quantification and Assessment*. Palgrave Macmillan. https://doi.org/10.1057/9780230511804.

Malvern, D. D., & Richards, B. (2012). Measures of lexical richness. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell/Wiley. https://doi.org/10.1002/9781405198431.wbeal0755.

McCarthy, P. M., & Jarvis, S. (2007). vocd: a theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767.

Newman, R. S., Rowe, M. L., & Ratner, N. B. (2016). Input and uptake at 7 months predicts toddler vocabulary: the role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, *43*(5), 1158–1173. https://doi.org/10.1017/S0305000915000446.

Pace-Sigge, M. (2018). How homo economicus is reflected in fiction – a corpus linguistic analysis of 19th and 20th century capitalist societies. *Language Sciences*, *70*, 103–117.

Reynolds, D. W. (1995). Repetition in nonnative speaker writing. *Studies in Second Language Acquisition*, *17*(2), 185–209. https://doi.org/10.1017/S0272263100014157.

Schmid, M. S. (2010). Language attrition and identity. In S. Han, & E. Poppel (Eds.), *Culture and Neural Frames of Cognition and Communication*, (pp. 185–205). Springer.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*(1), 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x.

Shin, D. (2019). Analyzing media discourse on the development of the National English Ability Test (NEAT) in South Korea. *Language Testing in Asia*, *9*(1), 1–14.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, *70*, 542–547.

Silverman, S., & Ratner, N. B. (2002). Measuring lexical diversity in children who stutter: application of vocd. *Journal of Fluency Disorders*, *27*(4), 289–304. https://doi.org/10.1016/S0094-730X(02)00162-6.

Siskova, Z. (2012). Lexical richness in EFL students' narratives. *University of Reading Language Studies Working Papers*, *4*, 26–36.

Skehan, P. (2009). Modelling second language performance: integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047.

Smith, J. A., & Kelly, C. (2002). Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities*, *36*(4), 411–430. https://doi.org/10.1023/A:1020201615753.

Sultana, N. (2019). Language assessment literacy: an uncharted area for the English language teachers in Bangladesh. *Language Testing in Asia*, *9*(1), 1–14. https://doi.org/10.1186/s40468-019-0077-8.

Tweedie, F., & Baayen, R. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323–352. https://doi.org/10.1023/A:1001749303137.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards, & B. Laufer (Eds.), *Vocabulary in a second language*, (pp. 173–189). Amsterdam: Jonh Benjamins. https://doi.org/10.1075/lllt.10.13ver.

Wright, T. S., & Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly*, *52*(2), 203–226. https://doi.org/10.1002/rrq.163.

Xiao, Z., McEnery, A., Baker, P., & Hardie, A. (2004). Developing Asian language corpora: standards and practice. In *Proceedings of the 4th Workshop on Asian Language Resources, Hainan*, (pp. 1–8).

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Zhang, Y. (2014). A corpus based analysis of lexical richness of Beijing Mandarin speakers: variable identification and model construction. *Language Sciences*, *44*, 60–69. https://doi.org/10.1016/j.langsci.2013.12.003.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.